Algebra & Number Theory

Volume 9

2015

No. 4

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

The subscription price for 2015 is US $255/year for the electronic version, and $440/year (+$55, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscribers address should be sent to MSP.

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

ANT peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.767

Motivic Donaldson–Thomas invariants of small crepant resolutions

Andrew Morrison and Kentaro Nagao

We compute the motivic Donaldson–Thomas theory of a small crepant resolution of a toric Calabi–Yau 3-fold.

Introduction

This paper is a continuation of [Morrison et al. 2012]. We study the *motivic Donaldson–Thomas invariants* of noncommutative and commutative crepant resolutions of the affine toric Calabi–Yau 3-fold $\{XY - Z^{N_0}W^{N_1}\} \subset \mathbb{C}^4$.

A *Donaldson–Thomas* (DT) *invariant* of a Calabi–Yau 3-fold Y is a counting invariant of coherent sheaves on Y, introduced in [Thomas 2000] as a holomorphic analogue of the Casson invariant of a real 3-manifold. A component of the moduli space of stable coherent sheaves on Y carries a symmetric obstruction theory and a virtual fundamental cycle [Behrend and Fantechi 1997; 2008]. A DT invariant of a compact Y is then defined as the integral of the constant function 1 over the virtual fundamental cycle of the moduli space.

MSC2010: 14N35.
Keywords: motivic Donaldson–Thomas invariants, small crepant resolutions.

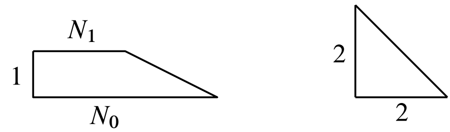


Figure 1. Polygons for $\mathcal{X}_{N_0,N_1}$ and $\mathcal{X}_{(\mathbb{Z}/2\mathbb{Z})^2}$.

It is known that the moduli space of coherent sheaves on Y can be locally described as the critical locus of a function, the *holomorphic Chern–Simons* (CS) *functional* (see [Joyce and Song 2012]). Behrend [2009] provided a description of DT invariants in terms of the Euler characteristic of the *Milnor fiber* of the CS functional. Inspired by this result, the proposal of [Kontsevich and Soibelman 2008; Behrend et al. 2013] was to study the *motivic Milnor fiber* of the CS functional as a motivic refinement of the DT invariant. Such a refinement had been expected in string theory [Iqbal et al. 2009; Dimofte and Gukov 2010].

On the other hand, in [Szendrői 2008] it was proposed to study counting invariants for the noncommutative crepant resolution (NCCR) of the conifold, which are called noncommutative Donaldson–Thomas (NCDT) invariants. It was also conjectured there that NCDT and DT invariants are related by wall-crossing. The paper [Nagao and Nakajima 2011] realized this, by:

- Describing the chamber structure on the space of stability parameters for the NCCR.
- Finding chambers which correspond to geometric DT and stable pair (PT) invariants, as well as NCDT invariants.
- Computing the generating function of DT-type invariants for each chamber.

For the conifold, the dimension of the fiber of the crepant resolution is less than 2 (we say that the resolution is small). This condition plays an important role in many places of the paper. Affine toric Calabi–Yau 3-folds which have small crepant resolutions are classified as follows:

(1) $\mathcal{X} = \mathcal{X}_{N_0,N_1} := \{XY - Z^{N_0}W^{N_1}\}$ for $N_0 > 0$ and $N_1 \geq 0$.

(2) $\mathcal{X} = \mathcal{X}_{(\mathbb{Z}/2\mathbb{Z})^2} := \mathbb{C}^3/(\mathbb{Z}/2\mathbb{Z})^2$, where $(\mathbb{Z}/2\mathbb{Z})^2$ acts on $\mathbb{C}^3$ with weights $(1,0)$, $(0,1)$ and $(1,1)$.

In [Nagao 2012], counting invariants for noncommutative and commutative crepant resolutions of $\{XY - Z^{N_0}W^{N_1}\}$ were studied. First, we provided descriptions of NCCRs of $\{XY - Z^{N_0}W^{N_1}\}$ in terms of a quiver with potential. Given N_0 and N_1, the quivers with potential are not unique. However it was also shown that any such quivers with potential are related by a sequence of mutations. Finally, generalizations of the results in [Nagao and Nakajima 2011] are given.

In [Morrison et al. 2012], we provided motivic refinements of formulae in [Nagao and Nakajima 2011]. For the proof, we needed one explicit evaluation of the "universal" series [Morrison et al. 2012, §2] and a wall-crossing argument [Morrison et al. 2012, §3].

In this paper, we will show similar formulae for $\{XY - Z^{N_0}W^{N_1}\}$, that is, motivic refinements of the formulae in [Nagao 2012]. The wall-crossing argument works without modifications (Section 6), while the evaluation part is more involved (Theorem 0.1). Our strategy is as follows:

- First, in Section 4, we evaluate the universal series for a specific NCCR using a generalization of the calculation of [Morrison et al. 2012, §2.2].
- Then, in Section 5, we evaluate the universal series for a general NCCR. Nagao [2011c] provided a formula which describes how the universal series changes under mutation. Although it is assumed that the quiver has no loops and 2-cycles in [Nagao 2011c], we can apply a parallel argument in our setting as well.

Since any two NCCRs are related by a sequence of mutations, the evaluation is done.

Main result

Let Γ be a quadrilateral (or a triangle in the case $N_1 = 0$) as in Figure 1 and σ a partition of Γ, that is, a division of Γ into an N-tuple of triangles with area $1/2$, as in Figure 2. We will associate σ to a quiver with superpotential $(Q_\sigma, \omega_\sigma)$ (see Section 2A). The set of vertices of the quiver Q_σ is $\widehat{I} := \mathbb{Z}/N\mathbb{Z}$, which is identified with $\{0, \dots, N-1\}$. A vertex has a loop if and only if it is in the subset $\widehat{I}_r \subset \widehat{I}$ (see (2-1) for the definition). It is shown in [Nagao 2012, §1] that the Jacobian algebra $J_\sigma := J(Q_\sigma, \omega_\sigma)$ is an NCCR of

$$\mathcal{X} := \operatorname{Spec}(\mathbb{C}[X, Y, Z, W]/(XY - Z^{N_0}W^{N_1})).$$

Let Δ be the set of roots of type $\widetilde{A}_N$ and let $\Delta_{\sigma,+}$ $\Delta^{\mathrm{re}}_{\sigma,+}$, and $\Delta^{\mathrm{im}}_{\sigma,+}$ denote the set of positive, positive real, and positive imaginary roots, respectively.[1]

For $\alpha \in \mathbb{N}^{\widehat{I}}$, let $\mathfrak{M}(J_\sigma, \alpha)$ be the moduli stack of J_σ-modules V with $\underline{\dim}\, V = \alpha$. We define the generating series of the motivic DT invariants of (Q_σ, W_σ) by

$$A^\sigma_U(y) = A^\sigma_U(y_0, \dots, y_{N-1}) := \sum_{\alpha \in \mathbb{N}^{Q_0}} [\mathfrak{M}(J_\sigma, \alpha)]_{\mathrm{vir}} \cdot y^\alpha \in \mathcal{M}_{\mathbb{C}}[\![y_0, \dots, y_{N-1}]\!].^{2}$$

[1] From the view point of the root system, a choice of a partition σ corresponds to a choice of a set of simple roots.

[2] For the wall-crossing of motivic DT theory, a twisted product on the y_α twisted by the Euler form plays a crucial role. In this case, the twisted product coincides with the usual commutative product since the Euler form is trivial.

Here $y^\alpha := \prod (y_i)^{\alpha_i}$ and $[\bullet]_{\mathrm{vir}}$ denotes the *virtual motive* (see Section 3A), an element of a suitable ring of motives $\mathcal{M}_{\mathbb{C}}$. The subscript refers to the fact that we think of this series as the universal series.

To each root $\alpha \in \Delta_{\sigma,+}$ we associate an infinite product as follows:

- For a real root $\alpha \in \Delta^{\mathrm{re}}_{\sigma,+}$ such that $\sum_{k \notin \widehat{I}_r} \alpha_k$ is odd, put

$$A^\alpha(y) := \mathrm{Exp}\left(\frac{-\mathbb{L}^{-\frac{1}{2}}}{1-\mathbb{L}^{-1}} y^\alpha\right) = \prod_{j \geq 0} (1 - \mathbb{L}^{-j-\frac{1}{2}} y^\alpha).$$

- For a real root $\alpha \in \Delta^{\mathrm{re}}_{\sigma,+}$ such that $\sum_{k \notin \widehat{I}_r} \alpha_k$ is even, put

$$A^\alpha(y) := \mathrm{Exp}\left(\frac{1}{1-\mathbb{L}^{-1}} y^\alpha\right) = \prod_{j \geq 0} (1 - \mathbb{L}^{-j} y^\alpha)^{-1}.$$

- For an imaginary root $\alpha \in \Delta^{\mathrm{im}}_{\sigma,+}$, put

$$A^\alpha(y) := \mathrm{Exp}\left(\frac{N-1+\mathbb{L}}{1-\mathbb{L}^{-1}} y^\alpha\right) = \prod_{j \geq 0} (1 - \mathbb{L}^{-j} y^\alpha)^{1-N} \cdot (1 - \mathbb{L}^{-j+1} y^\alpha)^{-1}.$$

The main result of this paper is the following formula:

Theorem 0.1.
$$A_U^\sigma(y) = \prod_{\alpha \in \Delta_{\sigma,+}} A^\alpha(y).$$

This is proved in Section 4 and Section 5B.

Corollaries

Let $\widetilde{J}_\sigma = J(\widetilde{Q}_\sigma, W_\sigma)$ be the framed algebra given by adding the new vertex ∞ and the new arrow from ∞ to 0 to the quiver of J_σ. Nagao and Nakajima [2011] introduced a notion of ζ-(semi)stability of $\widetilde{J}$-modules $\widetilde{V}$ with $\dim \widetilde{V}_\infty \leq 1$ for a stability parameter $\zeta \in \mathbb{R}^{\widehat{I}}$.

For $\alpha \in \mathbb{N}^{\widehat{I}}$, let $\mathfrak{M}_\zeta(\widetilde{J}, \alpha)$ be the moduli space of ζ-stable $\widetilde{J}$-modules $\widetilde{V}$ with $\underline{\dim}\, \widetilde{V} = (\alpha, 1)$. We want to compute the motivic generating series

$$Z_\zeta(y) = Z_\zeta(y_0, \ldots, y_{N-1}) := \sum_{\alpha \in \mathbb{N}^{\widehat{I}}} [\mathfrak{M}_\zeta(\widetilde{J}, \alpha)]_{\mathrm{vir}} \cdot y^\alpha \in \mathcal{M}_{\mathbb{C}}[\![y_0, \ldots, y_{N-1}]\!].$$

For each root $\alpha \in \Delta_{\sigma,+}$, we put

$$Z_\alpha(y_0, \ldots, y_{N-1}) := \frac{A^\alpha(-\mathbb{L}^{\frac{1}{2}} y_0, y_1, \ldots, y_{N-1})}{A^\alpha(-\mathbb{L}^{-\frac{1}{2}} y_0, y_1, \ldots, y_{N-1})}.$$

These are given as follows:

- For a real root $\alpha \in \Delta^{\mathrm{re}}_{\sigma,+}$ such that $\sum_{k \notin \widehat{I}_r} \alpha_k$ is odd, we have

$$Z_\alpha(-y_0, \dots, y_{N-1}) = \prod_{i=0}^{\alpha_0 - 1} (1 - \mathbb{L}^{-\frac{\alpha_0}{2} + \frac{1}{2} + i} y^\alpha).$$

- For a real root $\alpha \in \Delta^{\mathrm{re}}_{\sigma,+}$ such that $\sum_{k \notin \widehat{I}_r} \alpha_k$ is even, we have

$$Z_\alpha(-y_0, \dots, y_{N-1}) = \prod_{i=0}^{\alpha_0 - 1} (1 - \mathbb{L}^{-\frac{\alpha_0}{2} + 1 + i} y^\alpha)^{-1}.$$

- For an imaginary root $\alpha \in \Delta^{\mathrm{im}}_{\sigma,+}$, we have

$$Z_\alpha(-y_0, \dots, y_{N-1}) = \prod_{i=0}^{\alpha_0 - 1} (1 - \mathbb{L}^{-\frac{\alpha_0}{2} + 1 + i} y^\alpha)^{1-N} \cdot (1 - \mathbb{L}^{-\frac{\alpha_0}{2} + 2 + j} y^\alpha)^{-1}.$$

Applying the same argument as [Morrison et al. 2012, §3], we get the following formula (Section 6A):

Corollary 0.2. *For $\zeta \in \mathbb{R}^{\widetilde{I}}$ not orthogonal to any root, we have*

$$Z_\zeta(y) = \prod_{\substack{\alpha \in \Delta_{\sigma,+} \\ \zeta \cdot \alpha < 0}} Z_\alpha(y_0, \dots, y_{N-1}).$$

By [Behrend 2009; Behrend et al. 2013], the specialization $Z_\zeta(y)|_{\mathbb{L}^{\frac{1}{2}} \to 1}$ is the DT-type series at the generic stability parameter ζ, computed in [Nagao 2012].

Let $\mathcal{Y}_\sigma \to \mathcal{X}$ be the crepant resolution corresponding to σ. The noncommutative crepant resolution J_σ is derived equivalent to $\mathcal{Y}_\sigma$. In [Nagao and Nakajima 2011, §3], we found a stability parameter ζ_{DT} (resp. ζ_{PT}) such that the moduli space coincides with the Hilbert scheme (resp. the stable pair moduli space) for $\mathcal{Y}_\sigma$.

Let $Z^\sigma_{\mathrm{DT}}(s, T_1, \dots, T_{N-1})$ and $Z^\sigma_{\mathrm{DT}}(s, T_1, \dots, T_{N-1})$ be the generating functions of DT and PT invariants of $\mathcal{Y}_\sigma$, respectively. Here s is the variable for the homology class of a point and T_i is the variable for the homology class of the i-th component C_i of the exceptional curve. The variable change induced by the derived equivalence is given by

$$s := y_0 \cdot y_1 \cdots y_{N-1}, \quad T_i = y_i.$$

For $1 \le a \le b \le N-1$, we put

$$C_{[a,b]} := [C_a] + \cdots + [C_b] \in H_2(\mathcal{Y}_\sigma, \mathbb{Z}),$$

where C_i is a component of the exceptional curve, and let

$$T_{[a,b]} = T_a \cdots T_b$$

be the corresponding monomial. Let $c(a,b)$ denote the number of $(-1,-1)$-curves in $\{C_i \mid a \leq i \leq b\}$. We define infinite products as follows:

- If $c(a,b)$ is odd, we put

$$Z_{[a,b]} = Z_{[a,b]}(s, T_{[a,b]}) := \prod_{n=1}^{\infty}\left(\prod_{i=0}^{n-1}\left(1 - \mathbb{L}^{-\frac{n}{2}+\frac{1}{2}+i} \cdot (-s)^n \cdot T_{[a,b]}\right)\right).$$

- If $c(a,b)$ is even, we put

$$Z_{[a,b]} = Z_{[a,b]}(s, T_{[a,b]}) := \prod_{n=1}^{\infty}\left(\prod_{i=0}^{n-1}\left(1 - \mathbb{L}^{-\frac{n}{2}+1+i} \cdot (-s)^n \cdot T_{[a,b]}\right)^{-1}\right).$$

- For imaginary roots, we put

$$Z_{\mathrm{im}} = Z_{\mathrm{im}}(s) := \prod_{n=1}^{\infty}\left(\prod_{i=0}^{n-1}\left(1 - \mathbb{L}^{-\frac{n}{2}+1+i}(-s)^n\right)^{1-N}\left(1 - \mathbb{L}^{-\frac{n}{2}+2+i}(-s)^n\right)^{-1}\right).$$

Corollary 0.3. (1) *The refined DT and PT series of $\mathcal{Y}_\sigma$ are given by the formulae*

$$Z_{\mathrm{DT}}(s, T_1, \ldots, T_{N-1}) = Z_{\mathrm{im}}(s) \cdot \prod_{1\leq a\leq b\leq N-1} Z_{[a,b]}(s, T_{[a,b]}),$$
$$Z_{\mathrm{PT}}(s, T_1, \ldots, T_{N-1}) = \prod_{1\leq a\leq b\leq N-1} Z_{[a,b]}(s, T_{[a,b]}).$$

(2) *The generating function of virtual motives of the Hilbert scheme of points on $\mathcal{Y}_\sigma$ is given by the formula*

$$Z_{0\text{-dim}}(s) := \sum_{n=0}^{\infty} [(\mathcal{Y}_\sigma)^{[n]}]_{\mathrm{vir}} \cdot s^n = Z_{\mathrm{im}}.$$

(3) *The refined version of the DT-PT correspondence for $\mathcal{Y}_\sigma$ holds*:

$$Z_{\mathrm{DT}}(s, T_1, \ldots, T_{N-1}) = Z_{0\text{-dim}}(s) \times Z_{\mathrm{PT}}(s, T_1, \ldots, T_{N-1}).$$

Remark. The formula in (2) is a direct consequence of the formula for Z_{DT} in (1), since the polynomial in the $T_{[a,b]}$ variables does not contribute.

1. Root system of type $\widetilde{A}_N$

Let $N_0 > 0$ and $N_1 \geq 0$ be integers such that $N_0 \geq N_1$, and set $N = N_0 + N_1$. We set

$$\begin{aligned} I &= \{1, \dots, N-1\}, \\ \widehat{I} &= \{0, 1, \dots, N-1\}, \\ \widetilde{I} &= \left\{\tfrac{1}{2}, \tfrac{3}{2}, \dots, N - \tfrac{1}{2}\right\}, \\ \widetilde{\mathbb{Z}} &= \left\{n + \tfrac{1}{2} \mid n \in \mathbb{Z}\right\}. \end{aligned}$$

For $l \in \mathbb{Z}$ and $j \in \widetilde{\mathbb{Z}}$, let $\underline{l} \in \widehat{I}$ and $\underline{j} \in \widetilde{I}$ be the elements such that $l - \underline{l} \equiv j - \underline{j} \equiv 0$ modulo N.

Let $\mathbb{Z}^{\widehat{I}}$ be the free abelian group with basis $\{\alpha_i \mid i \in \widehat{I}\}$; we call α_i a simple root. We put

$$\begin{aligned} \Delta_+^{\mathrm{fin}} &:= \{\alpha_{[a,b]} := \alpha_a + \cdots + \alpha_b \mid 1 \leq a \leq b \leq N-1\}, \\ \Delta_+^{\mathrm{re},+} &:= \{\alpha_{[a,b]} + n \cdot \delta \mid \alpha_{[a,b]} \in \Delta_+^{\mathrm{fin}},\ n \in \mathbb{Z}_{\geq 0}\}, \\ \Delta_+^{\mathrm{re},-} &:= \{-\alpha_{[a,b]} + n \cdot \delta \mid \alpha_{[a,b]} \in \Delta_+^{\mathrm{fin}},\ n \in \mathbb{Z}_{> 0}\}, \end{aligned}$$

and

$$\Delta_+^{\mathrm{re}} := \Delta_+^{\mathrm{re},+} \sqcup \Delta_+^{\mathrm{re},-}, \quad \Delta_+^{\mathrm{im}} := \{n \cdot \delta \mid n \in \mathbb{Z}_{>0}\},$$

where $\delta := \alpha_0 + \cdots + \alpha_{N-1}$ is the (positive minimal) imaginary root.

For $k \in \widehat{I}$, the simple reflection at k is the group homomorphism given by

$$\begin{aligned} \mathbb{Z}^{\widehat{I}} &\to \mathbb{Z}^{\widehat{I}}, \\ \alpha_i &\mapsto \alpha_i - C_{ik} \cdot \alpha_k, \end{aligned}$$

where C is the Cartan matrix of type $\widetilde{A}_N$. This gives a self-bijection of $\Delta_+^{\mathrm{re},+} \backslash \{\alpha_k\}$.

2. Noncommutative crepant resolutions

2A. ***Quivers with potential.*** We denote by Γ the quadrilateral (or the triangle in case $N_1 = 0$) with vertices $(0,0)$, $(0,1)$, $(N_0, 0)$ and $(N_1, 1)$. Note that the affine toric Calabi–Yau 3-fold corresponding to Γ is $\mathcal{X} = \{XY - Z^{N_0} W^{N_1}\}$.

A partition σ of Γ is a pair of functions $\sigma_x : \widetilde{I} \to \widetilde{\mathbb{Z}}$ and $\sigma_y : \widetilde{I} \to \{0,1\}$ such that:

- $\sigma(i) := (\sigma_x(i), \sigma_y(i))$ gives a bijection between $\widetilde{I}$ and
$$\left\{\left(\tfrac{1}{2}, 0\right), \left(\tfrac{3}{2}, 0\right), \dots, \left(N_0 - \tfrac{1}{2}, 0\right), \left(\tfrac{1}{2}, 1\right), \left(\tfrac{3}{2}, 1\right), \dots, \left(N_1 - \tfrac{1}{2}, 1\right)\right\}.$$
- If $i < j$ and $\sigma_y(i) = \sigma_y(j)$ then $\sigma_x(i) > \sigma_x(j)$.

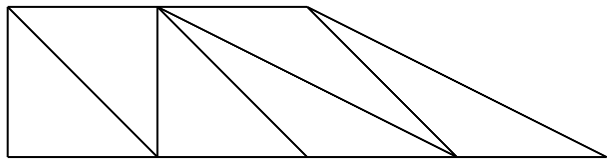


Figure 2. Γ_σ.

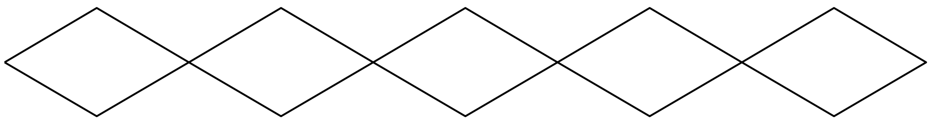


Figure 3. S.

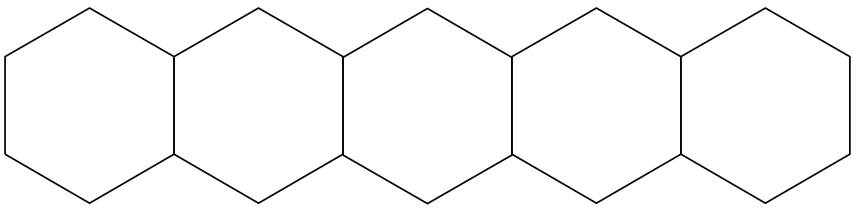


Figure 4. H.

Giving a partition σ of Γ is equivalent to dividing Γ into an N-tuple of triangles $\{T_i\}_{i\in\tilde{I}}$ with area $1/2$ so that T_i has $(\sigma_x(i) \pm 1/2, \sigma_y(i))$ as its vertices. Let Γ_σ be the corresponding diagram, Δ_σ the fan and $f_\sigma : \mathcal{Y}_\sigma \to \mathcal{X}$ the crepant resolution of $\mathcal{X}$. We put

$$\widehat{I}_r := \left\{k \in \widehat{I} \mid \sigma_y\left(k-\tfrac{1}{2}\right) = \sigma_y\left(k+\tfrac{1}{2}\right)\right\}. \tag{2-1}$$

Example 1. Let us consider as an example the case $N_0 = 4$, $N_1 = 2$ and

$$(\sigma(i))_{i\in\tilde{I}} = \left(\left(\tfrac{7}{2},0\right),\left(\tfrac{3}{2},1\right),\left(\tfrac{5}{2},0\right),\left(\tfrac{3}{2},0\right),\left(\tfrac{1}{2},1\right),\left(\tfrac{1}{2},0\right)\right).$$

We show the corresponding diagram Γ_σ in Figure 2.

Let S be the union of an infinite number of rhombi with edge length 1, as in Figure 3, located so that the centers of the rhombi are on a line parallel to the x-axis in $\mathbb{R}^2$, and let H be the union of an infinite number of hexagons with edge length 1, as in Figure 4 located so that the centers of the hexagons are in a line parallel to the x-axis in $\mathbb{R}^2$.

We form the sequence $\tau = \tau_\sigma : \mathbb{Z} \to \{S, H\}$ which maps l to S (resp. H) if l modulo N is not in $\widehat{I}_r$ (resp. is in $\widehat{I}_r$), and cover the whole plane $\mathbb{R}^2$ by arranging S's and H's according to this sequence (see Figure 5). We regard this as a graph

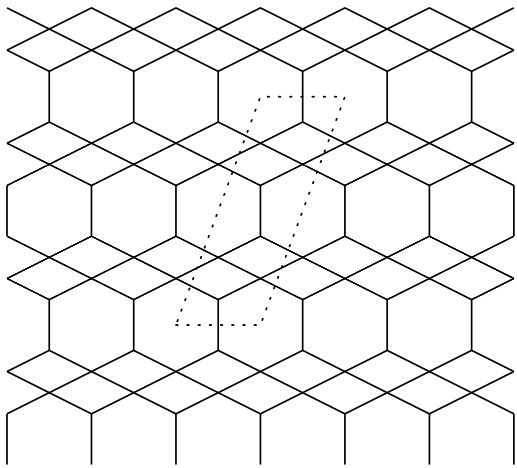


Figure 5. P_σ in the case of Example 1.

on the 2-dimensional torus $\mathbb{R}^2/\Lambda$, where Λ is the lattice generated by $(\sqrt{3},0)$ and $(N_0-N_1,(N_0-N_1)\sqrt{3}+N_1)$.

We can color the vertices of this graph black or white so that each edge connects a black vertex and a white one. Let P_σ denote this bipartite graph on the torus. For each edge $h^\vee$ in P_σ, we make its dual edge h directed so that we see the black end of $h^\vee$ on our right-hand side when we cross $h^\vee$ along h in the given direction. Let Q_σ denote the resulting quiver. The set of vertices of the quiver Q_σ is $\widehat{I}$, which is identified with $\mathbb{Z}/N\mathbb{Z}$. The set of edges of the quiver Q_σ is given by

$$H := \Bigl(\coprod_{i\in\widetilde{I}} h_i^+\Bigr) \sqcup \Bigl(\coprod_{i\in\widetilde{I}} h_i^-\Bigr) \sqcup \Bigl(\coprod_{k\in\widehat{I}_r} r_k\Bigr).$$

Here h_i^+ (resp. h_i^-) is an edge from $i-\frac{1}{2}$ to $i+\frac{1}{2}$ (resp. from $i+\frac{1}{2}$ to $i-\frac{1}{2}$), and r_k is an edge from k to itself.

For each vertex q of P_σ, let ω_q be the potential[3] which is the composition of all arrows in Q_σ corresponding to edges in P_σ with q as their ends. We define

$$\omega_\sigma := \sum_{q\text{ black}} \omega_q - \sum_{q\text{ white}} \omega_q .$$

The relations of the Jacobian algebra are as follows:

[3]A potential of a quiver Q is an element in $\mathbb{C}Q/[\mathbb{C}Q,\mathbb{C}Q]$, i.e., a linear combination of equivalence classes of cyclic paths in Q, where two paths are equivalent if they coincide after a cyclic rotation.

- For $i \in \widetilde{I}$ such that $i-\frac{1}{2}, i+\frac{1}{2} \in \widehat{I}_r$,

$$h_i^+ \circ r_{i-\frac{1}{2}} = r_{i+\frac{1}{2}} \circ h_i^+ \quad \text{and} \quad r_{i-\frac{1}{2}} \circ h_i^- = h_i^- \circ r_{i+\frac{1}{2}}.$$

- For $i \in \widetilde{I}$ such that $i-\frac{1}{2} \in \widehat{I}_r$ and $i+\frac{1}{2} \notin \widehat{I}_r$,

$$h_i^+ \circ r_{i-\frac{1}{2}} = h_{i+1}^- \circ h_{i+1}^+ \circ h_i^+ \quad \text{and} \quad r_{i-\frac{1}{2}} \circ h_i^- = h_i^- \circ h_{i+1}^- \circ h_{i+1}^+.$$

- For $i \in \widetilde{I}$ such that $i-\frac{1}{2} \notin \widehat{I}_r$ and $i+\frac{1}{2} \in \widehat{I}_r$,

$$h_i^+ \circ h_{i-1}^+ \circ h_{i-1}^- = r_{i+\frac{1}{2}} \circ h_i^+ \quad \text{and} \quad h_{i-1}^+ \circ h_{i-1}^- \circ h_i^- = h_i^- \circ r_{i+\frac{1}{2}}.$$

- For $i \in \widetilde{I}$ such that $i-\frac{1}{2}, i+\frac{1}{2} \notin \widehat{I}_r$,

$$h_i^+ \circ h_{i-1}^+ \circ h_{i-1}^- = h_{i+1}^- \circ h_{i+1}^+ \circ h_i^+ \quad \text{and} \quad h_{i-1}^+ \circ h_{i-1}^- \circ h_i^- = h_i^- \circ h_{i+1}^- \circ h_{i+1}^+.$$

- For $k \in \widehat{I}_r$,

$$h_{i-\frac{1}{2}}^+ \circ h_{i-\frac{1}{2}}^- = h_{i+\frac{1}{2}}^- \circ h_{i+\frac{1}{2}}^+.$$

2B. ***NCCR and derived equivalence.*** Let $\pi : \mathcal{Y}_\sigma \to \mathcal{X}$ be the crepant resolution corresponding to σ.

Theorem 2.1 [Nagao 2012, Theorems 1.15 and 1.20].

$$D^b(\operatorname{mod} J_\sigma) \simeq D^b(\operatorname{Coh} \mathcal{Y}_\sigma).$$

The equivalence is given by an explicit tilting vector bundle which is a direct sum of line bundles [Nagao 2012, Theorem 1.10]. In particular, the following map is compatible with the derived equivalence

$$\begin{aligned} H^0(Y_\sigma, \mathbb{Z}) \oplus H^2(Y_\sigma, \mathbb{Z}) &\to \mathbb{Z}^I, \\ [\mathrm{pt}] &\mapsto \delta, \\ [C_i] &\mapsto \alpha_i, \end{aligned}$$

where α_i is the i-th fundamental vector and $\delta := \alpha_0 + \alpha_1 + \cdots + \alpha_{N-1}$.

2C. ***Mutation and derived equivalence.*** The Derksen–Weyman–Zelevinsky mutation [Derksen et al. 2010] of a quiver with a potential induces a derived equivalence of the derived categories of Ginzburg's differential graded algebras [Keller and Yang 2011]. Moreover, the relation between the module categories of Jacobian algebras has a description in terms of torsion pair and tilting, which plays a crucial role for the wall-crossing formulae [Kontsevich and Soibelman 2008; Nagao 2013]. In this paper, we cannot apply [Derksen et al. 2010; Keller and Yang 2011] since we have loops and oriented 2-cycles in the quiver. In this subsection, we see derived equivalences and descriptions of module categories using the explicit computations given in [Nagao 2012, §3].

Let k be an edge of the partition σ which is a diagonal of a parallelogram. Note that such a k corresponds to a vertex without loops. Let σ' denote the partition which is obtained by a "flip" of the edge k.

Let P_i be the indecomposable projective J_σ-module associated to a vertex i. Note that, as a vector space, P_i is the space of linear combinations of paths ending at the vertex i. We define

$$P'_k := \operatorname{coker}(P_k \to P_{k-1} \oplus P_{k+1}),$$

and put $P'_i = P_i$ for $i \neq k$. Here the map $P_k \to P_{k\pm 1}$ above is induced by the arrow from k to $k \pm 1$.

Theorem 2.2 [Nagao 2012, Proposition 3.1].

(1) $$\operatorname{End}(\oplus P'_i)^{\mathrm{op}} \simeq J_{\sigma'}.$$

(2) *The map*

$$\Phi_k := \mathbf{R}\operatorname{Hom}(\oplus P'_i, \bullet) : D^b(\operatorname{mod} J_\sigma) \to D^b(\operatorname{mod} J_{\sigma'})$$

is an equivalence.

For a J_σ-module $V = \bigoplus_{i \in \hat{I}} V_i$, we have

$$\big(H^j_{\operatorname{mod} J_{\sigma'}}(\Phi_k(V))\big)_i = \begin{cases} V_i & i \neq k,\ j = 0, \\ \ker(V_{k-1} \oplus V_{k+1} \to V_k) & i = k,\ j = 0, \\ \operatorname{coker}(V_{k-1} \oplus V_{k+1} \to V_k) & i = k,\ j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The simple reflection is compatible with the derived equivalence for dimension vectors.

By the description above, we have

$$\begin{aligned} \operatorname{mod} J_\sigma \cap \Phi_k^{-1}(\operatorname{mod} J_{\sigma'}) &= \{V \in \operatorname{mod} J_\sigma \mid \operatorname{coker}(V_{k-1} \oplus V_{k+1} \to V_k) = 0\} \\ &= \{V \in \operatorname{mod} J_\sigma \mid \operatorname{Hom}(V, s_k) = 0\} \\ &=: (\operatorname{mod} J_\sigma)^k, \\ \operatorname{mod} J_\sigma \cap \Phi_k^{-1}(\operatorname{mod} J_{\sigma'})[1] &= \{V \in \operatorname{mod} J_\sigma \mid V_i = 0\ (i \neq k)\} \\ &=: \mathcal{S}_k. \end{aligned}$$

In other words, $((\mathrm{mod}\, J_\sigma)^k, \mathcal{S}_k)$ is a torsion pair of $\mathrm{mod}\, J_\sigma$, and $\Phi_k^{-1}(\mathrm{mod}\, J_{\sigma'})$ is obtained from $\mathrm{mod}\, J_\sigma$ by tilting with respect to this torsion pair (see [Nagao 2013, §3.1]). Then we have

$$\begin{aligned}\mathrm{mod}\, J_{\sigma'} \cap \Phi_k(\mathrm{mod}\, J_\sigma) &= \{V \in \mathrm{mod}\, J_{\sigma'} \mid \mathrm{Hom}(s'_k, V) = 0\}\\ &=: (\mathrm{mod}\, J_{\sigma'})_k.\end{aligned}$$

In summary, we have the following:

Proposition 2.3. *The equivalence Φ_k induces an equivalence of* $(\mathrm{mod}\, J_\sigma)^k$ *and* $(\mathrm{mod}\, J_{\sigma'})_k$.

Nagao [2012, proof of Proposition 3.1] gave the isomorphism in Theorem 2.2(1) explicitly. For $V \in \mathrm{mod}\, J_\sigma \cap \Phi_k^{-1}(\mathrm{mod}\, J_{\sigma'})$, the map

$$\bigl(H^0_{\mathrm{mod}\, J_{\sigma'}}(\Phi_k(V))\bigr)_{k-1} \to \bigl(H^0_{\mathrm{mod}\, J_{\sigma'}}(\Phi_k(V))\bigr)_k$$

is induced by the morphism

$$R_{k-1} \oplus R_{k-1,k+1} : V_{k-1} \to V_{k-1} \oplus V_{k+1},$$

where

$$R_{k-1} := \begin{cases} r_{k-1} & k-1 \in \widehat{I}_r,\\ h^+_{k-\frac{3}{2}} \circ h^-_{k-\frac{3}{2}} & k-1 \notin \widehat{I}_r,\end{cases}$$

and

$$R_{k-1,k+1} := h^-_{k+\frac{1}{2}} \circ h^+_{k+\frac{1}{2}} \circ h^+_{k-\frac{1}{2}}.$$

2D. *Cut and mutation.* Let (Q, W) be a quiver with potential. To each subset $C \subset Q_1$, we associate a grading g_C on Q by

$$g_C(a) = \begin{cases} 1 & a \in C,\\ 0 & a \in C.\end{cases}$$

A subset $C \subset Q_1$ is called a cut if W is homogeneous of degree 1 with respect to g_C. Denote by Q_C the subquiver of Q with vertex set Q_0 and arrow set $Q_1 \backslash C$. We define the truncated Jacobian algebra by

$$J(Q, W)_C := J(Q, W)/\langle C\rangle.$$

Let k be a vertex of Q_σ without loops and C be a cut of (Q_σ, w_σ) such that $g_C\bigl(h^+_{k+\frac{1}{2}}\bigr) = 1$[4]. We define a cut C' of $(Q_{\sigma'}, w_{\sigma'})$ by the following conditions:

[4]We can construct a cut of (Q_σ, w_σ) as follows: First, by coupling h_i^+ and h_i^- for each i, we group the arrows in Q_σ into $N + |\widehat{I}_r|$ groups. Note that $N + |\widehat{I}_r|$ is even. These groups have the natural cyclic order and we label each of them as odd or even. Choose (any) one arrow from each odd (or even) labeled group; then we get a cut.

- $g_{C'}\left(h^+_{k-\frac{1}{2}}\right) = 1$.
- $g_{C'}(h^\pm_i) = g_C(h^\pm_i)$ if $i \neq k-\frac{1}{2}, k+\frac{1}{2}$.

Proposition 2.4 [Nagao 2011c, Proposition 4.12]. *The equivalence Φ_k induces an equivalence of* $(\operatorname{mod} J_{\sigma,C})_k$ *and* $(\operatorname{mod} J_{\sigma',C'})^k$.

Proof. It is enough to show that if $h^+_{k+\frac{1}{2}}$ vanishes on V, then $h^+_{k-\frac{1}{2}}$ vanished on $\Phi_k(V)$.

Since $g_C\left(h^\pm_{k-\frac{1}{2}}\right) = 0$, we have

- $g_C(r_{k-1}) = 1$ if $k-1 \in \widehat{I}_r$, and
- $g_C\left(h^+_{k-\frac{3}{2}}\right) = 1$ or $g_C\left(h^-_{k-\frac{3}{2}}\right) = 1$ if $k-1 \notin \widehat{I}_r$,

and so R_{k-1} vanishes. Since $g_C\left(h^+_{k+\frac{1}{2}}\right) = 1$, we see that $R_{k-1,k+1}$ vanishes. □

3. Motivic Donaldson–Thomas invariants

3A. ***Motives.*** We are working in a version of the ring of motivic weights: let $\mathcal{M}_{\mathbb{C}}$ denote the K-group of the category of effective Chow motives over $\mathbb{C}$, extended by $\mathbb{L}^{-\frac{1}{2}}$, where $\mathbb{L}$ is the Lefschetz motive. It has a natural structure of a λ-ring [Getzler 1996; Heinloth 2007], with σ-operations defined by $\sigma_n([X]) = [X^n/S_n]$ and $\sigma_n(\mathbb{L}^{\frac{1}{2}}) = \mathbb{L}^{\frac{n}{2}}$. We put

$$\widetilde{\mathcal{M}}_{\mathbb{C}} = \mathcal{M}_{\mathbb{C}}[\![\mathbb{L}^{-1}]\!],$$

which is also a λ-ring. Note that in this latter ring the elements $(1-\mathbb{L}^n)$, and therefore the motives of general linear groups, are invertible. The rings $\mathcal{M}_{\mathbb{C}} \subset \widetilde{\mathcal{M}}_{\mathbb{C}}$ sit in larger rings $\mathcal{M}^{\hat{\mu}}_{\mathbb{C}} \subset \widetilde{\mathcal{M}}^{\hat{\mu}}_{\mathbb{C}}$ of equivariant motives, where $\hat{\mu}$ is the group of all roots of unity [Looijenga 2002].

Let $f : X \to \mathbb{C}$ be a regular function on a smooth variety X. Using arc spaces, Denef and Loeser [2001; Looijenga 2002] defined the motivic nearby cycle $[\psi_f] \in \mathcal{M}^{\hat{\mu}}_{\mathbb{C}}$ and the motivic vanishing cycle

$$[\varphi_f] := [\psi_f] - [f^{-1}(0)] \in \mathcal{M}^{\hat{\mu}}_{\mathbb{C}}$$

of f. Note that if $f = 0$, then $[\varphi_0] = -[X]$.

Theorem 3.1 [Behrend et al. 2013, Proposition 1.11]. *Let $f : X \to \mathbb{C}$ be a regular function on a smooth variety X. Assume that X admits a $\mathbb{C}^*$-action such that f is $\mathbb{C}^*$-equivariant, i.e., $f(tx) = tf(x)$ for $t \in \mathbb{C}^*$, $x \in X$, and such that there exist limits* $\lim_{t\to 0} tx$ *for all $x \in X$. Then*

$$[\varphi_f] = [f^{-1}(1)] - [f^{-1}(0)] \in \mathcal{M}_{\mathbb{C}} \subset \mathcal{M}^{\hat{\mu}}_{\mathbb{C}}.$$

Following [Behrend et al. 2013], we define the *virtual motive* of $\mathrm{crit}(f)$ to be

$$[\mathrm{crit}(f)]_{\mathrm{vir}} := -(-\mathbb{L}^{\frac{1}{2}})^{-\dim X}[\varphi_f] \in \mathcal{M}_{\mathbb{C}}^{\hat{\mu}}.$$

For a smooth variety X, we put

$$[X]_{\mathrm{vir}} := [\mathrm{crit}(0_X)]_{\mathrm{vir}} = (-\mathbb{L}^{\frac{1}{2}})^{-\dim X} \cdot [X].$$

3B. ***Quivers and moduli spaces.*** Let Q be a quiver, with vertex set Q_0 and edge set Q_1. For an arrow $a \in Q_1$, we denote by $s(a) \in Q_0$ and $t(a) \in Q_0$ the vertices at which a starts and ends, respectively. We define the Euler–Ringel form χ on $\mathbb{Z}^{Q_0}$ by the rule

$$\chi(\alpha,\beta) = \sum_{i \in Q_0} \alpha_i \beta_i - \sum_{a \in Q_1} \alpha_{s(a)}\beta_{t(a)}, \quad \alpha,\beta \in \mathbb{Z}^{Q_0}.$$

Given a Q-representation M, we define its dimension vector $\underline{\dim}\, M \in \mathbb{N}^{Q_0}$ by $\underline{\dim}\, M = (\dim M_i)_{i \in Q_0}$. Let $\alpha \in \mathbb{N}^{Q_0}$ be a dimension vector and let $V_i = \mathbb{C}^{\alpha_i}$ for $i \in Q_0$. We define

$$R(Q,\alpha) := \bigoplus_{a \in Q_1} \mathrm{Hom}(V_{s(a)}, V_{t(a)})$$

and

$$G_\alpha := \prod_{i \in Q_0} \mathrm{GL}(V_i).$$

Note that G_α naturally acts on $R(Q,\alpha)$, and the quotient stack

$$\mathfrak{M}(Q,\alpha) := [R(Q,\alpha)/G_\alpha]$$

gives the moduli stack of representations of Q with dimension vector α.

Let W be a potential on Q, i.e., a finite linear combination of cyclic paths in Q. Denote by $J = J_{Q,W}$ the Jacobian algebra, i.e., the quotient of the path algebra $\mathbb{C}Q$ by the two-sided ideal generated by formal partial derivatives of the potential W. Let

$$f_\alpha : R(Q,\alpha) \to \mathbb{C}$$

be the G_α-invariant function defined by taking the trace of the map associated to the potential W. As it is now well-known [Segal 2008, Proposition 3.8], a point in the critical locus $\mathrm{crit}(f_\alpha)$ corresponds to a J-module. The quotient stack

$$\mathfrak{M}(J,\alpha) := [\mathrm{crit}(f_\alpha)/G_\alpha]$$

gives the moduli stack of J-modules with dimension vector α.

Definition 3.2. A *central charge* is a group homomorphism $Z : \mathbb{Z}^{Q_0} \to \mathbb{C}$ such that

$$Z(\alpha) \in \mathbb{H}_+ = \{re^{i\pi\varphi} \mid r > 0,\ 0 < \varphi \le 1\}$$

for any $\alpha \in \mathbb{N}^{Q_0} \backslash \{0\}$. Given $\alpha \in \mathbb{N}^{Q_0} \backslash \{0\}$, the number $\varphi(\alpha) = \varphi \in (0, 1]$ such that $Z(\alpha) = re^{i\pi\varphi}$, for some $r > 0$, is called the phase of α.

Definition 3.3. For any nonzero Q-representation or J-module V, we define $\varphi(V) = \varphi(\underline{\dim}\, V)$. A Q-representation (resp. J-module) V is said to be Z-stable if for any proper nonzero Q-subrepresentation (resp. J-submodule) $U \subset V$ we have

$$\varphi(U) < \varphi(V),$$

and Z-semistable if for all such proper subrepresentations (resp. submodules) we have the weaker condition

$$\varphi(U) \le \varphi(V).$$

Definition 3.4. Given $\zeta \in \mathbb{R}^{Q_0}$, define the central charge $Z : \mathbb{Z}^{Q_0} \to \mathbb{C}$ by the rule

$$Z(\alpha) = -\zeta \cdot \alpha + i|\alpha|,$$

where $|\alpha| = \sum_{i \in Q_0} \alpha_i$. We call a Q-representation or J-module ζ-(semi)stable if it is Z-(semi)stable.

Remark 3.5. Let the central charge Z be as in Definition 3.4. Define the slope function $\mu : \mathbb{N}^{Q_0} \backslash \{0\} \to \mathbb{R}$ by $\mu(\alpha) = \zeta \cdot \alpha / |\alpha|$. If $l \subset \mathbb{H} = \mathbb{H}_+ \cup \{0\}$ is a ray such that $Z(\alpha) \in l$, then $l = \mathbb{R}_{\ge 0}(-\mu(\alpha), 1)$. This implies that $\varphi(\alpha) < \varphi(\beta)$ if and only if $\mu(\alpha) < \mu(\beta)$.

We say that $\zeta \in \mathbb{R}^{Q_0}$ is α-generic if for any $0 < \beta < \alpha$ we have $\varphi(\beta) \ne \varphi(\alpha)$. This condition implies that any ζ-semistable Q-representation or J-module is automatically ζ-stable.

Let $R_\zeta(Q, \alpha)$ denote the open subset of $R(Q, \alpha)$ consisting of ζ-semistable representations. Let $f_{\zeta,\alpha}$ denote the restriction of f_α to $R_\zeta(Q, \alpha)$. The quotient stacks

$$\mathfrak{M}_\zeta(Q, \alpha) := [R_\zeta(Q, \alpha)/G_\alpha] \quad \text{and} \quad \mathfrak{M}_\zeta(J, \alpha) := [\operatorname{crit}(f_{\zeta,\alpha})/G_\alpha] \tag{3-1}$$

give the moduli stacks of ζ-semistable Q-representations and J-modules with dimension vector α.

3C. ***Motivic DT invariants.*** Let (Q, W) be a quiver with a potential and let $J = J_{Q,W}$ be its Jacobian algebra. Recall that the degeneracy locus of the function $f_\alpha : R(Q, \alpha) \to \mathbb{C}$ defines the locus of J-modules, so that the quotient stack

$$\mathfrak{M}(J, \alpha) := [\operatorname{crit}(f_\alpha)/G_\alpha]$$

is the stack of J-modules with dimension vector α. We define motivic Donaldson–Thomas invariants by

$$[\mathfrak{M}(J,\alpha)]_{\mathrm{vir}} := \frac{[\mathrm{crit}(f_\alpha)]_{\mathrm{vir}}}{[G_\alpha]_{\mathrm{vir}}}.$$

For a stability parameter ζ, we define

$$[\mathfrak{M}_\zeta(J,\alpha)]_{\mathrm{vir}} = \frac{[\mathrm{crit}(f_{\zeta,\alpha})]_{\mathrm{vir}}}{[G_\alpha]_{\mathrm{vir}}}, \tag{3-2}$$

where, as before, $f_{\zeta,\alpha}$ denotes the restriction of $f_\alpha : R(Q,\alpha) \to \mathbb{C}$ to $R_\zeta(Q,\alpha)$.

3D. ***Generating series of motivic DT invariants.*** Let (Q,W) be a quiver with a potential admitting a cut, and let $J = J_{Q,W}$ be its Jacobian algebra.

Definition 3.6. We define the generating series of the motivic Donaldson–Thomas invariants of (Q,W) by

$$A_U(y) = \sum_{\alpha\in\mathbb{N}^{Q_0}} [\mathfrak{M}(J,\alpha)]_{\mathrm{vir}} \cdot y^\alpha = \sum_{\alpha\in\mathbb{N}^{Q_0}} \frac{[\mathrm{crit}(f_\alpha)]_{\mathrm{vir}}}{[G_\alpha]_{\mathrm{vir}}} \cdot y^\alpha \in \mathcal{T}_Q,$$

where the subscript refers to the fact that we think of this series as the universal series.

Given a cut C of (Q,W), we define a new quiver $Q_C = (Q_0, Q_1\backslash C)$. Let J_C be the quotient of $\mathbb{C}Q_C$ by the ideal

$$(\partial_C W) = (\partial W/\partial a, a \in C).$$

Proposition 3.7 [Morrison et al. 2012, Proposition 1.14]. *If (Q,W) admits a cut C, then*

$$A_U(y) = \sum_{\alpha\in\mathbb{N}^{Q_0}} (-\mathbb{L}^{\frac{1}{2}})^{\chi(\alpha,\alpha)+2d_I(\alpha)} \frac{[R(J_C,\alpha)]}{[G_\alpha]} y^\alpha,$$

where $d_C(\alpha) = \sum_{(a:i\to j)\in C} \alpha_i\alpha_j$ for any $\alpha \in \mathbb{Z}^{Q_0}$.

The quiver with potential (Q_σ, w_σ) introduced in Section 2 admits a cut (see Section 2D), and Proposition 3.7 can be applied. In the next section we use this to compute the universal series in a specific case.

4. The universal DT series: special case

Throughout this section we fix σ to be the unique partition defined such that

$$\widehat{I}_r = \{0,1,2,3,\dots,N'-1\},$$

in other words, the partition such that the quiver with potential (Q_σ, w_σ) has loops at the first N' vertices only.

The aim of this section is to prove Theorem 0.1 for this quiver with potential.

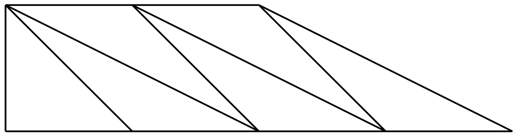


Figure 6. The case $N = 6$ and $N' = 2$.

We define three fixed subsets of the vertices:

$$I_1 := \{0, 1, \dots, N'-1\} \subset \mathbb{Z}/N,$$
$$I_2 := \{N', N'+2, N'+4, \dots, N-2\} \subset \mathbb{Z}/N,$$
$$I_3 := \{N'+1, N'+3, N'+5, \dots, N-1\} \subset \mathbb{Z}/N.$$

Then there exists a cut C given by the collection of arrows

$$C = \{h_i^- \mid i - \tfrac{1}{2} \notin I_2\}.$$

By Proposition 3.7, the universal DT series $A_U^\sigma(y) = \sum_{\alpha \in \mathbb{N}^Q} A_\alpha y^\alpha$ has coefficients given by

$$A_\alpha = (-\mathbb{L}^{\frac{1}{2}})^{\chi(\alpha,\alpha)+2d_C(\alpha)} \frac{[R(J_{\sigma,C}, \alpha)]}{[G_\alpha]} y^\alpha,$$

where $d_C(\alpha) = \sum_{(a:i\to j)\in C} \alpha_i \alpha_j$. To begin, we find a simple expression for the term $\chi(\alpha,\alpha) + 2d_C(\alpha)$ in the exponent. We know by definition that

$$\chi(\alpha,\alpha) = \sum_{i\in I_1\cup I_2\cup I_3} \alpha_i^2 - \sum_{i\in I_1} \alpha_i^2 - \sum_{i\in I_1\cup I_2\cup I_3} \alpha_i\alpha_{i+1} - \sum_{i\in I_1\cup I_2\cup I_3} \alpha_{i+1}\alpha_i,$$
$$d_I(\alpha) = \sum_{i\in I_1} \alpha_i\alpha_{i+1} + \sum_{i\in I_3} \alpha_{i+1}\alpha_i,$$

so it follows that

$$\chi(\alpha,\alpha) + 2d_C(\alpha) = \sum_{i\in I_2\cup I_3} \alpha_i^2 - 2\cdot \sum_{i\in I_2} \alpha_i\alpha_{i+1} = \sum_{i\in I_2} (\alpha_{i+1} - \alpha_i)^2.$$

Our next goal is to factorize $A_U^\sigma(y)$ into two simpler series. We proceed by analyzing the motivic classes $[R(J_{\sigma,C}, \alpha)]$.

Given a dimension vector $\alpha \in \mathbb{N}^{Q_0}$ and a representation of a $J_{\sigma,C}$-module

$$V = \bigoplus_{i\in I_1\cup I_2\cup I_3} V_i,$$

we focus on the specific element

$$H := h^+_{\frac{1}{2}} + h^+_{\frac{3}{2}} + \cdots + h^+_{N-\frac{1}{2}} \in \bigoplus_{i\in I_1\cup I_2\cup I_3} \mathrm{Hom}(V_i, V_{i+1}).$$

This map H acts as an endomorphism of the vector space V. Given any such linear map

$$H : V \to V,$$

there exists a unique splitting $V = V^I \oplus V^N$ with maps

$$\begin{aligned} H^I : V^I \to V^I & \quad \text{invertible}, \\ H^N : V^N \to V^N & \quad \text{nilpotent}, \end{aligned}$$

so that

$$H = H^I \oplus H^N.$$

Moreover, in our case the above splitting respects the grading by $i \in I_1 \cup I_2 \cup I_3$. To be explicit, we have that

$$V^I = \bigoplus_{i \in I_1 \cup I_2 \cup I_3} V_i^I,$$

where $V_i^I := V_i \cap V^I$ (similarly $V^N = \bigoplus_{i \in I_1 \cup I_2 \cup I_3} V_i^N$ with $V_i^N := V_i \cap V^N$). One immediate consequence of this is that

$$\dim(V_i^I) = \dim(V_{i+1}^I) \quad \text{for all } i \in I_1 \cup I_2 \cup I_3;$$

indeed, this is clear since the block form of H^I demands that it map V_i^I to V_{i+1}^I via an isomorphism. We are now ready to decompose the computation of $A_U^\sigma(y)$ into two simpler subproblems.

Definition 4.1 (invertible series). We define

$$R^I(a) := \{r \in R(J_{\sigma,C}, \alpha) \mid H \text{ is invertible}, \alpha_i = a \text{ for all } i\}$$

and the series

$$I^\sigma(x) := \sum_{a \geq 0} \frac{[R^I(a)]}{[\mathrm{GL}(a)]^N} x^a.$$

Definition 4.2 (nilpotent series). We define

$$R^N(\alpha) := \{r \in R(J_{\sigma,C}, \alpha) \mid H \text{ is nilpotent}\}$$

and the series

$$N^\sigma(y) := \sum_{\alpha \in \mathbb{N}^{Q_0}} (-\mathbb{L}^{\frac{1}{2}})^{\sum_{i \in I_2} (\alpha_{i+1} - \alpha_i)^2} \frac{[R^N(\alpha)]}{[G_\alpha]} y^\alpha.$$

The following lemma shows that the series $A_U^\sigma(y)$ factorizes into the product of the two series just defined:

Lemma 4.3. *Let $x = y_0 \cdots y_{N-1}$; then, in the notation above, we have*

$$A_U^\sigma(y) = I^\sigma(x) \cdot N^\sigma(y).$$

Proof. This formula follows directly from a stratification of the variety $R(J_{\sigma,C}, \alpha)$ by the dimension of V_i^I.

Fix $\alpha \in \mathbb{N}^{Q_0}$; we stratify $R(J_{\sigma,C}, \alpha)$ by $\dim(V_i^I) = a$. Let

$$\underline{a} := (a, a, \dots, a) \in \mathbb{N}^{Q_0},$$

and let

$$\alpha' \text{ be such that } \alpha = \underline{a} + \alpha' \in \mathbb{N}^{Q_0}.$$

There is a Zariski locally trivial fibration

$$\begin{array}{ccc} R^I(a) \times R^N(\alpha') & \longrightarrow & \{r \in R(J_{\sigma,C}, \alpha) \mid \dim(V_i^I) = a \text{ for } H \in r\} \\ & & \downarrow \\ & & \mathcal{M}(a, \alpha) \end{array}$$

Here $\mathcal{M}(a, \alpha)$ is the space parametrizing splittings $V_i = V_i^I \oplus V_i^N$. To see this, one checks that the arrows r_i, $h^-_{i+\frac{1}{2}}$ in the representation also preserve the splitting, so the entire representation splits into $V^I \oplus V^N$. This follows easily from the relations and some basic linear algebra.

Splittings of the vector space $V_i = V_i^I \oplus V_i^N$ are parametrized by

$$\mathrm{GL}(\alpha_i)/(\mathrm{GL}(a) \times \mathrm{GL}(\alpha_i')),$$

and hence the motivic class of the base is

$$[\mathcal{M}(a, \alpha)] = \frac{[G_\alpha]}{[\mathrm{GL}(a)]^N \cdot [G_{\alpha'}]}.$$

Summing over each stratum with $\dim(V_i^I) = a$, we get

$$[R(J_{\sigma,C}, \alpha)] = [G_\alpha] \cdot \sum_{a=0}^{\min_i\{\alpha_i\}} \frac{[R^I(a)]}{[\mathrm{GL}(a)]^N} \cdot \frac{[R^N(\alpha')]}{[G_{\alpha'}]}.$$

Multiplying both sides of this expression by $(-\mathbb{L}^{\frac{1}{2}})^{\sum_{i \in I_2} (\alpha_{i+1} - \alpha_i)^2} y^\alpha$ and summing gives

$$\begin{aligned} &A_U^\sigma(y) \\ &\quad = \Bigg(\sum_{a \ge 0} \frac{[R^I(a)]}{[\mathrm{GL}(a)]^N} \prod_{i=0}^{N-1} y_i^a \Bigg) \cdot \Bigg(\sum_{\alpha' \in \mathbb{N}^{Q_0}} (-\mathbb{L}^{\frac{1}{2}})^{\sum_{i \in I_2} (\alpha'_{i+1} - \alpha'_i)^2} \frac{[R^N(\alpha')]}{[G_{\alpha'}]} y^{\alpha'} \Bigg), \end{aligned}$$

proving the result. $\square$

In the next two sections we compute formulae for $I^\sigma(x)$ and $N^\sigma(y)$.

4A. ***Step I: the invertible case $I^\sigma(x)$.***

Proposition 4.4. *We have*

$$I^\sigma(x) = \operatorname{Exp}\left(\mathbb{L}\frac{x}{1-x}\right).$$

Proof. A $J_{\sigma,C}$-module $r \in R(J_{\sigma,C}, \alpha)$ is given by a vector space

$$V = \bigoplus_{i \in I_1 \cup I_2 \cup I_3} V_i$$

of dimension $\alpha \in \mathbb{N}^{Q_0}$ and a collection of linear maps

$$\begin{aligned} r_i &: V_i \to V_i && \text{for } i \in I_1, \\ h^-_{i+\frac{1}{2}} &: V_{i+1} \to V_i && \text{for } i \in I_2, \\ h^+_{i+\frac{1}{2}} &: V_i \to V_{i+1} && \text{for } i \in I_1 \cup I_2 \cup I_3, \end{aligned}$$

satisfying the relations coming from cyclic differentiation of the potential

$$\begin{aligned} r_i h^+_{i-\frac{1}{2}} &= h^+_{i-\frac{1}{2}} r_{i-1} && \text{for } i \in [1, N'-1] \cap I_1, \\ r_0 h^+_{N-\frac{1}{2}} &= h^+_{N-\frac{1}{2}} h^+_{N-\frac{3}{2}} h^-_{N-\frac{3}{2}}, \\ h^-_{N'+\frac{1}{2}} h^+_{N'+\frac{1}{2}} h^+_{N'-\frac{1}{2}} &= h^+_{N'-\frac{1}{2}} r_{N'-1}, \\ h^-_{i+\frac{3}{2}} h^+_{i+\frac{3}{2}} h^+_{i+\frac{1}{2}} &= h^+_{i+\frac{1}{2}} h^+_{i-\frac{1}{2}} h^-_{i-\frac{1}{2}} && \text{for } i = [N'+1, N-3] \cap I_3. \end{aligned}$$

Assuming moreover that $r \in R^I(a)$, we have

$$h^+_{i+\frac{1}{2}} : V_i \to V_{i+1} \ \text{ is invertible for all } i \in I_1 \cup I_2 \cup I_3.$$

This allows us to express $R^I(a)$ as a $\prod_{i=1}^{N-1} \mathrm{GL}(V_i)$-torsor over a commuting variety

$$\begin{aligned} \pi &: R^I(a) \to C(a), \\ \bigl(r_i, h^+_{i+\frac{1}{2}}, h^-_{i+\frac{1}{2}}\bigr) &\mapsto \bigl(r_0, h^+_{N-\frac{1}{2}} h^+_{N-\frac{3}{2}} \cdots h^+_{\frac{3}{2}} h^+_{\frac{1}{2}}\bigr), \end{aligned}$$

where

$$C(a) = \{(A, B) \in \operatorname{End}(V_0) \times \mathrm{GL}(V_0) \mid AB = BA\}.$$

The free action of $\prod_{i=1}^{N-1} \mathrm{GL}(V_i)$ on $R^I(a)$ is given by

$$\begin{aligned} (g_1,\dots,g_{N-1}) : r_i &\mapsto g_i r_i g_i^{-1} && \text{for } i \in [1, N'-1], \\ h^+_{\frac{1}{2}} &\mapsto g_1 h^+_{\frac{1}{2}}, \\ h^+_{N-\frac{1}{2}} &\mapsto h^+_{N-\frac{1}{2}} g^{-1}_{N-1}, \\ h^+_{i+\frac{1}{2}} &\mapsto g_{i+1} h^+_{i+\frac{1}{2}} g_i^{-1} && \text{for } i \in [1, N-2], \\ h^-_{i+\frac{1}{2}} &\mapsto g_i h^-_{i+\frac{1}{2}} g^{-1}_{i+1} && \text{for } i \in I_2. \end{aligned}$$

As $\mathrm{GL}(a)$ is a special group [Chevalley et al. 1958], the torsor splits in the Zariski topology, so motivically we have

$$[R^I(a)] = [\mathrm{GL}(a)]^{N-1} \cdot [C(a)].$$

Thus

$$I^\sigma(x) = \sum_{a \geq 0} \frac{[C(a)]}{[\mathrm{GL}(a)]} x^a.$$

The generating series for the commuting variety is obtained in [Bryan and Morrison 2015], giving the result. □

4B. ***Step II: the nilpotent case $N^\sigma(y)$.*** This section is the final step in the calculation. Here we compute $N^\sigma(y)$ and obtain the formula of $A^\sigma_U(y)$.

We fix a dimension vector $\alpha \in \mathbb{N}^{Q_0}$. As before, a $J_{\sigma,C}$-module is given by a vector space

$$V = \bigoplus_{i \in I_1 \cup I_2 \cup I_3} V_i$$

of dimension α and a collection of linear maps

$$\begin{aligned} r_i &: V_i \to V_i && \text{for } i \in I_1, \\ h^-_{i+\frac{1}{2}} &: V_{i+1} \to V_i && \text{for } i \in I_2, \\ h^+_{i+\frac{1}{2}} &: V_i \to V_{i+1} && \text{for } i \in I_1 \cup I_2 \cup I_3, \end{aligned}$$

satisfying the relations of the potential (see Proposition 4.4). Throughout this section we insist that the map

$$H = h^+_{\frac{1}{2}} + h^+_{\frac{3}{2}} + \dots + h^+_{N-\frac{1}{2}} \in \bigoplus_{i \in I_1 \cup I_2 \cup I_3} \mathrm{Hom}(V_i, V_{i+1})$$

is nilpotent. In fact, $R^N(\alpha)$ is exactly the collection of all such representations (see Definition 4.2). In particular, if we let $|\alpha| := \dim(V)$ then we know that $H^{|\alpha|} = 0$. This gives a filtration of the vector space

$$V = V^{|\alpha|} \supset V^{|\alpha|-1} \supset \cdots \supset V^1 \supset V^0 = \{0\},$$

where

$$V^j = \{v \in V \mid H^j(v) = 0\}.$$

Moreover, the filtration respects the grading by $i \in I_1 \cup I_2 \cup I_3$, by which we mean that

$$V^j = \bigoplus_{i \in I_1 \cup I_2 \cup I_3} (V^j \cap V_i),$$

where V_i is the summand at the i-th vertex of the quiver. By considering the vector space V as a representation of the nilpotent matrix H, we can identify V with a $\mathbb{C}[x]$-module supported at the origin. Modules for a principal ideal domain have a simple structure. In particular, we have

$$V \cong \bigoplus_{j=1}^{d} (\mathbb{C}[x]/(x^j))^{\oplus b_j}$$

as a $\mathbb{C}[x]$-module. The next proposition provides a more refined version of this statement, where each factor in this decomposition is generated by a vector from a vector space V_i:

Proposition 4.5. *For each $i \in I_1 \cup I_2 \cup I_3$, there exists a collection of integers b^i_j such that*

$$V \cong \bigoplus_{i \in I_1 \cup I_2 \cup I_3} \bigoplus_{i=1}^{d} (\mathbb{C}[x]/(x^j))^{\oplus b^i_j},$$

where the factor $(\mathbb{C}[x]/(x^j))^{\oplus b^i_j}$ is generated as a $\mathbb{C}[x]$-module by vectors in V_i. Moreover, the numbers b^i_j are uniquely determined by the above conditions.

Proof. We will argue by induction on d, the largest integer such that $b_d \neq 0$. As such, we can assume that for each $j \leq d-1$ the factor $\mathbb{C}[x]/(x^j)$ is generated by a vector in some V_i. Now let $e_1, \ldots, e_{b_d}$ be a generating set for the factor $(\mathbb{C}[x]/(x^d))^{\oplus b_d}$, and define $W := \mathrm{span}\{e_1, \ldots, e_{b_d}\}$. We consider the projection operators

$$p_i : V \to V_i / V_i \cap V^{d-1},$$

and set $W_i := p_i(W)$ and $b^i_d = \dim W_i$. We claim that

$$p_0 \oplus \cdots \oplus p_{N-1} : W \to W_0 \oplus \cdots \oplus W_{N-1}$$

is an isomorphism. The map is clearly onto and an injection since any vector in the kernel must lie in V^{d-1}. Now, considering a lifting of the vector space $V_i \supset W_i' \twoheadrightarrow W_i \subset V_i / V_i \cap V^{d-1}$, we have that

$$W_i' \oplus H W_i' \oplus \cdots \oplus H^{d-1} W_i' \subset V$$

is a submodule of V isomorphic to $(\mathbb{C}[x]/(x^d))^{\oplus b_d^i}$. Summing over all i, we have that $(\mathbb{C}[x]/(x^d))^{\sum_i b_d^i}$ is a submodule of V, and hence it follows that $\sum_i \dim W_i = \sum_i b_d^i \leq b_d = \dim W$, and so for dimension reasons we get

$$V \cong \left(\bigoplus_{i=0}^{N-1} (\mathbb{C}[x]/(x^d))^{\oplus b_d^i} \right) \bigoplus \left(\bigoplus_{j=1}^{d-1} (\mathbb{C}[x]/(x^j))^{\oplus b_j} \right).$$

Here each factor $(\mathbb{C}[x]/(x^d))^{\oplus b_d^i}$ is generated by vectors in V_i, so by our inductive hypothesis the entire module is generated by vectors in V_i.

Finally we prove the uniqueness statement. Assume we have two distinct such decompositions

$$V \cong \bigoplus_{i=0}^{N-1} \bigoplus_{j=1}^{d} (\mathbb{C}[x]/(x^j))^{\oplus b_j^i} \cong \bigoplus_{i=0}^{N-1} \bigoplus_{j=1}^{d} (\mathbb{C}[x]/(x^j))^{\oplus c_j^i}.$$

By restricting to subrepresentations if necessary, we can assume that $b_d^i \neq c_d^i$ for some i. However in this case

$$b_d^i = \dim(\ker(H^d : V_i \to V_{i+d}) / V_i \cap V^{d-1}) = c_d^i$$

is a contradiction. This proves the last part of the lemma. □

Next we organize this data in the way most helpful to our cause:

Definition 4.6. Let $0 \leq a, b \leq N-1$. We define

$$|b-a| = \min\{r \in \{0, 1, \ldots, N-1\} \mid b = a + r \bmod N\}.$$

Intuitively, this is the distance from a to b in the cyclic direction $i \to i+1$ corresponding to the map H.

Definition 4.7. Suppose we have a decomposition of V as a $\mathbb{C}[x]$-module as in Proposition 4.5. Define $V^{a,b}$ to be the vector subspace corresponding to the summand

$$\bigoplus_{l \geq 1} (\mathbb{C}[x]/(x^{N(l-1)+|b-a|+1}))^{b^a_{N(l-1)+|b-a|+1}},$$

and relabel the integers

$$b_l^{a,b} := b^a_{N(l-1)+|b-a|+1}$$

to define partitions

$$\pi^{[a,b]} := \left(1^{b_1^{a,b}}\ 2^{b_2^{a,b}}\ 3^{b_3^{a,b}}\ \cdots\right).$$

Notice that the above definition depends on the choice of the decomposition in Proposition 4.5. However, all such vector spaces are isomorphic abstractly as $\mathbb{C}[x]$-modules. We can think of these vector spaces as being generated by the nilpotent vectors that start at the a-th vertex and are annihilated at the $(b+1)$-st vertex under the action of the map H.

The next lemma makes explicit how to recover the dimension vector of a representation from the datum of the N^2 partitions $\{\pi^{[a,b]} \mid 0 \le a, b \le N-1\}$.

Lemma 4.8. *Given a representation $r \in R^N(\alpha)$ so that the endomorphism H has type $\{\pi^{[a,b]}\}$, the dimension vector of the representation r is given by*

$$\alpha_i = \sum_{a,b} |\pi^{[a,b]}| - \sum_{\substack{a,b \\ i \notin [a,b]}} l(\pi^{[a,b]}),$$

where $|\pi^{[a,b]}|$ and $l(\pi^{[a,b]})$ are the size and length of the partition $\pi^{[a,b]}$.

Proof. This is clear since

$$V = \bigoplus_{a,b} V^{a,b}$$

and

$$\dim(V^{a,b} \cap V_i) = \begin{cases} |\pi^{[a,b]}| & \text{if } i \in [a,b], \\ |\pi^{[a,b]}| - l(\pi^{[a,b]}) & \text{if } i \notin [a,b]. \end{cases} \qquad \square$$

We can use this to give a simple reformulation of the term $\chi(\alpha,\alpha) + 2d_C(\alpha)$ appearing in the series N^σ:

Corollary 4.9. *We have*

$$\chi(\alpha,\alpha) + 2d_C(\alpha) = \sum_{i \in I_2} \left(\sum_{b \neq i} l(\pi^{[i+1,b]}) - \sum_{c \neq i+1} l(\pi^{[c,i]}) \right)^2.$$

Proof. In our initial analysis of these terms we saw that

$$\chi(\alpha,\alpha) + 2d_C(\alpha) = \sum_{i \in I_2} (\alpha_{i+1} - \alpha_i)^2,$$

and now by Lemma 4.8 we have

$$\alpha_{i+1} - \alpha_i = \sum_{b \neq i} l(\pi^{[i+1,b]}) - \sum_{c \neq i+1} l(\pi^{[c,i]}). \qquad \square$$

The above classification has been for the purpose of breaking the variety $R^N(\alpha)$ down into simpler pieces.

Definition 4.10. Given N^2 partitions $\{\pi^{[a,b]} \mid 0 \le a, b \le N-1\}$ and a dimension vector α as in Lemma 4.8, we define

$$R(\{\pi^{[a,b]}\}) = \{r \in R^N(\alpha) \mid H \text{ has type } \{\pi^{[a,b]}\}\}.$$

This provides a stratification of $R^N(\alpha)$ into strata where the normal form of H has a fixed type. We will proceed to compute the motivic classes of each of these strata.

A representation in $R(\{\pi^{[a,b]}\})$ is given explicitly by a vector space $V = \bigoplus_{i \in I_1 \cup I_2 \cup I_3} V_i$ and a collection of linear maps corresponding to the arrows r_i with $i \in I_1$, $h^-_{i+\frac{1}{2}}$ with $i \in I_2$ and $h^+_{i+\frac{1}{2}}$ with $i \in I_1 \cup I_2 \cup I_3$. In addition, the linear maps satisfy relations

$$\begin{aligned}
r_i h^+_{i-\frac{1}{2}} &= h^+_{i-\frac{1}{2}} r_{i-1} && \text{for } i \in [1, N'-1] \cap I_1, \\
r_0 h^+_{N-\frac{1}{2}} &= h^+_{N-\frac{1}{2}} h^+_{N-\frac{3}{2}} h^-_{N-\frac{3}{2}}, \\
h^-_{N'+\frac{1}{2}} h^+_{N'+\frac{1}{2}} h^+_{N'-\frac{1}{2}} &= h^+_{N'-\frac{1}{2}} r_{N'-1}, \\
h^-_{i+\frac{3}{2}} h^+_{i+\frac{3}{2}} h^+_{i+\frac{1}{2}} &= h^+_{i+\frac{1}{2}} h^+_{i-\frac{1}{2}} h^-_{i-\frac{1}{2}} && \text{for } i = [N'+1, N-3] \cap I_3.
\end{aligned}$$

and we require that the map

$$H = h^+_{\frac{1}{2}} + h^+_{\frac{3}{2}} + \cdots + h^+_{N-\frac{1}{2}} \in \bigoplus_{i \in I_1 \cup I_2 \cup I_3} \mathrm{Hom}(V_i, V_{i+1})$$

has a type given by the partitions $\{\pi^{[a,b]} \mid 0 \le a, b \le N-1\}$. The linear map H contains all the information of the maps $h^+_{i+\frac{1}{2}}$. For brevity, we make the following definition, packaging all the remaining linear maps into one.

Definition 4.11. Given a representation as above, we define the linear map

$$\begin{aligned}
L := r_0 + r_1 + \cdots + r_{N'-1} + h^-_{N'+\frac{1}{2}} + \cdots + h^-_{N-\frac{3}{2}} \\
\in \bigoplus_{i \in I_1} \mathrm{Hom}(V_i, V_i) \bigoplus_{i \in I_2} \mathrm{Hom}(V_{i+1}, V_i).
\end{aligned}$$

From now on, in order to compute the motivic class of $R(\{\pi^{[a,b]}\})$ we will work with a choice of coordinates. Let

$$v_l^{a,b}(k) \in V_a$$

be such that $v_l^{a,b}(k)$ generates the k-th summand of $\mathbb{C}[x]/(x^{N(l-1)+|b-a|+1})^{\oplus b_l^{a,b}}$ in the decomposition of Proposition 4.5. Then we have that

$$\mathcal{B} := \{H^p v_l^{a,b}(k) \mid 1 \le k \le b_l^{a,b},\ 0 \le a, b \le N-1,\ 0 \le p \le N(l-1)+|b-a|+1\}$$

forms a basis of V.

Definition 4.12. We define $H(\pi^{[a,b]})$ to be the matrix representation of the map H with respect to the basis $\mathcal{B}$. Also define

$$F(\{\pi^{[a,b]}\}) := \{L \mid (L, H(\pi^{[a,b]})) \in R(\{\pi^{[a,b]}\})\},$$
$$N(\{\pi^{[a,b]}\}) := \{H \mid H \text{ has type } \{\pi^{[a,b]}\}\}.$$

Then $R(\{\pi^{[a,b]}\})$ has a decomposition as a vector bundle:

Lemma 4.13. *$R(\{\pi^{[a,b]}\})$ has the structure of a vector bundle*

$$\begin{tikzcd} F(\{\pi^{[a,b]}\}) \arrow[r] & R(\{\pi^{[a,b]}\}) \arrow[d] \\ & N(\{\pi^{[a,b]}\}) \end{tikzcd}$$

In particular, we have that

$$[R(\{\pi^{[a,b]}\})] = [F(\{\pi^{[a,b]}\})] \cdot [N(\{\pi^{[a,b]}\})]$$

in the Grothendieck ring of varieties.

Proof. The projection map

$$p : R(\{\pi^{[a,b]}\}) \to N(\{\pi^{[a,b]}\}),$$
$$(L, H) \mapsto H$$

defines the bundle structure with zero section $H \mapsto (0, H)$. The fiber is the linear space of all such L. □

Here the base of the vector bundle is the space of all matrices of type $\{\pi^{[a,b]}\}$; these are all conjugate to $H(\pi^{[a,b]})$, and therefore we have a torsor

$$\pi : G_\alpha \to N(\pi^{[a,b]}),$$
$$P \mapsto PH(\pi^{[a,b]})P^{-1}.$$

This is a torsor for the group $S'(\{\pi^{[a,b]}\}) := \mathrm{Stab}_{G_\alpha}(H(\pi^{a,b}))$. This group is given as the group of units in an algebra:

Definition 4.14. We identify $S'(\{\pi^{[a,b]}\})$ with the group of multiplicative units in the algebra

$$S(\{\pi^{[a,b]}\}) := \left\{ N \in \prod_{i=0}^{N-1} \mathrm{End}(\alpha_i) \;\middle|\; NH(\pi^{[a,b]}) = H(\pi^{[a,b]})N \right\}.$$

Since $S'(\{\pi^{[a,b]}\})$ is the group of units of an algebra it is a special group [Chevalley et al. 1958], and so the above torsor splits in the Zariski topology. For completeness we include a short proof of this fact.

Lemma 4.15. *The group $S'(\{\pi^{[a,b]}\})$ is a special algebraic group.*

Proof. The ring of units is defined by a single polynomial equation and is open in A. Consequently the units $S'(\{\pi^{[a,b]}\})$ act generically freely on the vector space $S(\{\pi^{[a,b]}\})$, so by Proposition 3.13 of [Merkurjev 2013] the group has essential dimension zero. Then by Proposition 3.16 of [Merkurjev 2013] it is a special algebraic group. □

The next lemma gives a formula of the motivic class of the group $S'(\{\pi^{[a,b,]}\})$, and via the splitting of the above torsor we deduce a formula for the class of $N(\{\pi^{[a,b]}\})$. Before stating the lemma we create some notation:

Definition 4.16. We denote the dimensions of the linear spaces $F(\{\pi^{[a,b]}\})$ and $S(\{\pi^{[a,b]}\})$ by

$$T(\{\pi^{[a,b]}\}) := \dim F(\{\pi^{[a,b]}\}),$$
$$B(\{\pi^{[a,b]}\}) := \dim S(\{\pi^{[a,b]}\}).$$

Lemma 4.17. *We have*

$$[S'(\{\pi^{[a,b]}\})] = [S(\{\pi^{[a,b]}\})] \cdot \prod_{0\le a,b\le N-1} \frac{1}{f(\pi^{[a,b]})},$$

where

$$f(\pi^{[a,b]}) := \prod_{l\ge 1} \frac{[\mathrm{End}(b_l^{a,b})]}{[\mathrm{GL}(b_l^{a,b})]}.$$

As a consequence,

$$[R(\{\pi^{a,b}\})] = [G_\alpha] \cdot \mathbb{L}^{T(\{\pi^{[a,b]}\})-B(\{\pi^{[a,b]}\})} \cdot \prod_{0\le a,b\le N-1} f(\pi^{[a,b]}).$$

Proof. Let

$$W_l^{a,b} := \mathrm{span}_{\mathbb{C}}\{v_l^{a,b}(k) \mid 1 \le k \le b_l^{a,b}\}$$

be the span of the basis elements $v_l^{a,b}(k)$ for $1 \le k \le b_l^{a,b}$. We have both inclusion and projection

$$W_l^{a,b} \hookrightarrow V \twoheadrightarrow W_l^{a,b}.$$

This gives a map of algebras

$$\pi : S(\{\pi^{[a,b]}\}) \to \prod_{a,b,l} \mathrm{End}(W_l^{a,b}),$$
$$N \mapsto \bigoplus_{a,b,l} N|_{W_l^{a,b}}.$$

This splits as a trivial vector bundle, whose rank is the dimension of the total space minus the dimension of the base. Since we have that the group $S'(\{\pi^{[a,b]}\})$ is the group of units in $S(\{\pi^{[a,b]}\})$, we can identify $S'(\{\pi^{[a,b]}\})$ as the inverse image of the units on the right-hand side. This is a trivial vector bundle of rank equal to $\dim S(\{\pi^{[a,b]}\}) - \dim \prod_{a,b,l} \mathrm{End}(W_l^{a,b})$. We have an isomorphism of varieties

$$S'(\{\pi^{[a,b]}\}) \equiv \frac{S(\{\pi^{[a,b]}\})}{\prod_{a,b,l} \mathrm{End}(W_l^{a,b})} \times \prod_{a,b,l} \mathrm{GL}(W_l^{a,b}),$$

so motivically we have

$$[S'(\{\pi^{[a,b]}\})] = [S(\{\pi^{[a,b]}\})] \cdot \prod_{0\le a,b\le N-1} \frac{1}{f(\pi^{[a,b]})}.$$

In Lemma 4.13 we saw that

$$[R(\{\pi^{[a,b]}\})] = [F(\{\pi^{[a,b]}\})] \cdot [N(\{\pi^{[a,b]}\})].$$

Now we know that $N(\{\pi^{[a,b]}\})$ is a torsor for the group $S'(\{\pi^{[a,b]}\})$. We have just computed the motive of this group. So we can deduce that

$$\begin{aligned}[R(\{\pi^{[a,b]}\})] &= [F(\{\pi^{[a,b]}\})] \cdot \frac{[G_\alpha]}{[S'(\{\pi^{[a,b]}\})]} \\ &= [F(\{\pi^{[a,b]}\})] \cdot \frac{[G_\alpha]}{[S(\{\pi^{[a,b]}\})]} \cdot \prod_{0\le a,b\le N-1} f(\pi^{[a,b]}) \\ &= [G_\alpha] \cdot \mathbb{L}^{T(\{\pi^{[a,b]}\})-B(\{\pi^{[a,b]}\})} \cdot \prod_{0\le a,b\le N-1} f(\pi^{[a,b]}). \qquad \square\end{aligned}$$

The next proposition computes the difference $T(\{\pi^{[a,b]}\}) - B(\{\pi^{[a,b]}\})$. Its proof is found in the Appendix:

Proposition 4.18. *We have that $T(\{\pi^{[a,b]}\}) - B(\{\pi^{[a,b]}\})$ is equal to*

$$-\frac{1}{2}\sum_{i\in I_2}\Bigg(\sum_{b\neq i} l(\pi^{[i+1,b]}) - \sum_{c\neq i+1} l(\pi^{[c,i]})\Bigg)^2 - \frac{1}{2}\sum_{\substack{a\in I_3\\ b\notin I_2}}\sum_{i\ge 1}(b_i^{a,b})^2 - \frac{1}{2}\sum_{\substack{a\notin I_3\\ b\in I_2}}\sum_{i\ge 1}(b_i^{a,b})^2.$$

Proof. The proof is a linear algebra calculation. See the Appendix. $\square$

As a corollary, we deduce the formula for $N^\sigma(y)$:

Proposition 4.19. *Let*

$$\begin{aligned} S &= \{[a,b] \mid a \in I_3, b \notin I_2 \text{ or } a \notin I_3, b \in I_2\}, \\ y_{[a,b]} &= y_a \cdot y_{a+1} \cdots y_b, \\ y' &= y_0 \cdot y_1 \cdots y_{N-1}, \end{aligned}$$

then we have

$$N^\sigma(y) = \operatorname{Exp}\biggl(\frac{\mathbb{L}}{\mathbb{L}-1}\frac{1}{1-y'}\biggl(\sum_{[a,b]\notin S} y_{[a,b]} - \mathbb{L}^{-\frac{1}{2}} \sum_{[a,b]\in S} y_{[a,b]}\biggr)\biggr).$$

Proof. Recall our initial definition of $N^\sigma(y)$:

$$N^\sigma(y) = \sum_{\alpha\in\mathbb{N}^{Q_0}} (-\mathbb{L}^{\frac{1}{2}})^{\chi(\alpha,\alpha)+2d_C(\alpha)} \frac{[R^N(\alpha)]}{[G_\alpha]} y^\alpha.$$

In Proposition 4.5 we saw that it was possible to stratify each of the varieties $R^N(\alpha)$ by the type $\{\pi^{[a,b]}\}$ of the cycle H. This gives

$$N^\sigma(y) = \sum_{\alpha\in\mathbb{N}^{Q_0}} (-\mathbb{L}^{\frac{1}{2}})^{\chi(\alpha,\alpha)+2d_C(\alpha)} [G_\alpha]^{-1} \biggl(\sum_{\{\pi^{[a,b]}\}\vdash\alpha} [R(\{\pi^{[a,b]}\})]\biggr) y^\alpha.$$

The motivic class of $R(\{\pi^{[a,b]}\})$ was computed in Lemma 4.17. Now substituting this class into the above formula gives

$$\sum_{\alpha\in\mathbb{N}^{Q_0}} (-\mathbb{L}^{\frac{1}{2}})^{\chi(\alpha,\alpha)+2d_C(\alpha)} \cdot \biggl(\sum_{\{\pi^{[a,b]}\}\vdash\alpha} \mathbb{L}^{T(\{\pi^{[a,b]}\})-B(\{\pi^{[a,b]}\})} \cdot \prod_{0\le a,b\le N-1} f(\pi^{[a,b]})\biggr) y^\alpha.$$

Lemma 4.8 showed how the dimension vector depended on the partitions: we had

$$\alpha_i = \sum_{0\le a,b\le N-1} |\pi^{[a,b]}| - \sum_{[a,b]\not\ni i} l(\pi^{[a,b]}),$$

and an immediate corollary was that

$$\chi(\alpha,\alpha) + 2d_C(\alpha) = \sum_{i\in I_2} \biggl(\sum_{b\neq i} l(\pi^{[i+1,b]}) - \sum_{c\neq i+1} l(\pi^{[c,i]})\biggr)^2.$$

Combining this with the formula for the difference $T(\{\pi^{[a,b]}\}) - B(\{\pi^{[a,b]}\})$ (Proposition 4.18) gives

$$N^{\sigma}(y) = \sum_{\{\pi^{[a,b]}\}} \Bigl(\prod_{[a,b]\notin S} f(\pi^{[a,b]})\Bigr) \cdot \Bigl(\prod_{[a,b]\in S} f(\pi^{[a,b]}) \prod_{l\geq 1} (-\mathbb{L}^{\frac{1}{2}})^{-(b_l^{a,b})^2}\Bigr) \cdot \prod_{i=0}^{N-1} y_i^{\sum_{0\leq a,b\leq N-1} |\pi^{[a,b]}| - \sum_{[a,b]\not\ni i} l(\pi^{[a,b]})}.$$

To simplify notation, set

$$g(\pi) := f(\pi)\cdot \prod_{l\geq 1} (-\mathbb{L}^{\frac{1}{2}})^{-b_l^2} \qquad \text{for } \pi = (1^{b_1}\, 2^{b_2}\, 3^{b_3} \cdots);$$

then rearranging the products and summations gives

$$N^{\sigma}(y) = \prod_{[a,b]\notin S} \sum_{\pi^{[a,b]}} f(\pi^{[a,b]}) \cdot y'^{|\pi^{[a,b]}| - l(\pi^{[a,b]})} \cdot y_{[a,b]}^{l(\pi^{[a,b]})} \cdot \prod_{[a,b]\in S} \sum_{\pi^{[a,b]}} g(\pi^{[a,b]}) \cdot y'^{|\pi^{[a,b]}| - l(\pi^{[a,b]})} \cdot y_{[a,b]}^{l(\pi^{[a,b]})}.$$

Both of these series are know to have product expansions [Macdonald 1995]

$$f(t,a) = \sum_{\pi} f(\pi) a^{l(\pi)} t^{|\pi|-l(\pi)} = \mathrm{Exp}\Bigl(\frac{1}{1-\mathbb{L}^{-1}} \cdot \frac{a}{1-t}\Bigr),$$

$$g(t,a) = \sum_{\pi} g(\pi) a^{l(\pi)} t^{|\pi|-l(\pi)} = \mathrm{Exp}\Bigl(\frac{(-\mathbb{L}^{\frac{1}{2}})^{-1}}{1-\mathbb{L}^{-1}} \cdot \frac{a}{1-t}\Bigr).$$

Now N^{σ} is a product of such series, and multiplying together the corresponding exponential generating series gives the desired result

$$N^{\sigma}(y) = \mathrm{Exp}\Bigl(\frac{\mathbb{L}}{\mathbb{L}-1} \frac{1}{1-y'} \Bigl(\sum_{[a,b]\notin S} y_{[a,b]} - \mathbb{L}^{-\frac{1}{2}} \sum_{[a,b]\in S} y_{[a,b]}\Bigr)\Bigr). \qquad \square$$

Now we have computed I^{σ} and N^{σ}, and so by Lemma 4.3

$$A_U^{\sigma}(y) = \mathrm{Exp}\Bigl(\mathbb{L}\frac{y'}{1-y'} + \frac{\mathbb{L}}{\mathbb{L}-1} \frac{1}{1-y'} \Bigl(\sum_{[a,b]\notin S} y_{[a,b]} - \mathbb{L}^{-\frac{1}{2}} \sum_{[a,b]\in S} y_{[a,b]}\Bigr)\Bigr).$$

Or, reformulating this as a product over the set of roots, we get

$$\mathrm{Exp}\Bigl(\frac{1}{1-\mathbb{L}^{-1}} \Bigl((\mathbb{L}+N-1) \sum_{\alpha\in\Delta^{\mathrm{im}}_{\sigma,+}} y^{\alpha} + \sum_{\substack{\alpha\in\Delta^{\mathrm{re}}_{\sigma,+}\\ \sum_{I_2\cup I_3}\alpha_i \text{ even}}} y^{\alpha} - \mathbb{L}^{-\frac{1}{2}} \sum_{\substack{\alpha\in\Delta^{\mathrm{re}}_{\sigma,+}\\ \sum_{I_2\cup I_3}\alpha_i \text{ odd}}} y^{\alpha}\Bigr)\Bigr),$$

thus proving Theorem 0.1 for the special case of the partition σ:

$$A_U^{\sigma}(y) = \prod_{\alpha \in \Delta_{\sigma,+}} A^{\alpha}(y).$$

5. The universal DT series: general case

In this section we will prove Theorem 0.1 for any partition σ.

5A. *Mutation and the root system.* Recall that the simple reflection provides a bijection between $\Delta_{\sigma,+} \backslash \{\alpha_k\}$ and $\Delta_{\sigma',+} \backslash \{\alpha'_k\}$ (see Section 2C). The simple root α_k maps to $-\alpha'_k$.

For $\alpha \in \Delta_+^{\mathrm{re}}$, let x_α be a simple module of $\underline{\dim}\,\alpha$. By [Nagao 2012, Proposition 2.14], $\sum_{i \notin \hat{I}_r} \alpha_i$ is odd (resp. even) if and only if $\mathrm{ext}^1(x, x) = 0$ (resp. $= 1$). In particular, the parity of $\sum_{i \notin \hat{I}_r} \alpha_i$ is preserved by the simple reflection.

5B. *Wall-crossing formula.*

Theorem 5.1 [Nagao 2011c, Theorem 4.9].

$$A_U^{\sigma'}(\boldsymbol{y}) = \frac{A_U^{\sigma}(\boldsymbol{y})}{\mathbb{E}(y_k)} \times \mathbb{E}(y_k^{-1}).$$

Proof. Step 1: By the observation in Section 2C, we have the factorization

$$A_U^{\sigma} = \mathbb{E}(y_k) \times A_k^{\sigma},$$

where

$$\mathbb{E}(y) := \sum_{n \geq 0} \frac{[\mathrm{pf}]}{[\mathrm{GL}_n]_{\mathrm{vir}}} \cdot y^n, \quad y_k := y_{\alpha_k}$$

and A_k^{σ} is the generating series of virtual motives of moduli stacks of objects in $(\mathrm{mod}\, J_\sigma)_k$. We also have

$$A_U^{\sigma'} = A^{\sigma',k} \times \mathbb{E}(y_k^{-1}),$$

where $A^{\sigma',k}$ is the generating series of virtual motives of moduli stacks of objects in $(\mathrm{mod}\, J_{\sigma'})^k$.

Step 2: By Proposition 2.4, we have $A_k^{\sigma} = A^{\sigma',k}$ (see [Nagao 2011c, Theorem 4.7]). □

Now Theorem 0.1 follows for any σ from the result in Section 4 combined with Theorem 5.1 and the remark in Section 5A.

5C. *Factorization of the universal series.* We will say that a stability parameter ζ is generic, if for any stable J_σ-module V, we have $\zeta \cdot \underline{\dim} V \neq 0$. For generic stability parameter ζ, let $\mathfrak{M}^+_\zeta(J_\sigma, \alpha)$ (resp. $\mathfrak{M}^-_\zeta(J_\sigma, \alpha)$) denote the moduli stacks of J_σ-modules V such that $\underline{\dim} V = \alpha$ and such that all the HN factors F of V with respect to the stability parameter ζ satisfy $\zeta \cdot \underline{\dim} F > 0$ (resp. < 0). Let $[\mathfrak{M}^\pm_\zeta(J_\sigma, \alpha)]_{\mathrm{vir}}$ denote the virtual motive of the moduli stack defined in the same way as (3-2). We put

$$A^\pm_\zeta(y) = \sum_{\alpha \in \mathbb{N}^{\widehat{I}}} [\mathfrak{M}^\pm_\zeta(J, \alpha)]_{\mathrm{vir}} \cdot y^\alpha.$$

Lemma 5.2 [Morrison et al. 2012, Lemma 2.6]. *The generating series $A^\pm_\zeta$ are given by*

$$A^\pm_\zeta(y) = \prod_{\substack{\alpha \in \Delta_{\sigma,+} \\ \pm\zeta\cdot\alpha<0}} A^\alpha(y).$$

6. Motivic DT with framing and DT/PT series

6A. *Motivic DT invariants with framing.* We denote by $\widetilde{Q}_\sigma$ the new quiver obtained from Q_σ by adding a new vertex ∞ and a single new arrow $\infty \to 0$. Let $\widetilde{J}_\sigma = J_{\widetilde{Q}_\sigma, w_\sigma}$ be the Jacobian algebra corresponding to the quiver with potential $(\widetilde{Q}_\sigma, w_\sigma)$, where we view w_σ as a potential for $\widetilde{Q}_\sigma$ in the obvious way.

Let $\zeta \in \mathbb{R}^{\widehat{I}}$ be a vector, which we will refer to as the stability parameter. A $\widetilde{J}_\sigma$-representation $\widetilde{V}$ with $\dim \widetilde{V}_\infty = 1$ is said to be ζ-(semi)stable, if it is (semi)stable with respect to $(\zeta, \zeta_\infty) \in \mathbb{R}^{\widehat{I} \sqcup \{\infty\}}$ (see Definition 3.3), where $\zeta_\infty = -\zeta \cdot \underline{\dim} V$. As in Section 3B, a stability parameter $\zeta \in \mathbb{R}^{Q_0}$ is said to be *generic* if for any stable J-module V we have $\zeta \cdot \underline{\dim} V \neq 0$.

For a stability parameter $\zeta \in \mathbb{R}^{Q_0}$ and a dimension vector $\alpha \in (\mathbb{Z}_{\geq 0})^{\widehat{I}}$, let $\mathfrak{M}_\zeta(\widetilde{J}_\sigma, \alpha)$ denote the moduli stack of ζ-semistable $\widetilde{J}_\sigma$-representations with dimension vector $(\alpha, 1)$. As in the introduction, we define the generating function

$$Z_\zeta(y_0, \dots, y_{N-1}) = Z_\zeta(y) := \sum_{\alpha \in (\mathbb{Z}_{\geq 0})^{\widehat{I}}} [\mathfrak{M}_\zeta(\widetilde{J}_\sigma, \alpha)]_{\mathrm{vir}} \cdot y^\alpha.$$

Theorem 6.1 [Morrison et al. 2012, Proposition 4.6]. *For a generic stability parameter ζ, we have*

$$Z_\zeta(y) = \frac{A^-_\zeta(-\mathbb{L}^{\frac{1}{2}} y_0, y_1, \dots, y_{N-1})}{A^-_\zeta(-\mathbb{L}^{-\frac{1}{2}} y_0, y_1, \dots, y_{N-1})}, \tag{6-1}$$

where A^-_ζ was defined in Section 5C.

Combined with Lemma 5.2, we get the formula in Corollary 0.2.

Remark 6.2. If we cross the wall W_α, we get (or lose) a factor $Z_\alpha(y)$ in the generating function. This is compatible with the result in [Nagao 2011b].

6B. ***Chambers in the moduli spaces.*** For a root $\alpha \in \Lambda$, let W_α denote the hyperplane in the space $\mathbb{R}^{\widehat{I}}$ of stability parameters which is orthogonal to α. We put

$$W = W_\delta \cup \bigcup_{\alpha \in \Delta^{\mathrm{re}}_{\sigma,+}} W_\alpha.$$

A connected component of the complement of W in $\mathbb{R}^{\widehat{I}}$ is called a chamber.

Theorem 6.3 [Nagao 2012, Proposition 2.10; Nagao and Nakajima 2011, Propositions 3.10, 3.11]. *The set of generic parameters in $\mathbb{R}^{\widehat{I}}$ is the compliment of W.*

(i) *For ζ with $\zeta_i < 0$ for all i, the moduli spaces $\mathfrak{M}_\zeta(\widetilde{J}, \alpha)$ are the NCDT moduli spaces, the moduli spaces of cyclic J-modules from [Szendrői 2008].*

(ii) *For ζ in the same chamber as* $(1 - N + \varepsilon, 1, 1, \dots, 1)$ $(0 < \varepsilon \ll 1)$, *the moduli spaces $\mathfrak{M}_\zeta(\widetilde{J}, \alpha)$ are the DT moduli spaces of Y_σ from [Maulik et al. 2006], the moduli spaces of subschemes on Y_σ with support in dimension at most* 1.

(iii) *For ζ in the same chamber as* $(1 - N - \varepsilon, 1, 1, \dots, 1)$ $(0 < \varepsilon \ll 1)$, *the moduli spaces $\mathfrak{M}_\zeta(\widetilde{J}, \alpha)$ are the PT moduli spaces of Y_σ introduced in [Pandharipande and Thomas 2009]; these are moduli spaces of stable rank-*1 *coherent systems.*

Remark 6.4. In the above statements ε depends on the dimension vector $(\alpha, 1)$.

6C. ***Motivic PT and DT invariants.*** Let

$$\zeta_{\mathrm{DT}} = (1 - N - \varepsilon, 1, 1, \dots, 1), \quad \zeta_{\mathrm{PT}} = (1 - N + \varepsilon, 1, 1, \dots, 1) \qquad (0 < \varepsilon \ll 1)$$

be stability parameters corresponding to DT and PT moduli spaces. Then we have

$$\{\alpha \in \Delta_{\sigma,+} \mid \zeta_{\mathrm{DT}} \cdot \alpha < 0\} = \Delta^{\mathrm{re},+}_{+},$$
$$\{\alpha \in \Delta_{\sigma,+} \mid \zeta_{\mathrm{PT}} \cdot \alpha < 0\} = \Delta^{\mathrm{re},+}_{+} \sqcup \Delta^{\mathrm{im}}_{+}.$$

As we mentioned in the introduction, the variable change induced by the derived equivalence is given by

$$s := y_0 \cdot y_1 \cdots y_{N-1}, \quad T_i = y_i.$$

Here s is the variable for the homology class of a point and T_i is the variable for the homology class of C_i. Then we get the formulae in Corollary 0.3.

6D. ***Connection with the refined topological vertex.*** As in [Nagao 2011a], we can apply the vertex operator method [Okounkov et al. 2006] to get a product expansion of the refined topological vertex for $\mathcal{Y}_\sigma$. Then we see that the PT generating function can be described by the refined topological vertices normalized by the refined MacMahon functions.[5]

Appendix

Throughout this appendix we will work with a fixed choice of basis $\mathcal{B}$. In Section 4B we chose a basis

$$\mathcal{B}=\{H^p v_l^{a,b}(k)\mid 1\le k\le b_l^{a,b},\ 0\le a,b\le N-1,\ 0\le p\le N(l-1)+|b-a|+1\}$$

and defined linear spaces

$$\begin{aligned}&F(\{\pi^{[a,b]}\})\\ &=\Bigl\{L\in\bigoplus_{i\in I_1}\mathrm{Hom}(V_i,V_i)\oplus\bigoplus_{i\in I_2}\mathrm{Hom}(V_{i+1},V_i)\;\Big|\;(L,H(\pi^{[a,b]}))\in R(\{\pi^{[a,b]}\})\Bigr\},\\ &S(\{\pi^{[a,b]}\})=\Bigl\{N\in\bigoplus_{i\in I_1\cup I_2\cup I_3}\mathrm{Hom}(V_i,V_i)\;\Big|\;[N,H(\pi^{[a,b]})]=0\Bigr\},\end{aligned}$$

with dimensions $T(\{\pi^{[a,b]}\})=\dim F(\{\pi^{[a,b]}\})$ and $B(\{\pi^{[a,b]}\})=\dim S(\{\pi^{[a,b]}\})$. The goal of the appendix is to prove Proposition 4.18, that is, to show that the difference $T(\{\pi^{[a,b]}\})-B(\{\pi^{[a,b]}\})$ is equal to

$$-\frac{1}{2}\sum_{i\in I_2}\Bigl(\sum_{b\neq i}l(\pi^{[i+1,b]})-\sum_{c\neq i+1}l(\pi^{[c,i]})\Bigr)^2 -\frac{1}{2}\sum_{a\in I_3,b\notin I_2}\sum_{i\ge 1}(b_i^{a,b})^2-\frac{1}{2}\sum_{a\notin I_3,b\in I_2}\sum_{i\ge 1}(b_i^{a,b})^2.$$

For some early examples it becomes clear that the dimensions of $F(\{\pi^{[a,b]}\})$ and $S(\{\pi^{[a,b]}\})$ are determined by solving a set of linearly independent equations. We will see that these dimensions are quadratic polynomials in the number of parts $b_l^{a,b}$ of the partitions $\{\pi^{[a,b]}\}$. An initial means of simplifying the calculation is to break the spaces $F(\{\pi^{[a,b]}\})$ and $S(\{\pi^{[a,b]}\})$ down into simpler spaces. One easy observation is that not only are the spaces $F(\{\pi^{[a,b]}\})$ and $S(\{\pi^{[a,b]}\})$ linear, but they come with a natural vector space structure, the origin corresponding

[5]Unfortunately, the DT generating function does not coincide with the refined topological vertex. See [Morrison et al. 2012, §4.3] for details.

to the zero matrix in both cases. This means that we have decompositions as vector spaces

$$F(\{\pi^{[a,b]}\}) = \bigoplus_{0\le a,b,c,d\le N-1} F(\pi^{[a,b]},\pi^{[c,d]}),$$
$$S(\{\pi^{[a,b]}\}) = \bigoplus_{0\le a,b,c,d\le N-1} S(\pi^{[a,b]},\pi^{[c,d]})$$

whose summands are given by the following definition:

Definition A.5. We define

$$F(\pi^{[a,b]},\pi^{[c,d]}) = F(\{\pi^{[a,b]}\}) \cap \bigoplus_{i\in I_1\cup I_2} \mathrm{Hom}(V^{a,b},V^{c,d}),$$
$$S(\pi^{[a,b]},\pi^{[c,d]}) = S(\pi^{[a,b]},\pi^{[c,d]}) \cap \bigoplus_{i\in I_1\cup I_2\cup I_3} \mathrm{Hom}(V^{a,b},V^{c,d}).$$

These subspaces are essentially given by the block matrices for the decomposition $V = \bigoplus_{0\le a,b\le N-1} V^{a,b}$.

Definition A.6. We define

$$T(\pi^{[a,b]},\pi^{[c,d]}) = \dim F(\pi^{[a,b]},\pi^{[c,d]})$$
$$B(\pi^{[a,b]},\pi^{[c,d]}) = \dim S(\pi^{[a,b]},\pi^{[c,d]}).$$

Both $T(\pi^{[a,b]},\pi^{[c,d]})$ and $B(\pi^{[a,b]},\pi^{[c,d]})$ can be written as quadratic expressions in the number of parts of $\pi^{[a,b]}$ and $\pi^{[c,d]}$. To do so, we introduce a quadratic form on the space of all partitions and a combinatorial operation that removes a box from each column of the partition.

Definition A.7. We define

$$M : \mathcal{P}\otimes\mathcal{P} \to \mathbb{Z}_{\ge 0},$$
$$(1^{b_1}\,2^{b_2}\,3^{b_3}\cdots)\otimes(1^{c_1}\,2^{c_2}\,3^{c_3}\cdots) \mapsto \sum_{i\ge 1}\Big(\sum_{j\ge i} b_j\Big)\Big(\sum_{j\ge i} c_j\Big),$$
$${}' : \mathcal{P}\to\mathcal{P},$$
$$\pi = (1^{b_1}\,2^{b_2}\,3^{b_3}\cdots) \mapsto \pi' = (1^{b_2}\,2^{b_3}\,3^{b_4}\cdots).$$

Let us begin with the easier case. We compute dimensions $B(\pi^{[a,b]},\pi^{[c,d]})$ of the spaces $S(\pi^{[a,b]},\pi^{[c,d]})$.

Lemma A.8. *Let $N \in S(\pi^{[a,b]}, \pi^{[c,d]})$; then the matrix N is uniquely determined by its value on the vectors $v_l^{a,b}(k)$. Moreover, the only restriction on the image of such a vector is that it lie in the linear subspace*

$$N(v_l^{a,b}) \in V_a \cap V^{c,d} \cap V^{N\cdot(l-1)+|b-a|+1}.$$

Proof. To define the linear map N on the space $V^{a,b}$, it suffices to define its value at each of the basis vectors

$$\{H^r v_l^{a,b}(k) \mid 0 \le r \le N \cdot (l-1) + |b-a|, 1 \le k \le b_l^{a,b}\}.$$

However, for $N \in S(\pi^{[a,b]}, \pi^{[c,d]})$, we have

$$N(H^r v_l^{a,b}(k)) = H^r(N v_l^{a,b}(k));$$

therefore the value of N at each $H^r v_l^{a,b}(k)$ is determined by $N v_l^{a,b}(k)$. This proves the first part of the lemma. Now we know that the matrix N maps the vector space at the a-th vertex to itself: $V_a \to V_a$; also, since $N \in S(\pi^{[a,b]}, \pi^{[c,d]})$, we insist that its image be in $V^{c,d}$. The only additional condition on the image of the vector $v_l^{a,b}(k)$ is

$$H^{N\cdot(l-1)+|b-a|+1}(N v_l^{a,b}(k)) = N(H^{N\cdot(l-1)+|b-a|+1} v_l^{a,b}(k)) = 0.$$

Combining these three conditions above, we have

$$N(v_l^{a,b}(k)) \in V_a \cap V^{c,d} \cap V^{N\cdot(l-1)+|b-a|+1}. \qquad \square$$

Corollary A.9. *We have*

$$B(\pi^{[a,b]}, \pi^{[c,d]}) = \begin{cases} M(\pi^{[a,b]}, \pi^{[c,d]}) & \text{if } a \in [c,d] \text{ and } |d-a| \le |b-a|, \\ M((\pi^{[a,b]})', \pi^{[c,d]}) & \text{if } a \in [c,d] \text{ and } |d-a| > |b-a|, \\ M(\pi^{[a,b]}, (\pi^{[c,d]})') & \text{if } a \notin [c,d] \text{ and } |d-a| \le |b-a|, \\ M((\pi^{[a,b]})', (\pi^{[c,d]})') & \text{if } a \notin [c,d] \text{ and } |d-a| > |b-a|. \end{cases}$$

Proof. Let $N \in S(\pi^{[a,b]}, \pi^{[c,d]})$. Each vector $v_l^{a,b}(k)$ with $1 \le k \le b_l^{a,b}$ can take any value in the vector space $V_a \cap V^{c,d} \cap V^{N\cdot(l-1)+|b-a|+1}$, and so the dimension of $S(\pi^{a,b}, \pi^{[c,d]})$ is given by

$$B(\pi^{[a,b]}, \pi^{[c,d]}) = \sum_{l \ge 0} b_l^{a,b} \cdot \dim(V_a \cap V^{c,d} \cap V^{N\cdot(l-1)+|b-a|+1}).$$

Counting the number of basis vectors of $V^{c,d}$ that lie in V_a, we see there are four possibilities for $\dim(V_a \cap V^{c,d} \cap V^{N\cdot(l-1)+|b-a|+1})$:

$$\begin{array}{ll} \sum_{i=1}^{l} i b_i^{c,d} + l \sum_{i\geq l} b_i^{c,d} & \text{if } a \in [c,d] \text{ and } |d-a| \leq |b-a|, \\ \sum_{i=1}^{l-1} i b_i^{c,d} + (l-1) \sum_{i\geq l} b_i^{c,d} & \text{if } a \in [c,d] \text{ and } |d-a| > |b-a|, \\ \sum_{i=1}^{l} i b_{i+1}^{c,d} + l \sum_{i\geq l} b_{i+1}^{c,d} & \text{if } a \notin [c,d] \text{ and } |d-a| \leq |b-a|, \\ \sum_{i=1}^{l-1} i b_{i+1}^{c,d} + (l-1) \sum_{i\geq l} b_{i+1}^{c,d} & \text{if } a \notin [c,d] \text{ and } |d-a| > |b-a|. \end{array}$$

Consider the first case $a \in [c,d]$ and $|d-a| \leq |b-a|$; then

$$\begin{aligned} B(\pi^{[a,b]}, \pi^{[c,d]}) &= \sum_{l\geq 1} b_l^{a,b} \cdot \Big(\sum_{i=1}^{l} i b_i^{c,d} + l \sum_{i\geq l} b_i^{c,d} \Big) \\ &= \sum_{i\geq 1} \Big(\sum_{l\geq i} b_l^{a,b} \Big) \cdot \Big(\sum_{l\geq i} b_l^{c,d} \Big) \\ &= M(\pi^{[a,b]}, \pi^{[c,d]}). \end{aligned}$$

The other three cases are identical. The relabeling of the partitions in these cases is encoded by the operation $\pi \mapsto \pi'$. □

Now we turn to computing the dimensions $T(\pi^{[a,b]}, \pi^{[c,d]})$ of the spaces $F(\pi^{[a,b]}, \pi^{[c,d]})$. This will be more intricate.

Lemma A.10. *Suppose $a \in I_1 \cup I_3$ and $L \in F(\pi^{[a,b]}, \pi^{[c,d]})$; then the map L is uniquely determined by its value on the vectors $v_l^{a,b}(k)$. Moreover the only restriction on the image of such a vector is that it lie in a linear subspace*:

$$L v_l^{a,b}(k) \in \begin{cases} V_a \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d} & \text{if } a \in I_1 \text{ and } b \notin I_2, \\ V_a \cap V^{N\cdot(l-1)+|b-a|} \cap V^{c,d} & \text{if } a \in I_1 \text{ and } b \in I_2, \\ V_{a-1} \cap V^{N\cdot(l-1)+|b-a|+2} \cap V^{c,d} & \text{if } a \in I_3 \text{ and } b \notin I_2, \\ V_{a-1} \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d} & \text{if } a \in I_3 \text{ and } b \in I_2. \end{cases}$$

Proof. To define the linear map L on the space $V^{a,b}$, it suffices to define its value at each of the basis vectors

$$\{ H^r v_l^{a,b}(k) \mid 0 \leq r \leq N\cdot(l-1) + |b-a|, 1 \leq k \leq b_l^{a,b} \}.$$

However, for $L \in F(\pi^{[a,b]}, \pi^{[c,d]})$, we know that the pair $(L, H(\pi^{[a,b]})) \in R(\{\pi^{[a,b]}\})$ satisfies the relations coming from the superpotential:

$$\begin{aligned} r_i h^+_{i-\frac{1}{2}} &= h^+_{i-\frac{1}{2}} r_{i-1} && \text{for } i \in [1, N'-1] \cap I_1, \\ r_0 h^+_{N-\frac{1}{2}} &= h^+_{N-\frac{1}{2}} h^+_{N-\frac{3}{2}} h^-_{N-\frac{3}{2}}, \\ h^-_{N'+\frac{1}{2}} h^+_{N'+\frac{1}{2}} h^+_{N'-\frac{1}{2}} &= h^+_{N'-\frac{1}{2}} r_{N'-1}, \\ h^-_{i+\frac{3}{2}} h^+_{i+\frac{3}{2}} h^+_{i+\frac{1}{2}} &= h^+_{i+\frac{1}{2}} h^+_{i-\frac{1}{2}} h^-_{i-\frac{1}{2}} && \text{for } i = [N'+1, N-3] \cap I_3. \end{aligned}$$

As in Lemma A.8, once the value of L is determined for $v_l^{a,b}(k)$ it is uniquely determined for all $H^r v_l^{a,b}(k)$ by the condition that the above relations be satisfied for the pair $(L, H(\pi^{[a,b]})$. To be precise, if $a \in I_1$ we have

$$L : H^r(v_l^{a,b}(k)) \mapsto \begin{cases} H^r L(v_l^{a,b}(k)) & \text{if } a+r \in I_1, \\ 0 & \text{if } a+r \in I_2, \\ H^{r-1} L(v_l^{a,b}(k)) & \text{if } a+r \in I_3, \end{cases}$$

and if $a \in I_3$ then

$$L : H^r(v_l^{a,b}(k)) \mapsto \begin{cases} H^{r+1} L(v_l^{a,b}(k)) & \text{if } a+r \in I_1, \\ 0 & \text{if } a+r \in I_2, \\ H^r L(v_l^{a,b}(k)) & \text{if } a+r \in I_3. \end{cases}$$

Since $L \in F(\pi^{[a,b]}, \pi^{[c,d]})$ by definition, its image must lie in the space $V^{c,d}$; also, if $a \in I_1$ then $L : V_a \to V_a$ and if $a \in I_3$ then $L : V_a \to V_{a-1}$. The only further condition on the image of a vector $v_l^{a,b}(k)$ is that its image be killed by a high-enough power of H. It is given that $H^{N\cdot(l-1)+|b-a|+1} v_l^{a,b}(k) = 0$, so then $H^t(Lv_l^{a,b}(k)) = 0$, where the exponent t is read off from the defining relations on L above. In the separate cases,

$$Lv_l^{a,b}(k) \in \begin{cases} V_a \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d} & \text{if } a \in I_1 \text{ and } b \notin I_2, \\ V_a \cap V^{N\cdot(l-1)+|b-a|} \cap V^{c,d} & \text{if } a \in I_1 \text{ and } b \in I_2, \\ V_{a-1} \cap V^{N\cdot(l-1)+|b-a|+2} \cap V^{c,d} & \text{if } a \in I_3 \text{ and } b \notin I_2, \\ V_{a-1} \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d} & \text{if } a \in I_3 \text{ and } b \in I_2, \end{cases}$$

proving the result. □

We have a result similar to Lemma A.10 when $a \in I_2$.

Lemma A.11. *Suppose $a \in I_2$ and $L \in F(\pi^{[a,b]}, \pi^{[c,d]})$; then the map L is uniquely determined by its value on the vectors $Hv_l^{a,b}(k)$. Moreover, the only restriction on the image of such a vector is that it lie in a linear subspace*:

$$L(Hv_l^{a,b}(k)) \in \begin{cases} V_a \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d} & \text{if } b \notin I_2, \\ V_a \cap V^{N\cdot(l-1)+|b-a|} \cap V^{c,d} & \text{if } b \in I_2. \end{cases}$$

Proof. Again, we know that to define the linear map L on the space $V^{a,b}$ it suffices to define its value at each of the basis vectors

$$\{H^r v_l^{a,b}(k) \mid 0 \le r \le N \cdot (l-1) + |b-a|, 1 \le k \le b_l^{a,b}\}.$$

Since by definition if $a \in I_2$ then $Lv_l^{a,b}(k) = 0$, the map is already trivially determined on these vectors and their image does not suffice to determine the map in general. However if we consider the vectors $Hv_l^{a,b}(k)$, then once the value of L is determined for $Hv_l^{a,b}(k)$ it is uniquely determined for all $H^r v_l^{a,b}(k)$ by the condition that the relations (see Lemma A.10) be satisfied by the pair $(L, H(\pi^{[a,b]})$. To be precise, if $a \in I_2$ we have

$$L : H^r(v_l^{a,b}(k)) \mapsto \begin{cases} H^r L(Hv_l^{a,b}(k)) & \text{if } a+r \in I_1, \\ 0 & \text{if } a+r \in I_2, \\ H^{r-1} L(Hv_l^{a,b}(k)) & \text{if } a+r \in I_3. \end{cases}$$

By definition, we know that the image of L lies in $V^{c,d}$ and also that for $a \in I_2$ we have $L : V_{a+1} \to V_a$. As before the only remaining condition on the image of $v_l^{a,b}(k)$ is that it be killed by a high-enough power of H. From the definition of L above we see that

$$L(Hv_l^{a,b}(k)) \in \begin{cases} V_a \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d} & \text{if } b \notin I_2, \\ V_a \cap V^{N\cdot(l-1)+|b-a|} \cap V^{c,d} & \text{if } b \in I_2, \end{cases}$$

proving the result. □

The following notation encapsulates the dimensions of all the vector spaces encountered in the last two lemmas.

Definition A.12. We define integers

$$d_{a,b:c,d}(l) = \begin{cases} \dim(V_a \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d}) & \text{if } a \in I_1 \cup I_2 \text{ and } b \notin I_2, \\ \dim(V_a \cap V^{N\cdot(l-1)+|b-a|} \cap V^{c,d}) & \text{if } a \in I_1 \cup I_2 \text{ and } b \in I_2, \\ \dim(V_{a-1} \cap V^{N\cdot(l-1)+|b-a|+2} \cap V^{c,d}) & \text{if } a \in I_3 \text{ and } b \notin I_2, \\ \dim(V_{a-1} \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d}) & \text{if } a \in I_3 \text{ and } b \in I_2. \end{cases}$$

From Lemma A.10 and Lemma A.11 we deduce the dimension of the spaces $F(\pi^{[a,b]}, \pi^{[c,d]})$.

Corollary A.13. *If $a \in I_1 \cup I_2$ and $b \notin I_2$ then*

$$T(\pi^{[a,b]}, \pi^{[c,d]}) = \begin{cases} M(\pi^{[a,b]}, \pi^{[c,d]}) & \textit{if } a \in [c,d] \textit{ and } |d-a| \le |b-a|, \\ M((\pi^{[a,b]})', \pi^{[c,d]}) & \textit{if } a \in [c,d] \textit{ and } |d-a| > |b-a|, \\ M(\pi^{[a,b]}, (\pi^{[c,d]})') & \textit{if } a \notin [c,d] \textit{ and } |d-a| \le |b-a|, \\ M((\pi^{[a,b]})', (\pi^{[c,d]})') & \textit{if } a \notin [c,d] \textit{ and } |d-a| > |b-a|. \end{cases}$$

If $a \in I_1 \cup I_2$ and $b \in I_2$ then

$$T(\pi^{[a,b]}, \pi^{[c,d]}) = \begin{cases} M(\pi^{[a,b]}, \pi^{[c,d]}) & \text{if } a \in [c,d] \text{ and } |d-a| \leq |b-a|-1, \\ M((\pi^{[a,b]})', \pi^{[c,d]}) & \text{if } a \in [c,d] \text{ and } |d-a| > |b-a|-1, \\ M(\pi^{[a,b]}, (\pi^{[c,d]})') & \text{if } a \notin [c,d] \text{ and } |d-a| \leq |b-a|-1, \\ M((\pi^{[a,b]})', (\pi^{[c,d]})') & \text{if } a \notin [c,d] \text{ and } |d-a| > |b-a|-1. \end{cases}$$

If $a \in I_3$ and $b \notin I_2$ then

$$T(\pi^{[a,b]}, \pi^{[c,d]}) = \begin{cases} M(\pi^{[a,b]}, \pi^{[c,d]}) & \text{if } a-1 \in [c,d] \text{ and } |d-(a-1)| \leq |b-a|+1, \\ M((\pi^{[a,b]})', \pi^{[c,d]}) & \text{if } a-1 \in [c,d] \text{ and } |d-(a-1)| > |b-a|+1, \\ M(\pi^{[a,b]}, (\pi^{[c,d]})') & \text{if } a-1 \notin [c,d] \text{ and } |d-(a-1)| \leq |b-a|+1, \\ M((\pi^{[a,b]})', (\pi^{[c,d]})') & \text{if } a-1 \notin [c,d] \text{ and } |d-(a-1)| > |b-a|+1. \end{cases}$$

If $a \in I_3$ and $b \in I_2$ then

$$T(\pi^{[a,b]}, \pi^{[c,d]}) = \begin{cases} M(\pi^{[a,b]}, \pi^{[c,d]}) & \text{if } a-1 \in [c,d] \text{ and } |d-(a-1)| \leq |b-a|, \\ M((\pi^{[a,b]})', \pi^{[c,d]}) & \text{if } a-1 \in [c,d] \text{ and } |d-(a-1)| > |b-a|, \\ M(\pi^{[a,b]}, (\pi^{[c,d]})') & \text{if } a-1 \notin [c,d] \text{ and } |d-(a-1)| \leq |b-a|, \\ M((\pi^{[a,b]})', (\pi^{[c,d]})') & \text{if } a-1 \notin [c,d] \text{ and } |d-(a-1)| > |b-a|. \end{cases}$$

Proof. We know that if $a \in I_1 \cup I_3$ (resp. $a \in I_2$) then the map $L \in F(\pi^{[a,b]}, \pi^{[c,d]})$ is determined by its value at the vectors $v_l^{a,b}(k)$ (resp. $Hv_l^{a,b}(k)$) for $1 \leq k \leq b_l^{a,b}$. In the notation of the previous definition, such a vector takes values in a space of dimension $d_{a,b;c,d}(l)$. So in all cases the total dimension of the space $F(\pi^{[a,b]}, \pi^{[c,d]})$ is equal to

$$T(\pi^{[a,b]}, \pi^{[c,d]}) = \sum_{l \geq 1} b_l^{a,b} \cdot d_{a,b;c,d}(l).$$

In the above definition of $d_{a,b;c,d}(l)$ there are four possible forms, depending on the value of a and b. Lets consider the first case, where $a \in I_1 \cup I_2$ and $b \notin I_2$. Then we have that

$$d_{a,b;c,d}(l) = \dim(V_a \cap V^{N\cdot(l-1)+|b-a|+1} \cap V^{c,d}).$$

Counting the number of basis vectors of $V^{c,d}$ that lie in V_a, we see there are four possibilities for $\dim(V_a \cap V^{c,d} \cap V^{N\cdot(l-1)+|b-a|+1})$:

$$\begin{cases} \sum_{i=1}^{l} i b_i^{c,d} + l \sum_{i\geq l} b_i^{c,d} & \text{if } a \in [c,d] \text{ and } |d-a| \leq |b-a|, \\ \sum_{i=1}^{l-1} i b_i^{c,d} + (l-1) \sum_{i\geq l} b_i^{c,d} & \text{if } a \in [c,d] \text{ and } |d-a| > |b-a|, \\ \sum_{i=1}^{l} i b_{i+1}^{c,d} + l \sum_{i\geq l} b_{i+1}^{c,d} & \text{if } a \notin [c,d] \text{ and } |d-a| \leq |b-a|, \\ \sum_{i=1}^{l-1} i b_{i+1}^{c,d} + (l-1) \sum_{i\geq l} b_{i+1}^{c,d} & \text{if } a \notin [c,d] \text{ and } |d-a| > |b-a|. \end{cases}$$

In the first case $a \in [c,d]$ and $|d-a| \leq |b-a|$, we have

$$\begin{aligned} T(\pi^{[a,b]}, \pi^{[c,d]}) &= \sum_{l\geq 1} b_l^{a,b} \cdot \left(\sum_{i=1}^{l} i b_i^{c,d} + l \sum_{i\geq l} b_i^{c,d} \right) \\ &= \sum_{i\geq 1} \left(\sum_{l\geq i} b_l^{a,b} \right) \cdot \left(\sum_{l\geq i} b_l^{c,d} \right) \\ &= M(\pi^{[a,b]}, \pi^{[c,d]}). \end{aligned}$$

In the second case $a \in [c,d]$ and $|d-a| > |b-a|$, we have

$$\begin{aligned} T(\pi^{[a,b]}, \pi^{[c,d]}) &= \sum_{l\geq 1} b_l^{a,b} \cdot \left(\sum_{i=1}^{l-1} i b_i^{c,d} + (l-1) \sum_{i\geq l} b_i^{c,d} \right) \\ &= \sum_{i\geq 1} \left(\sum_{l\geq i} b_{l+1}^{a,b} \right) \cdot \left(\sum_{l\geq i} b_l^{c,d} \right) \\ &= M((\pi^{[a,b]})', \pi^{[c,d]}). \end{aligned}$$

In the third case $a \notin [c,d]$ and $|d-a| \leq |b-a|$, we have

$$\begin{aligned} T(\pi^{[a,b]}, \pi^{[c,d]}) &= \sum_{l\geq 1} b_l^{a,b} \cdot \left(\sum_{i=1}^{l} i b_{i+1}^{c,d} + l \sum_{i\geq l} b_{i+1}^{c,d} \right) \\ &= \sum_{i\geq 1} \left(\sum_{l\geq i} b_l^{a,b} \right) \cdot \left(\sum_{l\geq i} b_{l+1}^{c,d} \right) \\ &= M(\pi^{[a,b]}, (\pi^{[c,d]})'). \end{aligned}$$

Finally, in the fourth case $a \notin [c,d]$ and $|d-a| > |b-a|$, we have

$$\begin{aligned} T(\pi^{[a,b]}, \pi^{[c,d]}) &= \sum_{l \geq 1} b_l^{a,b} \cdot \left(\sum_{i=1}^{l-1} i\, b_{i+1}^{c,d} + (l-1) \sum_{i \geq l} b_{i+1}^{c,d} \right) \\ &= \sum_{i \geq 1} \left(\sum_{l \geq i} b_{l+1}^{a,b} \right) \cdot \left(\sum_{l \geq i} b_{l+1}^{c,d} \right) \\ &= M((\pi^{[a,b]})', (\pi^{[c,d]})'). \end{aligned}$$

This completes the situation when $a \in I_1 \cup I_2$ and $b \notin I_2$. In the other situations, $a \in I_1 \cup I_2$ and $b \in I_2$, or $a \in I_3$ and $b \notin I_2$, or $a \in I_3$ and $b \in I_2$. All of these cases can be dealt with in a similar manner. □

Now we have computed all the dimensions $T(\pi^{[a,b]}, \pi^{[c,d]})$ and $B(\pi^{[a,b]}, \pi^{[c,d]})$. The next lemma combines Corollaries A.9 and A.13 to compute their difference. We see that in most cases there is an exact cancellation:

Lemma A.14. *The equality*

$$T(\pi^{[a,b]}, \pi^{[c,d]}) = B(\pi^{[a,b]}, \pi^{[c,d]})$$

holds, except in the following cases, where we give the possible values of the difference $T(\pi^{[a,b]}, \pi^{[c,d]}) - B(\pi^{[a,b]}, \pi^{[c,d]})$:

Case 1: $a \in I_1 \cup I_2$, $b = d \in I_2$:

$$\begin{aligned} &M((\pi^{[a,b]})', \pi^{[c,b]}) - M(\pi^{[a,b]}, \pi^{[c,b]}) && \text{if } a \in [c,b], \\ &M((\pi^{[a,b]})', (\pi^{[c,b]})') - M(\pi^{[a,b]}, (\pi^{[c,b]})') && \text{if } a \notin [c,b]. \end{aligned}$$

Case 2: $a \in I_3$, $b \notin I_2$, $d = a-1 \in I_2$:

$$\begin{aligned} &M(\pi^{[a,b]}, \pi^{[a,a-1]}) - M((\pi^{[a,b]})', \pi^{[a,a-1]}) && \text{if } a = c, \\ &M(\pi^{[a,b]}, \pi^{[c,a-1]}) - M((\pi^{[a,b]})', (\pi^{[c,a-1]})') && \text{if } a \neq c. \end{aligned}$$

Case 3: $a \in I_3$, $b \notin I_2$, $a = c$, $d \neq a-1$:

$$\begin{aligned} &M(\pi^{[a,b]}, (\pi^{[a,d]})') - M(\pi^{[a,b]}, \pi^{[a,d]}) && \text{if } |d-a| \leq |b-a|, \\ &M((\pi^{[a,b]})', (\pi^{[a,d]})') - M((\pi^{[a,b]})', \pi^{[a,d]}) && \text{if } |d-a| > |b-a|. \end{aligned}$$

Case 4: $a \in I_3$, $b \in I_2$, $d = a-1$:

$$\begin{aligned} &M(\pi^{[a,b]}, \pi^{[a,a-1]}) - M((\pi^{[a,b]})', \pi^{[a,a-1]}) && \text{if } a = c \text{ and } b \neq a-1, \\ &M(\pi^{[a,a-1]}, \pi^{[c,a-1]}) - M(\pi^{[a,a-1]}, (\pi^{[c,a-1]})') && \text{if } a \neq c \text{ and } b = a-1, \\ &M(\pi^{[a,b]}, \pi^{[c,a-1]}) - M((\pi^{[a,b]})', (\pi^{[c,a-1]})') && \text{if } a \neq c \text{ and } b \neq a-1. \end{aligned}$$

Case 5: $a \in I_3$, $b \in I_2$, $a-1 \in [c,d]$, $d \neq a-1$, $b = d$:

$$M((\pi^{[a,b]})', \pi^{[c,b]}) - M(\pi^{[a,b]}, \pi^{[c,b]}).$$

Case 6: $a \in I_3$, $b \in I_2$, $a-1 \notin [c,d]$, $a = c$, $|d-a| < |b-a|$:

$$M(\pi^{[a,b]}, (\pi^{[a,d]})') - M(\pi^{[a,b]}, \pi^{[a,d]}).$$

Case 7: $a \in I_3$, $b \in I_2$, $a-1 \notin [c,d]$:

$$\begin{array}{ll} M((\pi^{[a,b]})', (\pi^{[a,b]})') - M(\pi^{[a,b]}, \pi^{[a,b]}) & \text{if } a = c \text{ and } b = d, \\ M((\pi^{[a,b]})', (\pi^{[a,d]})') - M((\pi^{[a,b]})', \pi^{[a,d]}) & \text{if } a = c \text{ and } |d-a| > |b-a|, \\ M((\pi^{[a,b]})', (\pi^{[c,b]})') - M(\pi^{[a,b]}, (\pi^{[c,b]})') & \text{if } a \neq c \text{ and } b = d. \end{array}$$

Proof. Compare Corollaries A.9 and A.13. □

Our aim throughout this appendix has been to prove Proposition 4.18 and deduce that the difference $\sum_{0 \le a,b,c,d \le N-1} T(\pi^{[a,b]}, \pi^{[c,d]}) - B(\pi^{[a,b]}, \pi^{[c,d]})$ is equal to

$$-\frac{1}{2} \sum_{i \in I_2} \left(\sum_{b \neq i} l(\pi^{[i+1,b]}) - \sum_{c \neq i+1} l(\pi^{[c,i]}) \right)^2 - \frac{1}{2} \sum_{a \in I_3, b \notin I_2} \sum_{i \ge 1} (b_i^{a,b})^2 - \frac{1}{2} \sum_{a \notin I_3, b \in I_2} \sum_{i \ge 1} (b_i^{a,b})^2.$$

So all that remains is to check this sum agrees with the values we computed. First we will transform it into a expression in terms of the $M(\pi^{[a,b]}, \pi^{[c,d]})$. To do this we need the simple identities

$$\begin{aligned} M(\pi^{[a,b]}, \pi^{[c,d]}) - M((\pi^{[a,b]})', (\pi^{[c,d]})') &= \sum_{l \ge 1} \left(\sum_{i \ge l} b_i^{a,b} \cdot \sum_{i \ge l} b_i^{c,d} - \sum_{i \ge l} b_{i+1}^{a,b} \cdot \sum_{i \ge l} b_{i+1}^{c,d} \right) \\ &= \sum_{i \ge 1} b_i^{a,b} \cdot \sum_{i \ge 1} b_i^{c,d} \\ &= l(\pi^{[a,b]}) \cdot l(\pi^{[c,d]}) \end{aligned}$$

and

$$
\begin{aligned}
&M(\pi^{[a,b]}, \pi^{[a,b]}) - M((\pi^{[a,b]})', \pi^{[a,b]}) \\
&\qquad = \sum_{l\geq 1}\Bigl(\sum_{i\geq l} b_i^{a,b} \cdot \sum_{i\geq l} b_i^{a,b} - \sum_{i\geq l} b_{i+1}^{a,b} \cdot \sum_{i\geq l} b_i^{a,b}\Bigr) \\
&\qquad = \sum_{l\geq 1} b_l^{a,b} \cdot \sum_{i\geq l} b_i^{c,d} \\
&\qquad = \frac{1}{2} l(\pi^{[a,b]})^2 + \frac{1}{2}\sum_{l\geq 1}(b_l^{a,b})^2.
\end{aligned}
$$

Using these two identities and some simple algebraic manipulations, we can rewrite Proposition 4.18 as the statement that the difference

$$\sum_{0\leq a,b,c,d\leq N-1} T(\pi^{[a,b]}, \pi^{[c,d]}) - B(\pi^{[a,b]}, \pi^{[c,d]})$$

is equal to

$$
\begin{aligned}
&\sum_{i\in I_2} \sum_{\substack{b\neq i \\ c\neq i+1}} M(\pi^{[i+1,b]}, \pi^{[c,i]}) - M((\pi^{[i+1,b]})', (\pi^{[c,i]})') \\
&\quad + \sum_{i\in I_2} \sum_{\substack{b<d \\ b,d\neq i}} M((\pi^{[i+1,b]})', (\pi^{[i+1,d]})') - M(\pi^{[i+1,b]}, \pi^{[i+1,d]}) \\
&\quad + \sum_{i\in I_2} \sum_{\substack{a<c \\ a,c\neq i+1}} M((\pi^{[a,i]})', (\pi^{[c,i]})') - M(\pi^{[a,i]}, \pi^{[c,i]}) \\
&\quad + \sum_{\substack{a\in I_3, b\in I_2 \\ b\neq a-1}} M((\pi^{[a,b]})', (\pi^{[a,b]})') - M(\pi^{[a,b]}, \pi^{[a,b]}) \\
&\quad + \sum_{[a,b]\in S} M((\pi^{[a,b]})', \pi^{[a,b]}) - M(\pi^{[a,b]}, \pi^{[a,b]}).
\end{aligned}
$$

By a careful systematic approach, one shows that all these terms are exactly accounted for in Lemma A.14.

Acknowledgements

Morrison thanks his PhD supervisor Jim Bryan for his mathematical support over the last years. Both authors thank Sergey Mozgovoy for valuable discussion. Morrison is partially supported by a Four Year Doctoral Fellowship (4YF), University of British Columbia. Nagao was supported by the Grant-in-Aid for Research Activity Start-up (no. 22840023) and for Scientific Research (S) (no. 22224001).

Biographical note

by Prof. Hiraku Nakajima

Kentaro Nagao died on October 22, 2013, at age 31, from alveolar soft part sarcoma. He was a promising young mathematician, who started to make outstanding achievements in geometric representation theory. His too-early death was a sorrowful event and a big loss for mathematics.

Kentaro Nagao was born on April 18, 1982 in Tokyo, Japan. He first showed his mathematical ability by winning three gold medals (1998, 1999, 2000) and one silver (1997) at the International Mathematical Olympiad. He studied at Kyoto University, where he specialized in representation theory and geometry, and received a PhD in mathematics in 2008 under H. Nakajima. He then spent half a year at the University of Oxford, working with D. Joyce. In Spring 2010 he was appointed to an assistant professorship at Nagoya University. He was awarded the MSJ Takebe Katahiro Prize for his works on Donaldson–Thomas theory and cluster algebras in September 2013, one month before his death. He did not lose his enthusiasm for research under his difficult medical condition, as he expressed in response to the award. But his research was tragically stopped. We miss him.

References

[Behrend 2009] K. Behrend, "Donaldson–Thomas type invariants via microlocal geometry", *Ann. of Math.* (2) **170**:3 (2009), 1307–1338. MR 2011d:14098 Zbl 1191.14050

[Behrend and Fantechi 1997] K. Behrend and B. Fantechi, "The intrinsic normal cone", *Invent. Math.* **128**:1 (1997), 45–88. MR 98e:14022 Zbl 0909.14006

[Behrend and Fantechi 2008] K. Behrend and B. Fantechi, "Symmetric obstruction theories and Hilbert schemes of points on threefolds", *Algebra Number Theory* **2**:3 (2008), 313–345. MR 2009c:14006 Zbl 1170.14004

[Behrend et al. 2013] K. Behrend, J. Bryan, and B. Szendrői, "Motivic degree zero Donaldson–Thomas invariants", *Invent. Math.* **192**:1 (2013), 111–160. MR 3032328 Zbl 1267.14008

[Bryan and Morrison 2015] J. Bryan and A. Morrison, "Motivic classes of commuting varieties via power structures", *J. Algebraic Geom.* **24**:1 (2015), 183–199. MR 3275657 Zbl 06387936

[Chevalley et al. 1958] C. Chevalley, A. Grothendieck, and J.-P. Serre, *Séminaire C. Chevalley (2e année), III: Anneaux de Chow et applications*, Secrétariat Mathématique, Paris, 1958. MR 22 #1572 Zbl 0098.13101

[Denef and Loeser 2001] J. Denef and F. Loeser, "Geometry on arc spaces of algebraic varieties", pp. 327–348 in *European Congress of Mathematics, I* (Barcelona, 2000), edited by C. Casacuberta et al., Progr. Math. **201**, Birkhäuser, Basel, 2001. MR 2004c:14037 Zbl 1079.14003

[Derksen et al. 2010] H. Derksen, J. Weyman, and A. Zelevinsky, "Quivers with potentials and their representations, II: Applications to cluster algebras", *J. Amer. Math. Soc.* **23**:3 (2010), 749–790. MR 2012c:16044 Zbl 1208.16017

[Dimofte and Gukov 2010] T. Dimofte and S. Gukov, "Refined, motivic, and quantum", *Lett. Math. Phys.* **91**:1 (2010), 1–27. MR 2011m:14094 Zbl 1180.81112

[Getzler 1996] E. Getzler, "Mixed Hodge structures of configuration spaces", preprint, 1996. arXiv alg-geom/9510018

[Heinloth 2007] F. Heinloth, "A note on functional equations for zeta functions with values in Chow motives", *Ann. Inst. Fourier* (*Grenoble*) **57**:6 (2007), 1927–1945. MR 2009b:14041 Zbl 1154.14018

[Iqbal et al. 2009] A. Iqbal, C. Kozçaz, and C. Vafa, "The refined topological vertex", *J. High Energy Phys.* 10 (2009), 069, 58. MR 2011d:81266

[Joyce and Song 2012] D. Joyce and Y. Song, *A theory of generalized Donaldson–Thomas invariants*, vol. 217, Mem. Amer. Math. Soc. **1020**, Amer. Math. Soc., Providence, RI, 2012. MR 2951762 Zbl 1259.14054

[Keller and Yang 2011] B. Keller and D. Yang, "Derived equivalences from mutations of quivers with potential", *Adv. Math.* **226**:3 (2011), 2118–2168. MR 2012d:16043 Zbl 1272.13021

[Kontsevich and Soibelman 2008] M. Kontsevich and Y. Soibelman, "Stability structures, motivic Donaldson–Thomas invariants and cluster transformations", preprint, 2008. arXiv 0811:2435

[Looijenga 2002] E. Looijenga, "Motivic measures", exposé no. 874, 267–297 in *Séminaire Bourbaki* 1999/2000 (Exposé 865–879), Astérisque **276**, 2002. MR 2003k:14010 Zbl 0996.14011

[Macdonald 1995] I. G. Macdonald, *Symmetric functions and Hall polynomials*, 2nd ed., Clarendon Press, Oxford University Press, New York, 1995. MR 96h:05207 Zbl 0824.05059

[Maulik et al. 2006] D. Maulik, N. Nekrasov, A. Okounkov, and R. Pandharipande, "Gromov–Witten theory and Donaldson–Thomas theory, I", *Compos. Math.* **142**:5 (2006), 1263–1285. MR 2007i:14061 Zbl 1108.14046

[Merkurjev 2013] A. S. Merkurjev, "Essential dimension: a survey", *Transform. Groups* **18**:2 (2013), 415–481. MR 3055773 Zbl 1278.14066

[Morrison et al. 2012] A. Morrison, S. Mozgovoy, K. Nagao, and B. Szendrői, "Motivic Donaldson–Thomas invariants of the conifold and the refined topological vertex", *Adv. Math.* **230**:4-6 (2012), 2065–2093. MR 2927365 Zbl 1257.14028

[Nagao 2011a] K. Nagao, "Non-commutative Donaldson–Thomas theory and vertex operators", *Geom. Topol.* **15**:3 (2011), 1509–1543. MR 2012i:14069 Zbl 1219.14066

[Nagao 2011b] K. Nagao, "Refined open noncommutative Donaldson–Thomas invariants for small crepant resolutions", *Pacific J. Math.* **254**:1 (2011), 173–209. MR 2900667 Zbl 1245.14053

[Nagao 2011c] K. Nagao, "Wall-crossing of the motivic Donaldson–Thomas invariants", preprint, 2011. arXiv 1103.2922

[Nagao 2012] K. Nagao, "Derived categories of small toric Calabi–Yau 3-folds and curve counting invariants", *Q. J. Math.* **63**:4 (2012), 965–1007. MR 2999994 Zbl 1259.14017

[Nagao 2013] K. Nagao, "Donaldson–Thomas theory and cluster algebras", *Duke Math. J.* **162**:7 (2013), 1313–1367. MR 3079250 Zbl 06180315

[Nagao and Nakajima 2011] K. Nagao and H. Nakajima, "Counting invariant of perverse coherent sheaves and its wall-crossing", *Int. Math. Res. Not.* **2011**:17 (2011), 3885–3938. MR 2012h:14141 Zbl 1250.14021

[Okounkov et al. 2006] A. Okounkov, N. Reshetikhin, and C. Vafa, "Quantum Calabi–Yau and classical crystals", pp. 597–618 in *The unity of mathematics*, edited by P. Etingof et al., Progr. Math. **244**, Birkhäuser, Boston, 2006. MR 2006k:81342 Zbl 1129.81080

[Pandharipande and Thomas 2009] R. Pandharipande and R. P. Thomas, "Curve counting via stable pairs in the derived category", *Invent. Math.* **178**:2 (2009), 407–447. MR 2010h:14089 Zbl 1204.14026

[Segal 2008] E. Segal, "The A_∞ deformation theory of a point and the derived categories of local Calabi–Yaus", *J. Algebra* **320**:8 (2008), 3232–3268. MR 2009k:16016 Zbl 1168.18005

[Szendrői 2008] B. Szendrői, "Non-commutative Donaldson–Thomas invariants and the conifold", *Geom. Topol.* **12**:2 (2008), 1171–1202. MR 2009e:14100 Zbl 1143.14034

[Thomas 2000] R. P. Thomas, "A holomorphic Casson invariant for Calabi–Yau 3-folds, and bundles on $K3$ fibrations", *J. Differential Geom.* **54**:2 (2000), 367–438. MR 2002b:14049 Zbl 1034.14015

Communicated by Michel Van den Bergh
Received 2011-11-05 Revised 2012-04-26 Accepted 2015-03-27

andrewmo@math.ethz.ch *ETH Zurich, CH-8092 Zurich, Switzerland*

Kentaro Nagao *Graduate School of Mathematics, Nagoya University, Furu-cho, Chikusa-Ku, Nagoya 464-8602, Japan*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.815

Étale homotopy equivalence of rational points on algebraic varieties

Ambrus Pál

It is possible to talk about the étale homotopy equivalence of rational points on algebraic varieties by using a relative version of the étale homotopy type. We show that over p-adic fields rational points are homotopy equivalent in this sense if and only if they are étale-Brauer equivalent. We also show that over the real field rational points on projective varieties are étale homotopy equivalent if and only if they are in the same connected component. We also study this equivalence relation over number fields and prove that in this case it is finer than the other two equivalence relations for certain generalised Châtelet surfaces.

1. Introduction

For every field K, let $\overline{K}$ denote its separable closure. For every variety X defined over a K as above and every field extension $L|K$, let X_L denote the base change

MSC2010: primary 14F35; secondary 14G05.
Keywords: étale homotopy, rational points.

of X to L. Let K be a field, and let X be a variety defined over K. Harpaz and Schlank [2013] defined a relative version $\mathrm{Et}_{/K}(X)$ of the étale homotopy type of X by looking at the action of the absolute Galois group $\mathrm{Gal}(\overline{K}|K)$ of K on the étale hypercoverings of $X_{\overline{K}}$. With the aid of the action of $\mathrm{Gal}(\overline{K}|K)$ on $\mathrm{Et}_{/K}(X)$, they define a pro-object $X(hK)$ in the category of sets, which they call the homotopy fixed point set of X, that serves as a certain homotopical approximation of the set $X(K)$ of rational points. By slight abuse of notation we will use the same symbol to denote the projective limit of $X(hK)$, which we will consider as a topological space equipped with its pro-discrete topology. It is possible to define a natural map

$$\iota_{X/K} : X(K) \to X(hK),$$

which can be thought of as a homotopy-theoretic version of the section map in Grothendieck's anabelian geometry, which it also happens to refine. We say that $x, y \in X(K)$ are H-equivalent if $\iota_{X/K}(x) = \iota_{X/K}(y)$.

The aim of this paper is to describe the H-equivalence relation on $X(K)$ as explicitly as possible for many K and X. Let us first turn to the case when K is a finite extension of $\mathbb{Q}_p$. In this case, the map $\iota_{X/K}$ is not surjective in general; every abelian variety of positive dimension is a counterexample (see Proposition 9.7). However it is possible to describe the equivalence relation it induces on $X(K)$ in rather concrete terms. For any smooth variety X over any field K of characteristic zero, let $\mathrm{Br}(X) = H^2(X, \mathbb{G}_m)$ denote the cohomological Brauer group of X. We say that $x, y \in X(K)$ are Brauer equivalent if $x^*(b) = y^*(b)$ for all $b \in \mathrm{Br}(X)$. We say that $x, y \in X(K)$ are étale-Brauer equivalent if, for every finite, étale morphism $Y \to X$ of varieties over K and each $\tilde{x} \in Y(K)$ mapping to x, there is a $\tilde{y} \in Y(K)$ which maps to y and which is Brauer equivalent to $\tilde{x}$. Then we have the following:

Theorem 1.1. *Let K be a finite extension of $\mathbb{Q}_p$, and let X be a smooth quasi-projective variety over K. Then étale-Brauer equivalence and H-equivalence coincide on $X(K)$.*

It is important to note that this claim is not true for more general fields; Châtelet surfaces over number fields provide counterexamples (see the remark following Theorem 14.8 below). The main ingredients of the theorem above, besides obstruction theory, are duality for the Galois cohomology of K and Gabber's theorem on the existence of Azumaya algebras. We also provide examples to show that the theorem cannot be strengthened by substituting Brauer equivalence for étale-Brauer equivalence; see Theorem 9.5 below. We can also characterise H-equivalence for the field of real numbers:

Theorem 1.2. *Let K be the field $\mathbb{R}$ of real numbers, and let X be a smooth affine or projective variety over K. Then two K-rational points of X are H-equivalent if and only if they are in the same connected component of the topological space $X(K)$.*

The main tools of the proof of this result are a celebrated theorem of Mahé (see [Houdebine and Mahé 1982; Mahé 1982]), the theory of Stiefel–Whitney classes for quadratic bundles (see [Esnault et al. 1993; Milnor 1970]), and an equivariant version of a basic comparison result of Artin and Mazur [1969]. The reader should note that Quick [2011] developed a general theory of homotopy fixed point spaces for simplicial pro-sets equipped with a continuous action of a profinite group, which can be applied to Friedlander's étale topological type functor [1982]. His construction offers an alternative route for the foundations of our investigations.

Contents. In the next section, we review the relative étale homotopy type and homotopy fixed points of varieties as defined by Harpaz and Schlank and their relation to the Artin–Mazur construction. In the third section, we introduce a pointed version of the relative étale homotopy type and compare it with the previously defined constructions. In the fourth section, we study the relationship between the étale homotopy groups of finite étale coverings. In the following section, we show that the étale homotopy types of abelian varieties and smooth curves which are not projective of genus zero are Eilenberg–MacLane spaces over algebraically closed fields of characteristic zero. The fact presented in these two sections might be well-known to the experts, but we could not find a convenient reference. In the sixth section, we prove two useful lemmas about lifting a pair of rational points on certain principal bundles. Then we prove the fundamental theorem of obstruction theory for H-equivalence in the seventh section. We study the analogue of the Manin pairing for homotopy fixed points in the eighth section. In the ninth section, we prove that étale-Brauer equivalence is strictly finer than Brauer equivalence on $X(K)$ when K is a p-adic field and X is a bielliptic surface over K, using a rather standard set of tools. Theorem 1.1 is proved in the tenth section, while in the eleventh we prove Theorem 1.2. In the twelfth section, we introduce a natural homotopy version of Grothendieck's section and the Shafarevich–Tate conjectures over number fields by substituting the arithmetic fundamental group with the relative version of the étale homotopy type, which we call the homotopy section principle (HSP), and prove that it is equivalent to its well-established analogues in the special case of curves and abelian varieties. We provide further examples of varieties which satisfy HSP (see Theorems 13.3, 13.7 and 14.8) in the final two sections, including generalised Châtelet surfaces.

2. Basic definitions

Definition 2.1. Let Γ be a profinite group. By a Γ-set we mean a set with a Γ-action such that each element has an open stabiliser. Let $\Gamma\text{-Sets}$ denote the category whose objects are Γ-sets and whose morphisms are Γ-equivariant maps between them. By a simplicial Γ-set we mean a simplicial object in $\Gamma\text{-Sets}$. These form a category

Γ-SSets in the usual way. Note that, for every simplicial Γ-set S and every $n \in \mathbb{N}$, the n-skeleton $\mathrm{sk}_n(S)$, the n-coskeleton $\mathrm{cosk}_n(S)$ and the Kan replacement $\mathrm{Ex}^\infty(S)$ are all naturally equipped with a Γ-action. Since with respect to this action the stabiliser of each simplex of $\mathrm{sk}_n(S)$, $\mathrm{cosk}_n(S)$ and $\mathrm{Ex}^\infty(S)$ is open, these constructions furnish three functors: the n-skeleton $\mathrm{sk}_n : \Gamma\text{-SSets} \to \Gamma\text{-SSets}$, the n-coskeleton $\mathrm{cosk}_n : \Gamma\text{-SSets} \to \Gamma\text{-SSets}$, and the Kan replacement $\mathrm{Ex}^\infty : \Gamma\text{-SSets} \to \Gamma\text{-SSets}$ functors. Moreover, let $P_n : \Gamma\text{-SSets} \to \Gamma\text{-SSets}$ denote the corresponding analogue of the simplicial version of the n-th Postnikov piece given by the rule

$$P_n(S) = \mathrm{cosk}_{n+1}(\mathrm{sk}_{n+1}(\mathrm{Ex}^\infty S))$$

for every simplicial Γ-set S.

Notation 2.2. For every category $\mathscr{C}$ let Pro-$\mathscr{C}$ be the category of pro-objects of $\mathscr{C}$. For every pair of categories $\mathscr{C}, \mathscr{D}$ let $\mathscr{C} \times \mathscr{D}$ denote their direct product, and for every category $\mathscr{C}$ let $\mathscr{C}^{\mathrm{op}}$ denote its opposite category. Clearly there is a natural equivalence between $(\mathscr{C} \times \mathscr{D})^{\mathrm{op}}$ and $\mathscr{C}^{\mathrm{op}} \times \mathscr{D}^{\mathrm{op}}$; for the sake of simplicity we will not distinguish between these categories. We will consider every directed set, and in particular every ordered set, to be a category in the usual way.

Definition 2.3. Goerss [1995] constructs a model category structure on Γ-SSets, called the strict model structure. The corresponding homotopy category will be denoted by Ho(Γ-SSets) and will be called the homotopy category of simplicial Γ-sets. Similarly to the construction in Chapter 4 of [Artin and Mazur 1969], we may define a Postnikov tower functor

$$(\cdot)^\natural : \text{Pro-}\Gamma\text{-SSets} \to \text{Pro-}\Gamma\text{-SSets}$$

as follows: if I is a small filtering index category and $\Pi : I^{\mathrm{op}} \to \Gamma\text{-SSets}$ is a pro-object of Γ-SSets, then the functor $\Pi^\natural : I^{\mathrm{op}} \times \mathbb{N}^{\mathrm{op}} \to \Gamma\text{-SSets}$ is given by

$$\Pi^\natural(\alpha, n) = P_n(\Pi(\alpha)) \quad \text{for all } \alpha \in \mathrm{ob}(I) \text{ and } n \in \mathbb{N}.$$

We will denote by the same symbol the variant of the Postnikov tower functor in the category Pro-Ho(Γ-SSets), by the usual abuse of notation.

Definition 2.4. Next we recall the definition of the relative étale homotopy type, following [Harpaz and Schlank 2013]. Let K be a field, let $\Gamma_K = \mathrm{Gal}(\overline{K}|K)$ denote the absolute Galois group of K and let $\mathrm{Sch}_{/K}$ denote the category of locally Noetherian schemes over $\mathrm{Spec}(K)$. Let

$$\pi_{0/K} : \mathrm{Sch}_{/K} \to \Gamma_K\text{-Sets}$$

denote the functor which takes the K-scheme X to the Γ_K-set of connected components of $X_{\overline{K}}$. By applying this functor level-wise and composing it with the localisation functor $\Gamma_K\text{-SSets} \to \mathrm{Ho}(\Gamma_K\text{-SSets})$, we get a functor from the category of étale hypercoverings of the K-scheme X to the homotopy category of simplicial Γ_K-sets. This construction furnishes, similarly to what is done in Chapter 9 of [Artin and Mazur 1969], another functor

$$\mathrm{Et}_{/K} : \mathrm{Sch}_{/K} \to \text{Pro-Ho}(\Gamma_K\text{-}\mathrm{SSets}),$$

which we will call the relative étale homotopy type of X over K. Note that by functoriality we get a natural map

$$\iota_{X/K} : X(K) \to [\mathrm{Et}_{/K}(\mathrm{Spec}(K)), \mathrm{Et}_{/K}(X)] \to [\mathrm{Et}_{/K}(\mathrm{Spec}(K))^{\natural}, \mathrm{Et}_{/K}(X)^{\natural}],$$

where the second map is furnished by applying the Postnikov tower functor. We will call the pro-set $[\mathrm{Et}_{/K}(\mathrm{Spec}(K))^{\natural}, \mathrm{Et}_{/K}(X)^{\natural}]$ the homotopy fixed points of X and we will denote it by the symbol $X(hK)$.

Our next aim is to describe the target of this map more explicitly (and to justify the terminology which we have just introduced).

Definition 2.5. Let Sets and SSets denote the category of sets and the category of simplicial sets, respectively. Let Γ be as above. The category Γ-SSets is equipped with a natural concept of homotopy fixed points (see [Goerss 1995]); we will denote this functor $\Gamma\text{-}\mathrm{SSets} \to \mathrm{SSets}$ by $\Pi \mapsto \Pi^{h\Gamma}$. For every small filtering index category I and pro-object $\Pi : I^{\mathrm{op}} \to \mathrm{Ho}(\Gamma\text{-}\mathrm{SSets})$, we define the Γ-homotopy fixed point set of Π, denoted by $\Pi(E\Gamma)$, to be

$$\Pi(E\Gamma) = \lim_{\alpha \in \mathrm{ob}(I)} \pi_0(\Pi(\alpha)^{h\Gamma}).$$

We will frequently consider the limit $\Pi(E\Gamma)$ as a topological space via its natural pro-discrete topology. This structure is enough to reconstruct the underlying pro-set. By a formula of Harpaz and Schlank there is a natural identification

$$[(E\Gamma)^{\natural}, \Pi^{\natural}] = \Pi(E\Gamma),$$

where $E\Gamma$ is an analogue of the total space of the universal Γ-bundle in this setting (see [Harpaz and Schlank 2013, Definition 2.3]). They also show that when $\Gamma = \Gamma_K$ is the absolute Galois group of a field K then $\mathrm{Et}_{/K}(\mathrm{Spec}(K)) = E\Gamma_K$, hence we have $X(hK) = \mathrm{Et}_{/K}(X)(E\Gamma_K)$ for every K-scheme X, justifying our terminology.

Notation 2.6. Let Ho(SSets) denote the homotopy category of simplicial sets, let Sch denote the category of locally Noetherian schemes, and let

$$\mathrm{Et} : \mathrm{Sch} \to \text{Pro-Ho}(\mathrm{SSets})$$

denote the Artin–Mazur étale homotopy type functor. For every X as above and every $n \in \mathbb{N}$ let $\mathrm{Et}^n(X)$ denote the n-th Postnikov piece $P_n(\mathrm{Et}(X))$ and let $\mathrm{Et}(X)^\natural$ denote the Postnikov tower of $\mathrm{Et}(X)$. For every field K and every scheme X over K let $\overline{X}$ denote the base change $X_{\overline{K}}$. Moreover, for every such K, every $X \in \mathrm{ob}(\mathrm{Sch}_{/K})$ and every $n \in \mathbb{N}$, let $\mathrm{Et}^n_{/K}(X)$ denote the n-th Postnikov piece by $P_n(\mathrm{Et}_{/K}(X))$.

Lemma 2.7. *Let K be a field and X be a variety over K. Then there are natural isomorphisms*

$$f^n(X) : \mathrm{Et}^n(\overline{X}) \to \mathrm{Et}^n_{/K}(X) \quad \textit{and} \quad f^\natural(X) : \mathrm{Et}(\overline{X})^\natural \to \mathrm{Et}_{/K}(X)^\natural$$

in the category Pro-Ho(SSets).

Proof. The first half of the claim is Proposition 2.14 of [Harpaz and Schlank 2013]. The second half is an immediate consequence of the first half and the compatibility of the maps $f^n(X)$. □

Notation 2.8. For every $X \in \mathrm{ob}(\mathrm{Ho}(\mathrm{SSets}))$ let $X^\wedge \in \mathrm{ob}(\text{Pro-Ho}(\mathrm{SSets}))$ denote its profinite completion. The basic result about the homotopy type of complex algebraic varieties is the following classical theorem of Artin and Mazur:

Theorem 2.9. *Let X be a geometrically unibranch algebraic variety defined over $\mathbb{C}$. Then there is a canonical weak homotopy equivalence*

$$\eta_X : \mathrm{Et}(X)^\natural \to (X(\mathbb{C})^\wedge)^\natural$$

in Pro-Ho(SSets).

Proof. This is [Artin and Mazur 1969, Corollary 12.10, p. 143]. □

Proposition 2.10. *Assume that K is algebraically closed. Then $\iota_{X/K}$ is surjective. Two points $x, y \in X(K)$ are H-equivalent if and only if they lie in the same connected component of X with respect to the Zariski topology.*

Proof. When K is algebraically closed its absolute Galois group is trivial. Therefore $\mathrm{Ho}(\Gamma_K\text{-SSets})$ is the homotopy category of simplicial sets and $\mathrm{Et}_{/K}(X)$ is just the usual Artin–Mazur étale homotopy type of X. Because K is algebraically closed, $\mathrm{Et}_{/K}(\mathrm{Spec}(K))$ is contractible. Therefore there is a natural bijection

$$X(hK) \cong [\,*\,, \mathrm{Et}_{/K}(X)^\natural] \cong \pi_0(\mathrm{Et}_{/K}(X)^\natural) \cong \pi_0(X),$$

where the third identification is the consequence of a fundamental comparison theorem of Artin and Mazur [1969, Corollary 10.8, p. 122]. The claim now follows from the naturality of $\iota_{X/K}$. □

3. The pointed relative étale homotopy type

Definition 3.1. Let Γ be as above. By a pointed simplicial Γ-set we mean a simplicial Γ-set S_* with a point $p \in S_0$ fixed by Γ. These form a category, Γ-SSets_*, in the usual way. Note that, for every pointed simplicial Γ-set S and every $n \in \mathbb{N}$, the n-skeleton $\mathrm{sk}_n(S)$, the n-coskeleton $\mathrm{cosk}_n(S)$, the Kan replacement $\mathrm{Ex}^\infty(S)$ and n-th Postnikov piece $P_n(S)$ are all naturally equipped with a point fixed by Γ, and we will denote the corresponding four functors, the n-skeleton $\mathrm{sk}_n : \Gamma\text{-}\mathrm{SSets}_* \to \Gamma\text{-}\mathrm{SSets}_*$, the n-coskeleton $\mathrm{cosk}_n : \Gamma\text{-}\mathrm{SSets}_* \to \Gamma\text{-}\mathrm{SSets}_*$, the Kan replacement $\mathrm{Ex}^\infty : \Gamma\text{-}\mathrm{SSets}_* \to \Gamma\text{-}\mathrm{SSets}_*$ and the n-th Postnikov piece $P_n : \Gamma\text{-}\mathrm{SSets}_* \to \Gamma\text{-}\mathrm{SSets}_*$ by the same symbols by a slight abuse of notation.

Definition 3.2. The homotopy category of Γ-SSets_* with respect to the pointed version of weak equivalences of Goerss' strict model structure, called the homotopy category of pointed simplicial Γ-sets, will be denoted by $\mathrm{Ho}(\Gamma\text{-}\mathrm{SSets}_*)$. Similarly to the construction recalled in Definition 2.3, we may define a Postnikov tower functor

$$(\,\cdot\,)^\natural : \mathrm{Pro}\text{-}\Gamma\text{-}\mathrm{SSets}_* \to \mathrm{Pro}\text{-}\Gamma\text{-}\mathrm{SSets}_*,$$

and we will denote by the same symbol the corresponding Postnikov tower functor in the category $\mathrm{Pro}\text{-}\mathrm{Ho}(\Gamma\text{-}\mathrm{SSets}_*)$ by the usual abuse of notation. This is of course justified as the formations of these invariants commute with the forgetful functor $\mathrm{Pro}\text{-}\mathrm{Ho}(\Gamma\text{-}\mathrm{SSets}_*) \to \mathrm{Pro}\text{-}\mathrm{Ho}(\Gamma\text{-}\mathrm{SSets})$.

Definition 3.3. Since Γ-SSets is a model category, it has all colimits, in particular pushouts and equalisers. If Y_* is a simplicial Γ-set and $X_* \subseteq Y_*$ is a subsimplicial Γ-set, then let Y_*/X_* denote the simplicial Γ-set which is the pushout of the inclusion map $X_* \to Y_*$. We call Y_*/X_* the contraction of Y_* by X_*. If X_*, Y_* are simplicial Γ-sets and $f_* : X_* \to Y_*$, $g_* : X_* \to Y_*$ are maps of simplicial Γ-sets, then let $Y_*(f_* = g_*)$ denote the simplicial Γ-set which is the coequaliser of f_* and g_*.

Definition 3.4. Note that, for every pair of simplicial Γ-sets X_* and Y_*, the product $X_* \times Y_*$ equipped with the natural (diagonal) Γ-action is also a simplicial Γ-set. Similarly the coproduct (disjoint union) $X_* \sqcup Y_*$ of simplicial Γ-sets X_* and Y_* is also a simplicial Γ-set equipped with the tautological Γ-action. Let I denote the 1-simplex $\Delta^1 = [0, 1]$ with the trivial Γ-action; this choice makes it into a simplicial Γ-set. For every morphism $f : X_* \to Y_*$ of simplicial Γ-sets the mapping cylinder $\mathrm{Cyl}(f)$ is the coequaliser of the two maps

$$f' : X \to Y_* \sqcup X_* \times I \quad \text{and} \quad p : X \to Y_* \sqcup X_* \times I,$$

where f' is the composition of f and the tautological inclusion $Y_* \subseteq Y_* \sqcup X_* \times I$ and p is the composition of the map identifying X_* with $X_* \times \{1\} \subseteq X_* \times I$ and the tautological inclusion $X_* \times I \subset Y_* \sqcup X_* \times I$. We define the mapping cone $\mathrm{Cone}(f)$ of an $f : X_* \to Y_*$ as above as the contraction of $\mathrm{Cyl}(f)$ by the image of the map

$$q : X_* \to \mathrm{Cyl}(f),$$

where q is the composition of the map identifying X_* with $X_* \times \{0\} \subseteq X_* \times I$ with the tautological inclusion $X_* \times I \subset Y_* \sqcup X_* \times I$ composed with the natural surjection $Y_* \sqcup X_* \times I \mapsto \mathrm{Cyl}(f)$. Note that $\mathrm{Cone}(f)$ is canonically a pointed simplicial Γ-set, where the base point is the image of $q(X_0) \subset \mathrm{Cyl}(f)_0$ under the contraction map $\mathrm{Cyl}(f) \to \mathrm{Cone}(f)$.

Definition 3.5. By a pointed K-scheme (X, x) we will mean a locally Noetherian scheme X over K with a K-valued point $x : \mathrm{Spec}(K) \to X$ on X. These form the objects of a category $\mathrm{Sch}_{/K*}$, where a morphism f from an object (X, x) to another object (Y, y) is a map $f : X \to Y$ of schemes over K such that $f(x) = y$. Now let (X, x) be a pointed K-scheme and let H_* be an étale hypercovering of X. Then the pullback $x^*(H_*)$ is an étale hypercovering of $\mathrm{Spec}(K)$, and the map x induces a morphism $x_*(H_*) : \pi_{0/K}(x^*(H_*)) \to \pi_{0/K}(H_*)$ of simplicial Γ-sets. Let $\pi_{0/K}(H_*, x)$ denote the mapping cone of the composition of this map $x_*(H_*)$ and the canonical inclusion $\pi_{0/K}(H_*) \subseteq \mathrm{Ex}^\infty(\pi_{0/K}(H_*))$; it is a pointed simplicial Γ-set. A map $f : H_* \to J_*$ between étale hypercoverings of X induces a map $\pi_{0/K}(f, x) : \pi_{0/K}(H_*, x) \to \pi_{0/K}(J_*, x)$ between pointed simplicial Γ-sets, and a homotopy between two maps $f : H_* \to J_*, g : H_* \to J_*$ induces a pointed Γ-equivariant homotopy between $\pi_{0/K}(f, x)$ and $\pi_{0/K}(g, x)$. Therefore we may apply Corollary 8.13(i) of [Artin and Mazur 1969, p. 105] to conclude that the functor

$$H_* \mapsto \pi_{0/K}(H_*, x)$$

above induces an object $\mathrm{Et}_{/K}(X, x)$ of $\text{Pro-Ho}(\Gamma\text{-SSets}_*)$. We will call the latter the pointed relative étale homotopy type of (X, x).

Notation 3.6. Let (X, x) be a pointed K-scheme. For every étale hypercovering H_* of X let

$$i(H_*, x) : \pi_{0/K}(H_*) \to \pi_{0/K}(H_*, x)$$

be the composition of the natural inclusion map

$$\pi_{0/K}(H_*) \to \mathrm{Cone}(x_*(H_*))$$

and the map $\mathrm{Cone}(x_*(H_*)) \to \pi_{0/K}(H_*, x)$ induced by the functoriality of mapping cones. This is a natural transformation between two functors from the category of étale hypercoverings over X into Γ-SSets, and hence it induces a map

$$i(X,x):\mathrm{Et}_{/K}(X,x)\to\mathrm{Et}_{/K}(X)$$

of pro-objects of the homotopy category Ho(Γ-SSets), where by slight abuse of notation we let $\mathrm{Et}_{/K}(X,x)$ also denote the image of the pointed relative étale homotopy type of (X,x) with respect to the forgetful functor

$$\text{Pro-Ho}(\Gamma\text{-SSets}_*)\to\text{Pro-Ho}(\Gamma\text{-SSets}).$$

The map $i(X,x)$ is obviously natural.

Proposition 3.7. *The map* $i(X,x):\mathrm{Et}_{/K}(X,x)\to\mathrm{Et}_{/K}(X)$ *induces a bijection on homotopy fixed points.*

Proof. Note that the functor $\pi_{0/K}$ induces an equivalence between the category of étale coverings over Spec(K) and Γ-Sets, so $\pi_{0/K}(x^*(H_*))$ is a contractible simplicial Γ-set. Therefore $i(H_*,x)$ is a weak equivalence for every étale hypercovering H_* of X with respect to Goerss' *weak* model structure (see Theorem A on p. 189 and Definition 1.11 on p. 194 in [Goerss 1995]), and hence induces a bijection $\pi_0(\pi_{0/K}(H_*)^{h\Gamma})\to\pi_0(\mathrm{Cone}(x_*(H_*))^{h\Gamma})$. So the same holds for $i(X,x)$, too. □

Definition 3.8. Let Sch_* be the category of pointed locally Noetherian schemes, and, as usual, denote the objects of Sch_* by pairs (X,x), where X is a locally Noetherian scheme and x is a geometric point of X. By slight abuse of notation let

$$\mathrm{Et}:\mathrm{Sch}_*\to\text{Pro-Ho}(\mathrm{SSets}_*)$$

denote the pointed version of the Artin–Mazur étale homotopy type functor. For every object (X,x) of Sch_* and every $n\geq 1$, let $\pi_n(X,x)$ denote the n-th homotopy group of $\mathrm{Et}(X,x)$ when X is connected.

Notation 3.9. For every pointed scheme (X,x) and every $n\in\mathbb{N}$, let $\mathrm{Et}^n(X,x)$ denote the n-th Postnikov piece $P_n(\mathrm{Et}(X,x))$ and let $\mathrm{Et}(X,x)^\natural$ denote the Postnikov tower of $\mathrm{Et}(X,x)$. Similarly, for every field K, every pointed K-scheme (X,x) and every $n\in\mathbb{N}$, let $\mathrm{Et}^n_{/K}(X,x)$ denote the n-th Postnikov piece $P_n(\mathrm{Et}_{/K}(X,x))$ and let $\mathrm{Et}_{/K}(X,x)^\natural$ denote the Postnikov tower of $\mathrm{Et}_{/K}(X,x)$. Since we fixed a separable closure $\overline{K}$ of K, we may associate to every K-valued point $x:\mathrm{Spec}(K)\to X$ of a K-scheme X a $\overline{K}$-valued point $\overline{x}:\mathrm{Spec}(K)\to X$ which is the composition of the map $\mathrm{Spec}(\overline{K})\to\mathrm{Spec}(K)$ induced by the inclusion $K\subseteq\overline{K}$ and x.

Similarly to above, by slight abuse of notation we let $\mathrm{Et}_{/K}(X,x)$ and $\mathrm{Et}^n_{/K}(X,x)$ also denote respectively the image of the pointed relative étale homotopy type of (X,x) and of the n-th truncation of the latter with respect to the forgetful functor

$$\text{Pro-Ho}(\Gamma\text{-SSets}_*)\to\text{Pro-Ho}(\mathrm{SSets}_*).$$

Proposition 3.10. *Let K be a field and (X,x) a pointed K-scheme such that X is a variety over K. Then there are natural isomorphisms*

$$f^n(X,x):\mathrm{Et}^n(\overline{X},\overline{x})\to\mathrm{Et}^n_{/K}(X,x)\quad\textit{and}\quad f^{\natural}(X,x):\mathrm{Et}(\overline{X},\overline{x})^{\natural}\to\mathrm{Et}_{/K}(X,x)^{\natural}$$

in the category Pro-Ho(SSets$_*$).

Proof. Since the second half is an immediate consequence of the first half and the compatibility of the maps $f^n(X,x)$, it will be enough to prove the former. Let Δ^0 denote the 0-simplex, as usual. Let (H_*,h) be a pointed étale hypercovering of $(X,\overline{x})$. By definition h is a map $\Delta^0\to\pi_{0/K}(x^*(H_*))$ of simplicial sets. For every (H_*,h) as above let $\pi_{0/K}(H_*,h,x)$ be the contraction of $\pi_{0/K}(H_*,x)$ by the image of the map

$$c_h:\Delta^1\cong\Delta^0\times\Delta^1\to\pi_{0/K}(H_*,x)$$

of simplicial sets, where c_h is the composition $\pi\circ\iota\circ(h\times\mathrm{id}_I)$, where ι is the inclusion

$$\pi_{0/K}(x^*(H_*))\times\Delta^1\subset\mathrm{Ex}^{\infty}(\pi_{0/K}(H_*))\sqcup\pi_{0/K}(x^*(H_*))\times\Delta^1$$

and π is the canonical surjection

$$\mathrm{Ex}^{\infty}(\pi_{0/K}(H_*))\sqcup\pi_{0/K}(x^*(H_*))\times\Delta^1\to\pi_{0/K}(H_*,x).$$

Let

$$a(H_*,h):\pi_{0/K}(H_*,x)\to\pi_{0/K}(H_*,h,x)$$

be the contraction map. We will consider $\pi_{0/K}(H_*,h,x)$ a pointed simplicial set, where its distinguished point $b(h)\in\pi_{0/K}(H_*,h,x)_0$ is the image of the base point of the pointed simplicial Γ-set under $a(H_*,h)$. Note that this map is a weak equivalence in Pro $-$ Ho(SSets$_*$), since we contracted a contractible subsimplicial set. Therefore by [Artin and Mazur 1969, Corollary 8.13(i), p. 105] the functor

$$(H_*,h)\mapsto(\pi_{0/K}(H_*,h,x),b(h))$$

induces an object $\mathrm{Et}'(\overline{X},\overline{x})$ of Pro-Ho(SSets$_*$) which is isomorphic to $\mathrm{Et}_{/K}(X,x)$ in this category. For every (H_*,h) as above let $\pi_{0/K}(h)\in\pi_{0/K}(H_0)$ be the point corresponding to h, and let

$$b(H_*,h):\pi_{0/K}(H_*)\to\pi_{0/K}(H_*,x)$$

be the map of pointed simplicial sets which is the composition of the natural inclusion map

$$\pi_{0/K}(H_*)\subset\pi_{0/K}(H_*)\sqcup\pi_{0/K}(x^*(H_*))\times\Delta^1$$

with the natural surjection

$$\pi_{0/K}(H_*)\sqcup\pi_{0/K}(x^*(H_*))\times\Delta^1\to\mathrm{Cone}(x_*)$$

composed with the map $\mathrm{Cone}(x_*) \to \pi_{0/K}(H_*, x)$ induced by the functoriality of mapping cones. Let $c(H_*, h)$ be the composition of $b(H_*, h)$ with $a(H_*, h)$; then this map is a morphism

$$c(H_*, h) : (\pi_{0/K}(H_*), \pi_{0/K}(h)) \to (\pi_{0/K}(H_*, h, x), b(h))$$

of pointed simplicial sets. Since $\pi_{0/K}(x^*(H_*))$ is a contractible simplicial Γ-set, the map $c(H_*, h)$ is a weak equivalence for every pointed étale hypercovering (H_*, h) of X. Therefore by [Artin and Mazur 1969, Corollary 8.13(i), p. 105] the functor

$$(H_*, h) \mapsto (\pi_{0/K}(H_*), \pi_{0/K}(h))$$

induces an object $\mathrm{Et}''(\overline{X}, \overline{x})$ of Pro-Ho(SSets$_*$) which is isomorphic to $\mathrm{Et}'(\overline{X}, \overline{x})$, and hence to $\mathrm{Et}_{/K}(X, x)$ in this category. Therefore it will be sufficient to prove that there are natural isomorphisms

$$g^n(X, x) : \mathrm{Et}_n(\overline{X}, \overline{x}) \to P_n(\mathrm{Et}''(\overline{X}, \overline{x}))$$

in the category Pro-Ho(SSets$_*$) which are compatible with each other and with truncation.

Note that the inclusion of the indexing category of $\mathrm{Et}''(\overline{X}, \overline{x})$ in the indexing category of $\mathrm{Et}(\overline{X}, \overline{x})$ furnishes a natural map

$$g(X, x) : \mathrm{Et}(\overline{X}, \overline{x}) \to \mathrm{Et}''(\overline{X}, \overline{x}).$$

In order to prove that $g(X, x)$ is an isomorphism after taking n-th truncations, we can argue the same way as in the proof of [Harpaz and Schlank 2013, Proposition 2.14]. □

4. Homotopy groups of finite étale covers

Recall that for every object X of Sch with a geometric point x and every $n \in \mathbb{N}$ the symbol $\pi_n(X, x)$ denotes the homotopy group $\pi_n(\mathrm{Et}(X), x)$. In this section assume that K is an algebraically closed field.

Proposition 4.1. *Let $f : (X, x) \to (Y, y)$ be a finite étale map of pointed smooth connected quasiprojective varieties over K. Then the induced map*

$$\pi_n(f) : \pi_n(X, x) \to \pi_n(Y, y)$$

is an isomorphism for every $n \geq 2$.

Proof. Because $\pi_1(Y, y)$ is topologically finitely generated, its open normal subgroups are cofinal, and hence there is a finite étale map $g : (Z, z) \to (X, x)$ of pointed smooth connected quasiprojective varieties over K such that the image of $\pi_1(f \circ g) : \pi_1(Z, z) \to \pi_1(Y, y)$ is an open normal subgroup. In this case the image of $\pi_1(f): \pi_1(Z,z) \to \pi_1(X,x)$ is an open normal subgroup, too. It will be enough to show that the maps $\pi_n(f \circ g)$ and $\pi_n(f)$ are isomorphisms for every $n \geq 2$. Since the composition $f \circ g$ is also a finite étale map, we've reduced the claim to the special case when the image of $\pi_1(f)$ is an open normal subgroup.

In this case f is a finite Galois covering; let G denote the covering group and let $\alpha: G \to \mathrm{Aut}(X)$ be the action corresponding to the deck transformations. Let (H_*, h) be a pointed hypercovering of Y with respect to the étale site of Y pointed with y. Let C_* denote the étale Čech hypercovering

$$X \;\substack{\longrightarrow\\ \longleftarrow\\ \longleftarrow}\; X \times_Y X \;\substack{\longleftarrow\\ \longleftarrow\\ \longleftarrow}\; X \times_Y X \times_Y X \;\substack{\longleftarrow\\ \longleftarrow\\ \longleftarrow\\ \longleftarrow}\; \cdots$$

generated by the cover $X \to Y$, and equip it with a point c with respect to the same pointed site. Let (I_*, i) be the fibre product of (H_*, h) and (C_*, c) over Y; this pointed simplicial object is also a hypercovering. Let (J_*, j) be the pullback $f^*(I_*, i)$ of the pointed hypercovering (I_*, i) onto the étale site of X pointed with x with respect to f and let $f_*: (\pi_0(J_*), \pi_0(j)) \to (\pi_0(I_*), \pi_0(i))$ be the map induced by f between the pointed simplicial sets of connected components of J_* and I_*.

The action α induces an action of G on J_*, and hence an action of G on $\pi_0(J_*)$. If we equip $\pi_0(I_*)$ with the trivial action then f_* is G-equivariant. Because f is a finite étale cover the map f_* is surjective. Moreover, for every n the étale cover $I_n \to Y$ factors through $f: X \to Y$, and hence the action of G on the connected components of the base change $J_n \to X$ of this map to X is free. We get that $f_*: (\pi_0(J_*), \pi_0(j)) \to (\pi_0(I_*), \pi_0(i))$ is a G-cover and hence the induced maps

$$\pi_n(f_*): \pi_n(\pi_0(J_*), \pi_0(j)) \to \pi_n(\pi_0(I_*), \pi_0(i))$$

are isomorphisms for all $n \geq 2$.

Let (L_*, l) be a pointed hypercovering of X with respect to the étale site of X pointed with x. By composing the structure maps with f we get a pointed hypercovering of Y with respect to the étale site of Y pointed with y, which we will denote by (H_*, h) by slight abuse of notation. By applying the same construction to (H_*, h) as above we get a pointed hypercovering (J_*, j) of the étale site of X pointed with x which dominates (L_*, l). Therefore pointed hypercovers of X of the form as (J_*, j) above are cofinal, so the injectivity of the maps $\pi_n(f)$ for all $n \geq 2$ follows.

Let γ be an element of $\pi_n(Y, y)$ (where $n \geq 2$). For every pointed hypercovering (L_*, l) of X with respect to the étale site of X pointed with x we will construct an element $\gamma'_{(L_*, l)} \in \pi_n(\pi_0(L_*), \pi_0(l))$ as follows. Let (H_*, h), (I_*, i) and (J_*, j) be the same as in the paragraph above. Let $\gamma_{(I_*, i)} \in \pi_n(\pi_0(I_*), \pi_0(i))$ be the image of γ under the tautological map $\pi_n(Y, y) \to \pi_n(\pi_0(I_*), \pi_0(i))$, let $\pi_n(f_*)^{-1}(\gamma_{(I_*, i)})$

be the unique preimage of $\gamma_{(I_*,i)}$ under the isomorphism

$$\pi_n(f_*):\pi_n(\pi_0(J_*),\pi_0(j))\to\pi_n(\pi_0(I_*),\pi_0(i)),$$

and let $\gamma'_{(L_*,l)}$ be the image of $\pi_n(f_*)^{-1}(\gamma_{(I_*,i)})$ under the natural map

$$\pi_n(\pi_0(J_*),\pi_0(j))\to\pi_n(\pi_0(L_*),\pi_0(l)).$$

It is easy to check that the elements $\gamma'_{(L_*,l)}$ glue together to an element γ' of $\pi_n(X,x)$ whose image under $\pi_n(f)$ is γ. The surjectivity of the maps $\pi_n(f)$ for all $n\geq 2$ follows. $\square$

Notation 4.2. For every group Γ let $\widehat{\Gamma}$ be its profinite completion. For every object X of Sch, for every $n\in\mathbb{N}$ and for every pro-abelian group A let $H_n(X,A)$ denote the homology group $H_n(\mathrm{Et}(X),A)$. For every object (X,x) of Sch_* and n as above let

$$h_n(X,x):\pi_n(X,x)\to H_n(X,\widehat{\mathbb{Z}})$$

denote the Hurewicz map. Let X be a smooth, geometrically irreducible, quasiprojective variety over K. Let x be a K-valued point of X and let $\mathrm{Fet}(X,x)$ denote the category of finite étale pointed connected covers (Y,y) of (X,x) such that the image of the induced map $\pi_1(f):\pi_1(Y,y)\to\pi_1(Y,y)$ is an open characteristic subgroup for every object $f:(Y,y)\to(X,x)$. Since for every $f:(Y,y)\to(X,x)$ as above the induced map $\pi_2(f):\pi_2(Y,y)\to\pi_2(X,x)$ is an isomorphism by Proposition 4.1, the projective limit of the inverses of these maps is an isomorphism

$$a_X:\pi_2(X,x)\to\lim_{(Y,y)\in\mathrm{Fet}(X,x)}\pi_2(Y,y).$$

Moreover, we may take the projective limit of the Hurewicz maps

$$b_X:\lim_{(Y,y)\in\mathrm{Fet}(X,x)}\pi_2(Y,y)\to\lim_{(Y,y)\in\mathrm{Fet}(X,x)}H_2(Y,\widehat{\mathbb{Z}}).$$

Theorem 4.3. *The map*

$$b_X\circ a_X:\pi_2(X,x)\to\lim_{(Y,y)\in\mathrm{Fet}(X,x)}H_2(Y,\widehat{\mathbb{Z}})$$

is an isomorphism.

Proof. First we are going to prove that b_X is injective. Let γ be a nonzero element of $\pi_2(X,x)$. Then there is a pointed hypercovering (H_*,h) of X with respect to the étale site of X pointed with x such that the image of γ under the natural map $\pi_2(X,x)\to\pi_2(\pi_0(H_*),\pi_0(h))$ is nonzero. Let N be the kernel of the natural map $\pi_1(X,x)\to\pi_1(\pi_0(H_*),\pi_0(h))$; it is an open normal subgroup. Because $\pi_1(X,x)$ is topologically finitely generated, its open characteristic subgroups are cofinal, and hence there is an object $f:(Y,y)\to(X,x)$ of $\mathrm{Fet}(X,x)$ such that the image of $\pi_1(f):\pi_1(Y,y)\to\pi_1(X,x)$ lies in N.

Note that f is a finite Galois covering; let G denote the covering group and let $\alpha : G \to \mathrm{Aut}(X)$ be the action corresponding to the deck transformations. Let C_* denote the étale Čech hypercovering generated by the cover $Y \to X$, and equip it with a point c with respect to the same pointed site. Let (I_*, i) be the pointed simplicial object which is the fibre product of (H_*, h) and (C_*, c) and let (J_*, j) be the pullback $f^*(I_*, i)$ of the pointed hypercovering (I_*, i) onto the étale site of X pointed with x with respect to f. Let $f_* : (\pi_0(J_*), \pi_0(j)) \to (\pi_0(I_*), \pi_0(i))$ be the map induced by f between the pointed simplicial sets of connected components of J_* and I_*.

We may argue as above to conclude that $f_* : (\pi_0(J_*), \pi_0(j)) \to (\pi_0(I_*), \pi_0(i))$ is a G-cover with respect to the action induced by α on $\pi_0(J_*)$ and the trivial action on $\pi_0(I_*)$. Therefore the induced map

$$\pi_1(f_*) : \pi_1(\pi_0(J_*), \pi_0(j)) \to \pi_1(\pi_0(I_*), \pi_0(i))$$

is injective. There is a commutative diagram

$$\begin{tikzcd}
\pi_1(Y, y) \arrow[r] \arrow[d, "\pi_1(f)"'] & \pi_1(\pi_0(J_*), \pi_0(j)) \arrow[d, "\pi_1(f_*)"'] \\
\pi_1(X, x) \arrow[r] & \pi_1(\pi_0(I_*), \pi_0(i)).
\end{tikzcd}$$

By assumption the composition of $\pi_1(f)$ and the lower horizontal map has trivial image. Since the upper horizontal map is surjective by [Artin and Mazur 1969, Corollary 10.6, pp. 121–122], we get that $\pi_1(f_*)$ has trivial image, too, and hence $\pi_1(\pi_0(J_*), \pi_0(j))$ is the trivial group. There is a similar commutative diagram

$$\begin{tikzcd}
\pi_2(Y, y) \arrow[r] \arrow[d, "\pi_2(f)"'] & \pi_2(\pi_0(J_*), \pi_0(j)) \arrow[d, "\pi_2(f_*)"'] \\
\pi_2(X, x) \arrow[r] & \pi_2(\pi_0(I_*), \pi_0(i))
\end{tikzcd}$$

for π_2. The image of γ, considered as an element of $\pi_2(Y, y)$, is nonzero under the composition of $\pi_2(f)$ and the lower horizontal map by assumption. Therefore its image γ' under the upper horizontal map is also nonzero. Since $\pi_1(\pi_0(J_*), \pi_0(j))$ is trivial, we get that the image of γ' under the Hurewicz map $\pi_2(\pi_0(J_*), \pi_0(j)) \to H_2(\pi_0(J_*), \widehat{\mathbb{Z}})$ is nonzero. By naturality this implies that the image of γ under the Hurewicz map $h_2(Y, y) : \pi_2(Y, y) \to H_2(Y, \widehat{\mathbb{Z}})$ is nonzero, too.

Next we are going to prove that b_X is surjective. For every $(Z, z) \in \mathrm{Fet}(X, x)$ let $\mathcal{C}(Z, z)$ be the category of morphisms $(Y, y) \to (Z, z)$ in $\mathrm{Fet}(X, x)$, and let

$$c_{(Z,z)} : \lim_{(Y,y)\to(Z,z)\in\mathcal{C}(Z,z)} H_2(Y, \widehat{\mathbb{Z}}) \to H_2(Z, \widehat{\mathbb{Z}})$$

be the tautological map. Because the preimage $(b_X \circ a_X)^{-1}(\mathrm{Im}(c_{(Z,z)})) \subseteq \pi_2(X,x)$ is closed and $\pi_2(X,x)$ is profinite, by compactness it will be enough to show that for every object (Z,z) as above $(b_X \circ a_X)^{-1}(\mathrm{Im}(c_{(Z,z)}))$ is nonempty. Now fix an element $\gamma \in \mathrm{Im}(c_{(Z,z)})$ and choose a

$$\gamma' \in \lim_{(Y,y)\to(Z,z)\in\mathscr{C}(Z,z)} H_2(Y,\widehat{\mathbb{Z}})$$

such that $c_{(Z,z)}(\gamma') = \gamma$. For every pointed hypercovering (H_*,h) of Z with respect to the étale site of Z pointed with z, we are going to construct an element $\gamma_{(H_*,h)} \in \pi_2(\pi_0(H_*),\pi_0(h))$ as follows.

Let N be the kernel of the natural map $\pi_1(Z,z) \to \pi_1(\pi_0(H_*),\pi_0(h))$. Using that $\pi_1(Z,z)$ is topologically finitely generated as above, we get that there is a morphism $f:(Y,y) \to (Z,z)$ of $\mathrm{Fet}(X,x)$ such that the image of $\pi_1(f): \pi_1(Y,y) \to \pi_1(Z,z)$ lies in N. Let C_* denote the étale Čech hypercovering generated by the cover $Y \to Z$, and equip it with a point c with respect to the same pointed site. Let (I_*,i) be the pointed hypercovering which is the fibre product of (H_*,h) and (C_*,c), and let (J_*,j) be the pullback $f^*(I_*,i)$ of (I_*,i) onto the étale site of Y pointed with y with respect to f. Let $f_*:(\pi_0(J_*),\pi_0(j)) \to (\pi_0(I_*),\pi_0(i))$ be the map induced by f between the pointed simplicial sets of connected components of J_* and I_*.

As we saw in the proof of injectivity, the group $\pi_1(\pi_0(J_*),\pi_0(j))$ is trivial and hence the Hurewicz map $\pi_2(\pi_0(J_*),\pi_0(j)) \to H_2(\pi_0(J_*),\widehat{\mathbb{Z}})$ is an isomorphism. Therefore there is a unique $\gamma_f \in \pi_2(\pi_0(J_*),\pi_0(j))$ whose image under this Hurewicz map is the image of γ' under the composition of $c_{(Y,y)}$ and the natural map $H_2(Y,\widehat{\mathbb{Z}}) \to H_2(\pi_0(J_*),\widehat{\mathbb{Z}})$. Let $\gamma_{(H_*,h)} \in \pi_2(\pi_0(H_*),\pi_0(h))$ be the image of γ_f under the composition of the functorial maps $\pi_2(f_*):\pi_2(\pi_0(J_*),\pi_0(j)) \to \pi_2(\pi_0(I_*),\pi_0(i))$ and $\pi_2(\pi_0(I_*),\pi_0(i)) \to \pi_2(\pi_0(H_*),\pi_0(h))$.

First we are going to show that $\gamma_{(H_*,h)}$ is independent of the choice of the morphism f. Let $f':(Y',y') \to (Z,z)$ be another morphism of $\mathrm{Fet}(X,x)$ such that the image of $\pi_1(f'):\pi_1(Y',y') \to \pi_1(Z,z)$ lies in N. Then the fibre product of f and f' over (Z,z) is another morphism of $\mathrm{Fet}(X,x)$ with this property. Moreover, it can be factorised into a composition of a morphism g of $\mathrm{Fet}(X,x)$ and f, and into a composition of a morphism g' of $\mathrm{Fet}(X,x)$ and f' too. Therefore it will be enough to show that this construction applied to the morphism $f \circ g$ of $\mathrm{Fet}(X,x)$ will give the same element in $\pi_2(\pi_0(H_*),\pi_0(h))$ as f, where $g:(V,v) \to (Y,y)$ is a morphism of $\mathrm{Fet}(X,x)$.

Let (C_*,c) be the same Cech hypercovering as above, and let C'_* denote the étale Čech hypercovering generated by the cover $V \to Z$, and equip it with a point c' with respect to the same pointed site. Note that g furnishes a map $\delta:(C'_*,c') \to (C_*,c)$ of pointed hypercoverings. Let (C_*,c), (I_*,i) and (J_*,j) be as above, let (I'_*, i') be the fibre product of (H_*, h) and (C'_*, c'), and let (J'_*, j') be the pullback $(f \circ g)^*(I'_*, i')$. Let $f_* : (\pi_0(J_*), \pi_0(j)) \to (\pi_0(I_*), \pi_0(i))$ and $(f \circ g)_* : (\pi_0(J'_*), \pi_0(j')) \to (\pi_0(I'_*), \pi_0(i'))$ be the maps induced by f and by $f \circ g$, respectively. Note that g induces a map $g_* : g^*(\pi_0(J_*), \pi_0(j)) \to (\pi_0(J_*), \pi_0(j))$. Then there is a commutative diagram

$$\begin{tikzcd}
\pi_2(\pi_0(J'_*), \pi_0(j')) \arrow[r, "\pi_2((f\circ g)_*)"] \arrow[d, "\pi_2(g_* \circ (f \circ g)^*(\delta_H))"] & \pi_2(\pi_0(I'_*), \pi_0(i')) \arrow[r] \arrow[d, "\pi_2(\delta_H)"] & \pi_2(\pi_0(H_*), \pi_0(h)) \arrow[d, equal] \\
\pi_2(\pi_0(J_*), \pi_0(j)) \arrow[r, "\pi_2(f_*)"] & \pi_2(\pi_0(I_*), \pi_0(i)) \arrow[r] & \pi_2(\pi_0(H_*), \pi_0(h)),
\end{tikzcd}$$

where $\delta_H : (I'_*, i') \to (I_*, i)$ is the fibre product of δ with (H, h), the morphism $(f \circ g)^*(\delta_H) : (\pi_0(J'_*), \pi_0(j')) \to g^*(\pi_0(J_*), \pi_0(j)) = (f \circ g)^*(\pi_0(I_*), \pi_0(i))$ is the base change of δ_H with respect to $f \circ g$; while the middle and left vertical maps are induced by δ_H and $g_* \circ (f \circ g)^*(\delta_H)$, respectively. Since the image of $\gamma_{f\circ g}$ under the left vertical map is γ_f, the claim above follows.

In order to conclude the proof of the theorem itself, we only need to show that, for every morphism $\delta : (H_*, h) \to (H'_*, h')$ of pointed hypercoverings of Z with respect to the étale site of Z pointed with z, the induced map $\pi_2(\delta) : \pi_2(\pi_0(H_*), \pi_0(h)) \to \pi_2(\pi_0(H'_*), \pi_0(h'))$ takes $\gamma_{(H_*,h)}$ to $\gamma_{(H'_*,h')}$. Indeed, in this case these $\gamma_{(H_*,h)}$ glue together to an element of $\pi_2(Z, z)$ whose image is γ under the Hurewicz map, by construction. Now let $f : (Y, y) \to (Z, z)$ be a morphism of $\mathrm{Fet}(X, x)$ such that the image of $\pi_1(f) : \pi_1(Y, y) \to \pi_1(Z, z)$ lies in the kernel of the natural map $\pi_1(Z, z) \to \pi_1(\pi_0(H_*), \pi_0(h))$. Let (C_*, c), (I_*, i) and (J_*, j) be as above. Note that the image of $\pi_1(f)$ lies in the kernel of the natural map $\pi_1(Z, z) \to \pi_1(\pi_0(H'_*), \pi_0(h'))$, too. Let (I'_*, i') be the fibre product of (H'_*, h') and (C_*, c) and let (J'_*, j') be the pullback $f^*(I'_*, i')$ onto the étale site of Y pointed with y with respect to f. By slight abuse of notation let f_* denote both maps $(\pi_0(J_*), \pi_0(j)) \to (\pi_0(I_*), \pi_0(i))$ and $(\pi_0(J'_*), \pi_0(j')) \to (\pi_0(I'_*), \pi_0(i'))$ induced by f. Then there is a commutative diagram

$$\begin{tikzcd}
\pi_2(\pi_0(J_*), \pi_0(j)) \arrow[r, "\pi_2(f_*)"] \arrow[d, "\pi_2(f^*(\delta_C))"'] & \pi_2(\pi_0(I_*), \pi_0(i)) \arrow[r] \arrow[d, "\pi_2(\delta_C)"'] & \pi_2(\pi_0(H_*), \pi_0(h)) \arrow[d, "\pi_2(\delta)"'] \\
\pi_2(\pi_0(J'_*), \pi_0(j')) \arrow[r, "\pi_2(f_*)"] & \pi_2(\pi_0(I'_*), \pi_0(i')) \arrow[r] & \pi_2(\pi_0(H'_*), \pi_0(h')),
\end{tikzcd}$$

where the middle and left vertical maps are induced by the fibre product δ_C of δ with (C, c) and the pullback $f^*(\delta_C)$ of δ_C with respect to f, respectively. By construction, the image of $\gamma_f \in \pi_2(\pi_0(J_*), \pi_0(j))$ constructed for (H, h) under $\pi_2(f^*(\delta_C))$ is the element denoted by the same symbol and constructed for (H', h'). The claim is now clear. $\square$

5. The homotopy type of curves and abelian varieties

Definition 5.1. Following [Serre 1997, p. 16], we will say that a group Γ is good if the homomorphism of cohomology groups $H^n(\widehat{\Gamma}, M) \to H^n(\Gamma, M)$ induced by the natural homomorphism $\Gamma \to \widehat{\Gamma}$ is an isomorphism for every finite Γ-module M. For every smooth, connected, quasiprojective variety X over any field K and every $n \geq 1$, let $\pi_n(X)$ denote the isomorphism class of the n-th homotopy group of $\pi_n(X, x)$ for some geometric point x. As the notation indicates, these isomorphism classes do not depend on the choice of the base point.

Proposition 5.2. *Let X be a smooth variety over $\mathbb{C}$ such that $X(\mathbb{C})$ has the homotopy type of the Eilenberg–MacLane space $B\pi_1(X(\mathbb{C}))$ and the group $\pi_1(X(\mathbb{C}))$ is good. Then* $\mathrm{Et}(X)$ *is weakly homotopy equivalent to $B\pi_1(X)$.*

Proof. By [Artin and Mazur 1969, Corollary 6.6, p. 72] the profinite completion $(B\Gamma)^\wedge$ of $B\Gamma$ is weakly homotopy equivalent to $B\widehat{\Gamma}$ if and only if Γ is a good group. Because we assumed that X is smooth, $\mathrm{Et}(X)$ is weakly homotopy equivalent to $B\widehat{\pi_1(X(\mathbb{C}))}$ by Theorem 2.9. On the other hand, the profinite completion of $\pi_1(X(\mathbb{C}))$ is isomorphic to $\pi_1(X)$ by the Grauert–Remmert theorem. The claim is now clear. □

Remark 5.3. It is important to note that the condition requiring the fundamental group to be good is not only sufficient, but also necessary. In particular, there are algebraic varieties X over $\mathbb{C}$ such that $X(\mathbb{C})$ has the homotopy type of the Eilenberg–MacLane space $B\pi_1(X(\mathbb{C}))$, but the group $\pi_1(X(\mathbb{C}))$ is not good, therefore $\mathrm{Et}(X)$ is not an Eilenberg–MacLane space. For an important class of examples see [Mochizuki 2003, Lemma 3.16, p. 146].

Proposition 5.4. *Let X be a smooth geometrically irreducible quasiprojective variety over an algebraically closed field K of characteristic zero, and let F be another algebraically closed field containing K. Then* $\mathrm{Et}(X)$ *is weakly homotopy equivalent to* $\mathrm{Et}(X_F)$.

Proof. This claim is a special case of [Artin and Mazur 1969, Corollary 12.12, p. 144] when X is proper, using also [Artin and Mazur 1969, Theorem 11.1, p. 124]. We only need to add a little bit more when X is not proper. By Hironaka's resolution of singularities, there is a projective variety Y over K which contains X as an open subvariety such that the complement $C \subset Y$ is a normal crossings divisor. By the tame invariance theorem, the tame fundamental groups $\pi_1^C(Y)$ and $\pi_1^{C_F}(Y_F)$ are isomorphic. But since the base fields have characteristic zero we have

$$\pi_1^C(Y) \cong \pi_1(X) \quad \text{and} \quad \pi_1^{C_F}(Y_F) \cong \pi_1(X_F).$$

Therefore $\pi_1(X) \cong \pi_1(X_F)$, so the argument presented in [Artin and Mazur 1969] can be applied in this case, too. □

Corollary 5.5. (a) *Let X be a smooth, geometrically connected curve over an algebraically closed field K of characteristic zero which is not a projective curve of genus zero. Then* $\mathrm{Et}(X)$ *is weakly homotopy equivalent to* $B\pi_1(X)$.

(b) *Let X be an abelian variety over an algebraically closed field K of characteristic zero. Then* $\mathrm{Et}(X)$ *is weakly homotopy equivalent to* $B\pi_1(X)$.

Proof. Recall that a smooth, geometrically connected curve Y defined over a field has type (g, d) if g is the genus of the smooth projective completion Y^c of Y and d is the number of geometric points in the complement of Y in Y^c. Let X be a smooth, geometrically connected curve of type (g, d) such that $(g, d) \neq (0, 0)$ over an algebraically closed field K of characteristic zero. There is a subfield $F \subset K$ which is finitely generated over $\mathbb{Q}$ with X already defined over F, that is, there is a smooth, geometrically connected curve Y of type (g, d) over F whose base change to K is X.

Choose an embedding $i : \overline{F} \to \mathbb{C}$ of fields. Then the base change $Y_{\mathbb{C}}$ of the curve $Y_{\overline{F}}$ to $\mathbb{C}$ with respect to this embedding is also a smooth, geometrically connected curve of type (g, d). The topological space $Y_{\mathbb{C}}(\mathbb{C})$ has the homotopy type of the Eilenberg–MacLane space $B\pi_1(Y_{\mathbb{C}}(\mathbb{C}))$. The topological fundamental group of a smooth, connected complex curve is good (this fact follows at once from [Serre 1997, Problem 1(a), p. 15]) so we get from Proposition 5.2 that $\mathrm{Et}(Y_{\mathbb{C}})$ is weakly homotopy equivalent to $B\pi_1(Y_{\mathbb{C}})$. By a repeated application of Proposition 5.4 we get that $\mathrm{Et}(X) = \mathrm{Et}(Y_K)$ is weakly homotopy equivalent to $\mathrm{Et}(Y_{\mathbb{C}})$, and hence $\pi_1(X) \cong \pi_1(Y_{\mathbb{C}})$, so $\mathrm{Et}(X)$ is weakly homotopy equivalent to $B\pi_1(X)$.

The proof of claim (b) is essentially the same as the proof of claim (a); we only need to add that finitely generated free abelian groups are good (see [Serre 1997, Problem 2(d), p. 16]), so $\mathrm{Et}(A)$ is weakly homotopy equivalent to $B\pi_1(A)$ for every abelian variety A defined over $\mathbb{C}$ by Propositions 5.2 and 5.4. □

6. Grothendieck's short exact sequence

Notation 6.1. Let X be a geometrically connected variety defined over K. Let η be a $\overline{K}$-valued point of X. Then Grothendieck's short exact sequence of étale fundamental groups for X is

$$1 \longrightarrow \pi_1(\overline{X}, \eta) \longrightarrow \pi_1(X, \eta) \longrightarrow \Gamma_K \longrightarrow 1, \tag{6.1.1}$$

which is an exact sequence of profinite groups in the category of topological groups. Every K-rational point $x \in X(K)$ induces a section $\Gamma_K \to \pi_1(X, \eta)$ of the sequence (6.1.1), well-defined up to conjugation. Let $\mathrm{Sec}(X/K)$ denote the set of conjugacy classes of sections of (6.1.1) (in the category of profinite groups where morphisms are continuous homomorphisms). Then we have a map

$$s_{X/K} : X(K) \to \mathrm{Sec}(X/K)$$

which sends every point $x \in X(K)$ to the corresponding conjugacy class of sections.

Definition 6.2. For every characteristic open subgroup N of $\pi_1(\overline{X}, \eta)$, consider the short exact sequence

$$1 \longrightarrow \pi_1(\overline{X}, \eta)/N \longrightarrow \pi_1(X, \eta)/N \longrightarrow \Gamma_K \longrightarrow 1, \tag{6.2.1}$$

obtained by dividing out (6.1.1) by N. Let $\mathrm{Sec}(X/K, N)$ denote the set of conjugacy classes of sections of (6.2.1). Let

$$s_{X/K,N} : X(K) \to \mathrm{Sec}(X/K, N)$$

denote the composition of $s_{X/K,N}$ and the natural forgetful map

$$\phi_{X/K,N} : \mathrm{Sec}(X/K) \to \mathrm{Sec}(X/K, N).$$

Note that for every pair of characteristic open subgroups $N' \subseteq N$ of $\pi_1(\overline{X}, \eta)$ the composition of $\phi_{X/K,N'}$ and the forgetful map $\mathrm{Sec}(X/K, N') \to \mathrm{Sec}(X/K, N)$ is $\phi_{X/K,N}$. Therefore we may take the projective limit of the maps $\phi_{X/K,N}$ to get a map

$$\phi_{X/K} = \lim_N \phi_{X/K,N} : \mathrm{Sec}(X/K) \to \lim_N \mathrm{Sec}(X/K, N).$$

where the limit is over the set of characteristic open subgroups of $\pi_1(\overline{X}, \eta)$ directed with respect to reverse inclusion.

Proposition 6.3. *The map $\phi_{X/K}$ is a bijection.*

Proof. Let r, s be two sections $\Gamma_K \to \pi_1(X, \eta)$ such that for every characteristic open subgroup N of $\pi_1(\overline{X}, \eta)$ the compositions of r and s with the quotient map $\pi_1(X, \eta) \to \pi_1(X, \eta)/N$ are conjugates. Then for every such N the set

$$C_N = \{g \in \pi_1(\overline{X}, \eta) \mid g^{-1} r(h) g s(h)^{-1} \in N \text{ for all } h \in \Gamma_K\}$$

is nonempty. Since the sets C_N are closed in the compact topological space $\pi_1(\overline{X}, \eta)$, their intersection

$$\bigcap_N C_N = \left\{ g \in \pi_1(\overline{X}, \eta) \,\middle|\, g^{-1} r(h) g s(h)^{-1} \in \bigcap_N N \text{ for all } h \in \Gamma_K \right\}$$

is also nonempty. Because in a topologically finitely generated profinite group, such as $\pi_1(\overline{X}, \eta)$, the intersection of all characteristic open subgroups is the identity element, we get that r and s are conjugates. Therefore $\phi_{X/K}$ is injective.

Now let r be an element of $\lim_N \mathrm{Sec}(X/K, N)$, and for every N as above let $r_N \in \mathrm{Sec}(X/K, N)$ be the image of r under the projection $\lim_M \mathrm{Sec}(X/K, M) \to \mathrm{Sec}(X/K, N)$. For every positive integer m let $N(m)$ be the intersection of all open subgroups of $\pi_1(\overline{X}, \eta)$ of index at most m. We are going to construct a section $s_m : \Gamma_K \to \pi_1(X, \eta)/N(m)$ whose conjugacy class is $r_{N(m)}$ for every m by induction, as follows. When $m = 1$ this section is just the identity. Assume now that s_{m-1} is already constructed. Let s'_m be a section $\Gamma_K \to \pi_1(X, \eta)/N(m)$ whose conjugacy class is $r_{N(m)}$. Because $r_{N(m)}$ maps to $r_{N(m-1)}$ under the forgetful map $\mathrm{Sec}(X/K, N(m)) \to \mathrm{Sec}(X/K, N(m-1))$, we get that there is a $g \in \pi_1(\overline{X}, \eta)/N(m)$ such that the composition of $g^{-1}s'_m g$ and the quotient map $\pi_1(X, \eta)/N(m) \to \pi_1(X, \eta)/N(m-1)$ is s_{m-1}. Let s_m be $g^{-1}s'_m g$. These sections are compatible and their limit is a section

$$s : \Gamma_K \to \lim_{m=1}^{\infty} \pi_1(X, \eta)/N(m) = \pi_1(X, \eta)$$

whose image is r under $\phi_{X/K}$. So the latter is surjective, too. □

Definition 6.4. We say that X is well-equipped with K-rational points if the map $s_{X/K,N}$ is surjective for every characteristic open subgroup N of $\pi_1(\overline{X}, \eta)$. Note that for a different choice of a base point η' there is an isomorphism between $\pi_1(X, \eta)$ and $\pi_1(X, \eta')$ which maps $\pi_1(\overline{X}, \eta)$ onto $\pi_1(\overline{X}, \eta')$, canonical up to conjugacy. Therefore the sets $\mathrm{Sec}(X/K)$ and $\mathrm{Sec}(X/K, N)$ are independent of the choice of the base point η, as the notation indicates.

Proposition 6.5. *The algebraic groups* GL_n *and* PGL_n *are well-equipped with K-rational points over any characteristic-zero field K and positive integer n.*

Proof. Let 1 denote the unit of $\mathrm{GL}_n(\overline{K})$, and $\mathrm{PGL}_n(\overline{K})$ as well. The quotient map $p : \mathrm{GL}_n \to \mathrm{PGL}_n$ by the centre of GL_n induces a surjection

$$\pi_1(p) : \pi_1(\overline{\mathrm{GL}}_n, 1) \to \pi_1(\overline{\mathrm{PGL}}_n, 1),$$

and hence it will be enough to prove the claim for GL_n only. Let $i : \mathrm{GL}_1 \to \mathrm{GL}_n$ be the map which embeds GL_1 into GL_n as diagonal matrices with 1 on the diagonal except at the upper-left corner. This map induces an isomorphism

$$\pi_1(i) : \pi_1(\overline{\mathrm{GL}}_1, 1) \to \pi_1(\overline{\mathrm{GL}}_n, 1).$$

Therefore it will be enough to prove the claim for GL_1 only. There is a natural isomorphism

$$\pi_1(\overline{\mathrm{GL}}_1, 1) \cong \widehat{\mathbb{Z}},$$

and for every $k \in \mathbb{N}$ there is a natural bijection

$$\mathrm{Sec}(\mathrm{GL}_1 / K, k\widehat{\mathbb{Z}}) \cong H^1(K, \mu_k),$$

where $\mu_k \subseteq \overline{K}^*$ is the module of k-th roots of unity. Moreover, under this identification $s_{\mathrm{GL}_1/K,k\widehat{\mathbb{Z}}}$ corresponds to the coboundary map furnished by Kummer theory. Since, by Hilbert's theorem 90, $H^1(K, \overline{K}^*)$ is zero, the claim now follows. □

Proposition 6.6. *Assume that K is an algebraically closed field of characteristic zero. Let G be a geometrically connected algebraic group over K, and let $f : X \to Y$ be a principal G-bundle over a geometrically connected smooth variety Y over K. Let $x \in X$ and set $y = f(x)$. Then the sequence*

$$\pi_1(f^{-1}(y), x) \to \pi_1(X, x) \xrightarrow{\pi_1(f)} \pi_1(Y, y) \to 1 \tag{6.6.1}$$

is exact, where the first map is induced by the inclusion $f^{-1}(y) \subseteq X$.

Proof. There is a subfield $F \subset K$ which is finitely generated over $\mathbb{Q}$ and G, X, Y and f are already defined over F. Therefore, by the invariance theorem for the (tame) étale fundamental group, it will be sufficient to prove the claim for $\overline{F}$. By the axiom of choice there is an embedding $\overline{F} \to \mathbb{C}$ of fields, and hence we may assume that K is $\mathbb{C}$ without loss of generality, again by the invariance theorem. Because the map $X(\mathbb{C}) \to Y(\mathbb{C})$ induced by f is a Serre fibration, there is a short exact sequence

$$\pi_1(f^{-1}(y)(\mathbb{C}), x) \longrightarrow \pi_1(X(\mathbb{C}), x) \longrightarrow \pi_1(Y(\mathbb{C}), y) \longrightarrow 1 \tag{6.6.2}$$

of topological fundamental groups of complex analytic spaces. The profinite completion functor is right-exact, so the completion of (6.6.2) is also exact. By the Grauert–Remmert theorem the latter is the sequence (6.6.1). □

Proposition 6.7. *Let G be a geometrically connected algebraic group well-equipped with K-rational points over a field K of characteristic zero. Let $f : X \to Y$ be a principal G-bundle over a smooth variety Y over K and let $x, y \in Y(K)$ be such that:*

(i) $s_{Y/K}(x) = s_{Y/K}(y)$.

(ii) *The sets $f^{-1}(x)(K)$ and $f^{-1}(y)(K)$ are nonempty.*

Then, for every characteristic open subgroup N of $\pi_1(\overline{X}, \eta)$ (where $\eta \in X(\overline{K})$ is arbitrary), there are two points $x_N \in f^{-1}(x)(K)$ and $y_N \in f^{-1}(y)(K)$ such that $s_{X/K,N}(x_N) = s_{X/K,N}(y_N)$.

Proof. Pick two points $x' \in f^{-1}(x)(K)$, $y' \in f^{-1}(y)(K)$ and let $\eta \in f^{-1}(x)(\overline{K})$. Let $r \in s_{X/K}(x')$, $s \in s_{X/K}(y')$ be two sections of the short exact sequence (6.1.1). Because both x' and η lie in $f^{-1}(x)(K)$, we may assume that the image of r is in $\pi_1(\overline{f^{-1}(x)}, \eta)$ without loss of generality. Now let $\theta = f(\eta)$, and let r_0, s_0 be the composition of r, s and $\pi_1(f) : \pi_1(X, \eta) \to \pi_1(Y, \theta)$, respectively. As $r_0 \in s_{Y/K}(x)$ and $s_0 \in s_{Y/K}(y)$, we get that these sections are conjugate. By Proposition 6.6 the map $\pi_1(f)$ is surjective, and therefore we may assume that

r_0 and s_0 are the same, by conjugating s if necessary. This implies that s lies in $\pi_1(f^{-1}(x), \eta)$, just as r does, by Proposition 6.6. Let N be now a characteristic open subgroup of $\pi_1(\overline{X}, \eta)$, and let M' be the preimage of N with respect to the map $\pi_1(\overline{f^{-1}(x)}, \eta) \to \pi_1(\overline{X}, \eta)$. As M' is open and $\pi_1(\overline{f^{-1}(x)}, \eta)$ is topologically finitely generated, there is a characteristic open subgroup M of $\pi_1(\overline{f^{-1}(x)}, \eta)$ lying in M'. Because $f^{-1}(x)$ has a K-rational point, it is isomorphic to G. Therefore it is well-equipped with K-rational points, so there is an $x_N \in f^{-1}(x)(K)$ such that the composition of s and the quotient map $\pi_1(f^{-1}(x), \eta) \to \pi_1(f^{-1}(x), \eta)/M$ lies in $s_{f^{-1}(x)/K,M}(x_N)$. If we set $y_N = y'$ then it is clear that the pair x_N, y_N satisfies the required properties. $\square$

Proposition 6.8. *Let G be a geometrically connected algebraic group well-equipped with K-rational points over a field K of characteristic zero such that $\pi_1(\overline{G})$ is finite. Let $f : X \to Y$ be a principal G-bundle over a smooth variety Y over K and let $x, y \in Y(K)$ be such that*:

(i) $s_{Y/K}(x) = s_{Y/K}(y)$.

(ii) *The sets $f^{-1}(x)(K)$ and $f^{-1}(y)(K)$ are nonempty.*

Then there are two points $x' \in f^{-1}(x)(K)$ and $y' \in f^{-1}(y)(K)$ such that $s_{X/K}(x') = s_{X/K}(y')$.

Proof. The proof is the same as above, except that we look at sections of the full Grothendieck short exact sequence (6.1.1) for X. We leave the details to the reader. $\square$

Remark 6.9. By Proposition 6.5 we may apply Proposition 6.7 to principal GL_n-bundles. Note that $\mathrm{SL}_n \to \mathrm{PSL}_n$ is a finite étale cover and SL_n is simply connected. As PSL_n and PGL_n are isomorphic, we get that $\pi_1(\overline{\mathrm{PGL}_n})$ is finite. Therefore by Proposition 6.5 we may apply Proposition 6.8 to principal PGL_n-bundles.

7. Basic consequences of obstruction theory

Definition 7.1. For every $n \in \mathbb{N}$, by functoriality we get a natural map

$$\iota^n_{X/K} : X(K) \to [\mathrm{Et}_{/K}(\mathrm{Spec}(K)), \mathrm{Et}_{/K}(X)] \to [\mathrm{Et}_{/K}(\mathrm{Spec}(K))^\natural, \mathrm{Et}^n_{/K}(X)],$$

where the second map is furnished by applying the Postnikov tower functor and composing with the n-th truncation map $\mathrm{Et}_{/K}(X)^\natural \to \mathrm{Et}^n_{/K}(X)$. We will denote $[\mathrm{Et}_{/K}(\mathrm{Spec}(K))^\natural, \mathrm{Et}^n_{/K}(X)]$ by the symbol $X^n(hK)$. The n-th truncation map $S^\natural \to P_n(S)$ is natural, and hence induces a natural map

$$h^n_{X/K} : X(hK) \to X^n(hK)$$

such that $\iota^n_{X/K} = h^n_{X/K} \circ \iota_{X/K}$. For every positive integer n, let $\sim_n$ denote the following equivalence relation on $X(hK)$: for every pair $x, y \in X(hK)$ we have $x \sim_n y$ if and only if $h^n_{X/K}(x) = h^n_{X/K}(y)$. It is clear that the equivalence relation $\sim_{n+1}$ is finer than the equivalence relation $\sim_n$.

Definition 7.2. For every pointed K-scheme (X, x), let

$$(X, x)(hK) = \mathrm{Et}_{/K}(X, x)(E\Gamma_K).$$

Let $\chi(X, x) : (X, x)(hK) \to X(hK)$ denote the bijection induced by the map $i(X, x)$ (see Proposition 3.7). Then we have a unique natural map $\iota_{(X,x)/K} : X(K) \to (X, x)(hK)$ such that the diagram

$$\begin{tikzcd}
 & (X,x)(hK) \arrow[d, "\chi(X,x)"] \\
X(K) \arrow[ur, "\iota_{(X,x)/K}"] \arrow[r, "\iota_{X/K}"'] & X(hK)
\end{tikzcd}$$

is commutative.

Definition 7.3. We are going to need a variant of the equivalence relations $\sim_n$ for pointed relative étale homotopy types, too. Let (X, x) be a pointed K-scheme as above. We will denote $\mathrm{Et}^n_{/K}(X, x)(E\Gamma_K)$ by the symbol $(X, x)^n(hK)$. The n-th truncation map furnishes a natural map

$$h^n_{(X,x)/K} : (X, x)(hK) \to (X, x)^n(hK).$$

By slight abuse of notation, for every positive integer n let $\sim_n$ denote the following equivalence relation on $(X, x)(hK)$: for every pair $x, y \in (X, x)(hK)$ we have $x \sim_n y$ if and only if $h^n_{(X,x)/K}(x) = h^n_{(X,x)/K}(y)$.

This notation is justified because of the following:

Lemma 7.4. *Under the map $\chi(X, x)$ the equivalence relation $\sim_n$ on $(X, x)(hK)$ corresponds to the equivalence relation $\sim_n$ on $X(hK)$.*

Proof. For every étale hypercovering H_* of X there is a natural commutative diagram

$$\begin{tikzcd}
\pi_{0/K}(H_*) \arrow[r, "i(H_*,x)"] \arrow[d] & \pi_{0/K}(H_*, x) \arrow[d] \\
P_n(\pi_{0/K}(H_*)) \arrow[r, "P_n(i(H_*,x))"] & P_n(\pi_{0/K}(H_*, x))
\end{tikzcd}$$

where we use Notation 3.6, and where the vertical maps are induced by truncation. Since the upper horizontal map is a *weak* equivalence, so is the lower horizontal map, so it induces a bijection on homotopy fixed points. The claim now follows by taking the limit, similarly to the proof of Proposition 3.7. □

Notation 7.5. For every profinite group Γ and every pro-discrete Γ-module M, let $H^k(\Gamma, M)$ denote the projective limit of the (continuous) cohomology groups $H^k(\Gamma, N)$, where N runs through the directed system of discrete quotients of M. For the sake of simple notation, for the rest of the paper for every field K and every pro-discrete Γ_K-module M let $H^k(K, M)$ denote the group $H^k(\Gamma_K, M)$ introduced above. Note that these groups commute with projective limits.

Theorem 7.6. *Assume that X is a smooth geometrically connected variety over K, and assume that $X(K)$ is nonempty. For every $x, y \in X(hK)$:*

(i) *$x = y$ if and only if $x \sim_n y$ for every $n \in \mathbb{N}$.*

(ii) *There is a natural bijection*

$$j_{X/K} : X^1(hK) \to \mathrm{Sec}(X/K)$$

such that for every $p \in X(K)$ we have $j_{X/K}(\iota^1_{X/K}(p)) = s_{X/K}(p)$.

(iii) *For every positive integer n, if $x \sim_n y$ then there exists a natural obstruction class $\delta^X_n(x, y) \in H^{n+1}(K, \pi_{n+1}(\overline{X}))$ such that $x \sim_{n+1} y$ if and only if $\delta^X_n(x, y) = 0$.*

Proof. The first claim is an immediate consequence of the definition of $\mathrm{Et}_{/K}(X)^\natural$. Next we are going to prove (ii). Let $f : Y \to X$ be a torsor under a finite étale group $\mathcal{G}$ over K. Then the $\overline{K}$-valued points $G = \mathcal{G}(\overline{K})$ of $\mathcal{G}$ form a finite group equipped with an action of Γ_K. Let C_* denote the étale Čech hypercovering

$$Y \mathrel{\substack{\longrightarrow\\ \longleftarrow\\ \longleftarrow}} Y \times_X Y \mathrel{\substack{\longleftarrow\\ \longrightarrow\\ \longleftarrow\\ \longrightarrow\\ \longleftarrow}} Y \times_X Y \times_X Y \mathrel{\substack{\longleftarrow\\ \longrightarrow\\ \longleftarrow\\ \longrightarrow\\ \longleftarrow\\ \longrightarrow\\ \longleftarrow}} \cdots$$

generated by the cover $Y \to X$. It is explained at the beginning of Section 9 of [Harpaz and Schlank 2013] that there is a natural map

$$c_Y : \pi_{0/K}(C_*)(E\Gamma_K) \to H^1(\Gamma_K, G).$$

Moreover, by [Harpaz and Schlank 2013, Lemma 9.1], for every $p \in X(K)$ the image of the corresponding homotopy fixed point in $\pi_{0/K}(C_*)(E\Gamma_K)$ with respect to c_Y is the element which classifies the $\mathcal{G}$-torsor $Y_p = f^{-1}(p)$. When Y is geometrically connected there is a *weak* equivalence $\pi_\Gamma(C_*) \to BG$, where we equip the latter with the tautological Γ_K-action, and hence the map c_Y is a bijection (see the discussion after Lemma 9.7 of [Harpaz and Schlank 2013]). Moreover, the composition of the natural map $X(hK) \to \pi_{0/K}(C_*)(E\Gamma_K)$ and c_Y factors through $h^1_{X/K}$.

Let $\eta \in X(K)$ be a K-rational point and let $\overline{\eta}$ denote the $\overline{K}$-valued point associated to η (see Notation 3.9). Fix an element s of $s_{X/K}(\eta)$. For every characteristic open subgroup N of $\pi_1(X, \overline{x})$ let $N' \subseteq \pi_1(X, \overline{x})$ be the subgroup generated by N and the image of s. Since N' is an open subgroup, there is a connected finite étale cover $f_N : Y_N \to X$ such that the image of $\pi_1(Y_N)$ with respect to $\pi_1(f_N)$ is N'. Let $\mathcal{G}_N$ denote the unique finite étale group over K such that $\mathcal{G}_N(\overline{K})$ is

$$\pi_1(\overline{X}, \overline{\eta})/N \cong \pi_1(X, \overline{\eta})/N',$$

equipped with its natural Γ_K-action induced by conjugation. Then $f_N : Y_N \to X$ is a torsor under $\mathcal{G}_N$, and hence by applying the construction above to the étale Čech hypercovering generated by f_N we get a map

$$j_{X/K,N} : X^1(hK) \to \mathrm{Sec}(X/K, N)$$

such that for every $p \in X(K)$ we have $j_{X/K,N}(\iota^1_{X/K}(p)) = s_{X/K,N}(p)$. According to Proposition 6.3, by taking the limit over every characteristic open subgroup N of $\pi_1(X, \overline{\eta})$ we get a continuous map

$$j_{X/K} : X^1(hK) \to \mathrm{Sec}(X/K)$$

between compact Hausdorff topological spaces such that for every $p \in X(K)$ we have $j_{X/K}(\iota^1_{X/K}(p)) = s_{X/K}(p)$. By [Harpaz and Schlank 2013, Lemma 9.11] this map is a bijection.

For every connected pointed Γ_K-space S which has only finitely many nontrivial homotopy groups, there is a natural spectral sequence for homotopy groups of homotopy fixed points

$$E_2^{s,t} = H^s(K, \pi_t(S)) \Rightarrow \pi_{t-s}(S^{h\Gamma_K})$$

constructed by Goerss [1995, Theorem B, p. 189]. Therefore for objects of this category there is a natural obstruction class of the type described in the last claim. So claim (iii) follows from Lemma 7.4 and Proposition 3.10 by applying Goerss' results to $\mathrm{Et}_{/K}(X, \eta)$. ☐

Lemma 7.7. *Assume that X is a smooth geometrically connected variety over K and $\mathrm{Et}(\overline{X})$ is weakly homotopy equivalent to $B\pi_1(\overline{X})$. Also suppose that $X(K) \neq \varnothing$. Then for every $x, y \in X(hK)$ we have $x = y$ if and only if $j_{X/K}(x) = j_{X/K}(y)$.*

Proof. All the higher homotopy groups of $\overline{X}$ vanish, so the claim is immediate from the theorem above. ☐

We say that two points $x, y \in X(K)$ are directly A-equivalent if there is a map $f : \mathbb{A}^1_K \to X$ of K-varieties such that $f(0) = x$ and $f(1) = y$. The A-equivalence on $X(K)$ is the equivalence relation generated by direct A-equivalence.

Proposition 7.8. *Assume that K has characteristic zero. Then for every X over K the map*

$$\iota_{X/K} : X(K) \to X(hK)$$

factors through A-equivalence.

Proof. It will be sufficient to show that for every two points $x, y \in X(K)$ which are directly A-equivalent we have $\iota_{X/K}(x) = \iota_{X/K}(y)$. Let $f : \mathbb{A}^1_K \to X$ be a morphism of K-varieties such that $f(0) = x$ and $f(1) = y$. Both $\iota_{X/K}(x)$ and $\iota_{X/K}(y)$ lie in the image of the map $f_* : \mathbb{A}^1_K(hK) \to X(hK)$. By Corollary 5.5 and Lemma 7.7 the set $\mathbb{A}^1_K(hK)$ consists of one element, since $\mathbb{A}^1_{\overline{K}}$ has trivial étale fundamental group. The claim is now clear. $\square$

Remark 7.9. The validity of a such a claim was already suggested by Toën [2004], but his original claim is not true as stated in positive characteristic. In the special case when K is a finite field it was observed in [Tamagawa 1997, Proposition 2.8, pp. 151-152] that the map $s_{\mathbb{A}^1_K/K}$, and hence the map $\iota_{\mathbb{A}^1_K/K}$, is injective. Thus it is not true in general that for two points $x, y \in X(K)$ which are directly A-equivalent we have $\iota_{X/K}(x) = \iota_{X/K}(y)$ when K has positive characteristic.

8. The Manin pairing

Notation 8.1. For every object $X : I \to \mathrm{Ho}(\Gamma\text{-SSets})$ of $\text{Pro-Ho}(\Gamma\text{-SSets})$ such that I is countable, let $X_{h\Gamma} \in \mathrm{ob}(\text{Pro-Ho}(\mathrm{SSets}))$ denote the pro-homotopy quotient defined at the beginning of Section 6.2 of [Harpaz and Schlank 2013]. Note that this construction can be applied to the objects $\mathrm{Et}_{/K}(X), \mathrm{Et}^n_{/K}(X)$ and $\mathrm{Et}_{/K}(X)^\natural$ when $\Gamma = \Gamma_K$ is the absolute Galois group of the field K.

Proposition 8.2. *Let K be a field and X a variety over K. Then there are natural isomorphisms*

$$\mathrm{Et}_{/K}(X)_{h\Gamma_K} \cong \mathrm{Et}(X), \quad \mathrm{Et}^n_{/K}(X)_{h\Gamma_K} \cong \mathrm{Et}^n(X), \quad \mathrm{Et}_{/K}(X)^\natural_{h\Gamma_K} \cong \mathrm{Et}(X)^\natural$$

in the category Pro-Ho(SSets).

Proof. The second isomorphism is the content of Proposition 6.14 in [Harpaz and Schlank 2013]. The third isomorphism follows from the naturality of this isomorphism, and the first isomorphism is shown in the proof of the proposition mentioned above. $\square$

Definition 8.3. Let X be again a variety over K. Note that by functoriality we get a natural map

$$\lambda_{X/K} : X(K) \to [\mathrm{Et}(\mathrm{Spec}(K)), \mathrm{Et}(X)] \to [\mathrm{Et}(\mathrm{Spec}(K))^\natural, \mathrm{Et}(X)^\natural],$$

where the second map is furnished by applying the Postnikov tower functor. By applying the pro-homotopy quotient functor and the proposition above, we get that there is a natural map

$$\kappa_{X/K} : X(hK) \to [\mathrm{Et}(\mathrm{Spec}(K))^\natural, \mathrm{Et}(X)^\natural]$$

such that $\lambda_{X/K} = \kappa_{X/K} \circ \iota_{X/K}$.

Definition 8.4. By [Artin and Mazur 1969, Corollary 10.8, pp. 122-123] there is a natural equivalence between the category of locally constant étale sheaves of finite abelian groups on X and local coefficient systems of finite abelian groups on $\mathrm{Et}(X)^\natural$, and under this equivalence the étale cohomology of X with coefficients in a locally constant étale sheaf $\mathcal{F}$ of finite abelian groups is the same as the cohomology of $\mathrm{Et}(X)^\natural$ with coefficients in the local coefficient system corresponding to $\mathcal{F}$. We will not distinguish between these two categories in what follows. In particular for a finite, étale group scheme G over K we will identify $H^*(\mathrm{Et}(X)^\natural, G)$ and $H^*(X, G)$.

A basic, but important, corollary of these observations is the following. Let G and X be as above, and let $c \in H^i(X, G)$ be a cohomology class for some $i \in \mathbb{N}$.

Lemma 8.5. *Assume that $x, y \in X(K)$ are H-equivalent. Then the cohomology classes $x^*(c), y^*(c) \in H^i(K, G)$ are equal.*

Proof. The lemma follows from the commutativity of the diagram

$$\begin{array}{ccc} X(K) \times H^i(X,G) & \longrightarrow & H^i(K,G) \\ \downarrow{\scriptstyle \lambda_{X/K}} \quad \| & & \| \\ [\mathrm{Et}(\mathrm{Spec}(K))^\natural, \mathrm{Et}(X)^\natural] \times H^i(\mathrm{Et}(X)^\natural, G) & \longrightarrow & H^i(\mathrm{Et}(\mathrm{Spec}(K))^\natural, G) \end{array}$$

where the horizontal maps are the pairings furnished by pullback. $\square$

For every $n \in \mathbb{N}$ not divisible by the characteristic of K, let $(\cdot,\cdot)_n$ denote the pairing

$$(\cdot,\cdot)_n : X(hK) \times H^2(X, \mu_n) \to H^2(K, \mu_n)$$

given by the rule $(x, c)_n = \kappa_{X/K}(x)^*(c)$.

Lemma 8.6. *The diagram*

$$\begin{array}{ccc} X(K) \times \mathrm{Br}(X) & \xrightarrow{(\cdot,\cdot)} & \mathrm{Br}(K) \\ {\scriptstyle \iota_{X/K}}\downarrow \qquad \uparrow & & \uparrow{\scriptstyle \alpha_n} \\ X(hK) \times H^2(X, \mu_n) & \xrightarrow{(\cdot,\cdot)_n} & H^2(K, \mu_n) \end{array} \tag{8.6.1}$$

commutes, where the middle and right vertical arrows are induced by the inclusion $\mu_n \subset \mathbb{G}_m$ of sheaves.

Proof. This is immediate from the functoriality of the constructions involved. □

Lemma 8.7. *Let X be a geometrically connected variety over K and $x, y \in X(K)$. Then $\iota_{X/K}(x) \sim_1 \iota_{X/K}(y)$ if and only if for every finite, étale map $f : Y \to X$ of geometrically connected varieties over K such that there is an $\tilde{x} \in Y(K)$ with the property $f(\tilde{x}) = x$ there is a $\tilde{y} \in Y(K)$ such that $f(\tilde{y}) = y$.*

Proof. By part (ii) of Theorem 7.6 we have $\iota_{X/K}(x) \sim_1 \iota_{X/K}(y)$ if and only if $s_{X/K}(x) = s_{X/K}(y)$. It is well-known that the latter condition is equivalent to the second condition of the claim (see for example [Tamagawa 1997]). □

Lemma 8.8. *Let $f : Y \to X$ be a finite, étale map of varieties over K. Assume that $x, y \in X(K)$ are H-equivalent and that there is an $\tilde{x} \in Y(K)$ such that $f(\tilde{x}) = x$. Then there is a $\tilde{y} \in Y(K)$ such that $\tilde{x}, \tilde{y}$ are H-equivalent and $f(\tilde{y}) = y$.*

Proof. The connected component X' of X on which x lies is geometrically connected. By Proposition 2.10 the point y must lie on the same component. The connected component Y' of Y on which $\tilde{x}$ lies is also geometrically connected, and the restriction $f|_{Y'} : Y' \to X'$ is a finite, étale map. Therefore we may assume without loss of generality that X and Y are geometrically connected. Hence by Lemma 8.7 there is a $\tilde{y} \in Y(K)$ such that $f(\tilde{y}) = y$, and we may even assume that $\iota_{X/K}(\tilde{x}) \sim_1 \iota_{X/K}(\tilde{y})$.

It will be enough to show that $\tilde{x}$ and $\tilde{y}$ are H-equivalent. We will prove that $\iota_{X/K}(\tilde{x}) \sim_n \iota_{X/K}(\tilde{y})$ for every $n \geq 2$ by induction. Since the map $\pi_n(f) : \pi_n(\overline{Y}) \to \pi_n(\overline{X})$ is an isomorphism for every $n \geq 2$ by Proposition 4.1, we get that the induced map

$$H^n(\pi_n(f)) : H^n(K, \pi_n(\overline{Y})) \to H^n(K, \pi_n(\overline{X}))$$

is also an isomorphism. By naturality of the obstruction classes we have

$$H^n(\pi_n(f))(\delta_n^Y(\tilde{x}, \tilde{y})) = \delta_n^X(x, y).$$

Since the right-hand side is zero we get that $\delta_n^Y(\tilde{x}, \tilde{y})$ is also zero. □

Proposition 8.9. *Let X be a regular variety over K, and assume that $x, y \in X(K)$ are H-equivalent. Then x and y are étale-Brauer equivalent.*

Proof. Let $Y \to X$ be a finite, étale morphism of varieties over K such that there is an $\tilde{x} \in Y(K)$ mapping to x. By Lemma 8.8 there is a $\tilde{y} \in Y(K)$ such that $\tilde{x}, \tilde{y}$ are H-equivalent and $f(\tilde{y}) = y$. It will be enough to show that $\tilde{x}, \tilde{y}$ are Brauer equivalent. Because Y is the finite étale cover of a regular variety, it is also regular, so it will be enough show that, given a regular variety X over K, every pair of points $x, y \in X(K)$ which are H-equivalent are also Brauer equivalent. Because X is regular, the group $H^2(X, \mathbb{G}_m)$ is torsion (see [Grothendieck 1971, Proposition 1.4, p. 291]). Therefore for every $b \in H^2(X, \mathbb{G}_m)$ there is a natural number $n \in \mathbb{N}$ and a $c \in H^2(X, \mu_n)$ such that b is the image of c under the natural map $H^2(X, \mu_n) \to H^2(X, \mathbb{G}_m)$. The claim now follows from the commutativity of the diagram (8.6.1). $\square$

9. Brauer equivalence versus étale-Brauer equivalence

Definition 9.1. Let E, E' be two elliptic curves defined over a field K whose characteristic is not two and let $t \in E'(K)$ be a point of order two. Let $\rho : E \to E$ be the multiplication by -1 map and let $\sigma : E' \to E'$ be the translation by t. Let X denote the quotient of $E \times E'$ by the fixed-point-free involution (ρ, σ). Then X is a smooth projective geometrically irreducible surface over K. We call such surfaces bielliptic.

Proposition 9.2. *Let K be a finite extension of $\mathbb{Q}_p$, and let X be a bielliptic surface over K. Then the map $s_{X/K}$ is injective.*

Proof. Let $x, y \in X(K)$ be two different points and let η be $\overline{K}$-valued point of X. Let $\mathcal{K} \subseteq \pi_1(\overline{X}, \eta)$ be the characteristic subgroup such that the quotient $\pi_1(\overline{X}, \eta)/\mathcal{K}$ is the maximal 2-torsion abelian quotient of $\pi_1(\overline{X}, \eta)$. Fix an element s of $s_{X/K}(x)$ and let $\mathcal{K}' \subseteq \pi_1(X, \eta)$ be the subgroup generated by $\mathcal{K}$ and the image of s; this is an open subgroup. Let $f : Y \to X$ be the connected finite étale cover such that the image of $\pi_1(Y)$ with respect to $\pi_1(f)$ is $\mathcal{K}'$. Since $\overline{Y}$ is a finite étale cover of the abelian variety $\overline{E \times E'}$, where we use the notation of the definition above, it is also an abelian variety. Therefore Y is a principal homogeneous space over an abelian variety defined over K.

By construction there is an $x' \in Y(K)$ such that $f(x') = x$. So Y has a K-rational point, and hence it is also an abelian variety over K. If y does not have a lift to a K-valued point of Y then $s_{X/K}(x) \neq s_{X/K}(y)$. So we may assume that there is a $y' \in Y(K)$ such that $f(y') = y$. Let $G = \pi_1(X, \eta)/\pi_1(Y, \eta)$. Since we may take a finite extension of K during the proof of injectivity of $s_{X/K}$, we may assume that the action of Γ_K on G is trivial without loss of generality. In this case this finite group is the Galois group of the connected finite étale cover $f : Y \to X$. It also acts on $\mathrm{Sec}(Y/K)$, and two elements of $\mathrm{Sec}(Y/K)$ are in the same G-orbit if and only if they have the same image under the map $\mathrm{Sec}(Y/K) \to \mathrm{Sec}(X/K)$ induced by f. Moreover, the section map $s_{Y/K} : Y(K) \to \mathrm{Sec}(Y/K)$ is G-equivariant. Therefore, if $s_{X/K}(x) = s_{X/K}(y)$ then $s_{Y/K}(x') = s_{Y/K}(g(y'))$ for some $g \in G$. For abelian varieties over K the section map is injective, so $x' = g(y')$ in this case, which implies that $x = y$. This is a contradiction, therefore $s_{X/K}(x)$ is different from $s_{X/K}(y)$. $\square$

We continue to use the notation which we introduced above. Let D denote the quotient of E' by the fixed-point-free involution σ, and let $g : X \to D$ be the quotient map.

Proposition 9.3. *Assume that K is a finite extension of $\mathbb{Q}_p$. Then the map g^* : $H^2(D,\mathbb{G}_m) \to H^2(X,\mathbb{G}_m)$ induced by g has finite cokernel.*

Proof. For every variety Y over K, the Hochschild–Serre spectral sequence

$$E^2_{p,q} = H^p(K, H^q(\overline{Y},\mathbb{G}_m)) \Rightarrow H^{p+q}(Y,\mathbb{G}_m)$$

furnishes on $H^2(Y,\mathbb{G}_m)$ a natural filtration

$$0 = E^2_3 \subseteq E^2_2 \subseteq E^2_1 \subseteq E^2_0 = H^2(Y,\mathbb{G}_m) \tag{9.3.1}$$

such that

$$E^{\infty}_{p,2-p} \cong E^2_p/E^2_{p+1}, \quad p = 0, 1, 2.$$

The members E^2_2 and E^2_1 are usually denoted by $\mathrm{Br}_0(Y)$ and $\mathrm{Br}_1(Y)$, respectively. Because $H^3(K,\mathbb{G}_m) = 0$ (see [Serre 1997, Proposition 15, p. 93]), the coboundary map

$$d^2_{1,1} : E^2_{1,1} = H^1(K, H^1(\overline{Y},\mathbb{G}_m)) \to H^3(K,\mathbb{G}_m)$$

is zero and therefore

$$E^{\infty}_{1,1} = E^3_{1,1} = \mathrm{Ker}(d^2_{1,1}) = H^1(K, H^1(\overline{Y},\mathbb{G}_m)).$$

In short, we have a natural exact sequence

$$0 \longrightarrow \mathrm{Br}_0(Y) \longrightarrow \mathrm{Br}_1(Y) \longrightarrow H^1(K, H^1(\overline{Y},\mathbb{G}_m)) \longrightarrow 0.$$

The group $\mathrm{Br}_0(Y)$ is the image of the natural map $\mathrm{Br}(K) \to \mathrm{Br}(Y)$, therefore the map $g^* : \mathrm{Br}_0(D) \to \mathrm{Br}_0(X)$ is surjective. The Hochschild–Serre spectral sequence furnishes a natural injection

$$\mathrm{Br}(Y)/\mathrm{Br}_1(Y) \to \mathrm{Br}(\overline{Y}).$$

Since $\mathrm{Br}(\overline{X})$ is dual to the torsion subgroup of the Néron–Severi group $\mathrm{NS}(\overline{X})$ in our case (see [Skorobogatov 1999, p. 403]), which is finite, it will be enough to show that the map

$$g_* : H^1(K, H^1(\overline{D},\mathbb{G}_m)) \to H^1(K, H^1(\overline{X},\mathbb{G}_m))$$

induced by g has finite cokernel. Since g is the Albanese map for X (see [loc. cit.]), the map $\mathrm{Pic}^0(\overline{D}) \to \mathrm{Pic}^0(\overline{X})$ induced by g is an isomorphism. Therefore, by looking at the cohomological long exact sequence associated to the short exact sequence of Γ_K-modules

$$0 \longrightarrow \mathrm{Pic}^0(\overline{X})(\overline{K}) \longrightarrow H^1(\overline{X},\mathbb{G}_m) \longrightarrow \mathrm{NS}(\overline{X}) \longrightarrow 0,$$

we are reduced to show that $H^1(K, \mathrm{NS}(\overline{X}))$ is finite. The abelian group $\mathrm{NS}(\overline{X})$ is finitely generated, so there is a finite Galois extension $L|K$ such that the action of $\mathrm{Gal}(\overline{L}|L)$ on $\mathrm{NS}(\overline{X})$ is trivial. The abelianisation of $\mathrm{Gal}(\overline{L}|L)$ is isomorphic to the profinite completion of L^*, so it is topologically finitely generated. Therefore

$$H^1(L, \mathrm{NS}(\overline{X})) \cong \mathrm{Hom}(\mathrm{Gal}(\overline{L}|L), \mathrm{NS}(\overline{X}))$$

is finite. Therefore the inflation map

$$H^1(\mathrm{Gal}(L|K), \mathrm{NS}(\overline{X})) \to H^1(K, \mathrm{NS}(\overline{X}))$$

has finite cokernel. Since $\mathrm{Gal}(L|K)$ is finite we get that $H^1(K, \mathrm{NS}(\overline{X}))$ is finite. □

Let X be a smooth variety over a field K, and let $b \in H^2(X, \mathbb{G}_m)$. We say that $x, y \in X(K)$ are b-equivalent if $x^*(b) = y^*(b)$. This defines an equivalence relation of $X(K)$, which we will call b-equivalence.

Proposition 9.4. *Assume that K is a finite extension of $\mathbb{Q}_p$. Then b-equivalence classes are open in the p-adic topology.*

Proof. Let $x \in X(K)$. It will be enough to show that x has a p-adically open neighbourhood U in $X(K)$ such that x and y is b-equivalent for every $y \in U$. We may assume that X is affine by taking a Zariski-open neighbourhood of x. By a theorem of Gabber (see [de Jong 2013, Theorem 1.1]), there is an Azumaya algebra A of some rank n on X which represents b. Let $\pi : Y \to X$ be the PGL_n-torsor corresponding to A, that is, the torsor whose class in $H^1(X, \mathrm{PGL}_n)$ is the same as the class of A. Let $\sigma \in H^1(K, \mathrm{PGL}_n)$ be the class of the fibre of Y over x and let $\pi^\sigma : Y^\sigma \to X$ be the twist of $\pi : Y \to X$ by σ. Then the fibre of Y^σ over x is a trivial PGL_n-torsor. Therefore there is an $x' \in Y^\sigma(K)$ such that $\pi^\sigma(x') = x$. Because π^σ is a submersion, there is an p-adically open neighbourhood U of x in $X(K)$ and a p-adically analytical section $s : U \to Y^\sigma(K)$ of π^σ with $s(x) = x'$. Therefore for every $y \in U$ the fibre of Y^σ over y has a K-rational point, so it is a trivial PGL_n-torsor. The claim is now clear. □

Theorem 9.5. *Let K be as above, and let X be a bielliptic surface over K. Then étale-Brauer equivalence is strictly finer than Brauer equivalence on $X(K)$.*

Of course our choice of example is motivated by the classical paper [Skorobogatov 1999], and the result above can be considered its natural local counterpart. (Also compare with [Harari 2000], which uses similar ideas.)

Proof. Let X be a bielliptic surface over K, and let $g : X \to D$ be the map introduced above. By Proposition 9.2 the étale-Brauer equivalence-classes of $X(K)$ consists of points, so it will be enough to show that the Brauer equivalence-classes of $X(K)$ are infinite. Let $r \in D(K)$. Note that every $x, y \in g^{-1}(r)(K)$ are b-equivalent for every $b \in H^2(X, \mathbb{G}_m)$ in the image of the map $g^* : H^2(D, \mathbb{G}_m) \to H^2(X, \mathbb{G}_m)$. Therefore finitely many Brauer equivalence classes intersect $g^{-1}(r)(K)$ by Propositions 9.3

and 9.4. By the inverse function theorem both D and $g^{-1}(r)$ have infinitely many K-valued points, so the claim holds. □

In the rest of this section we study the somewhat independent question of the surjectivity of $\iota_{X/K}$ over p-adic fields.

Notation 9.6. Let Groups denote the category of groups. Let $\mathbb{N}_d$ denote the category whose objects are positive integers and whose morphisms are the following: for every pair of objects $m, n \in \mathrm{ob}(\mathbb{N}_d)$ the set of morphisms from m to n consists of the ordered pair $\phi_{m,n} = (m, n)$ if $n \mid m$, and otherwise is empty. For every abelian group A and natural number n let $A[n]$ denote the subgroup of n-torsion elements of A and let A^{tor} denote the pro-group $A^{\mathrm{tor}} : \mathbb{N}_d \to \mathrm{Groups}$ given by the rule

$$A^{\mathrm{tor}}(n) = A[n]$$

such that, for every pair of positive integers m, n such that $n \mid m$, the homomorphism

$$A^{\mathrm{tor}}(\phi_{m,n}) : A[m] \to A[n]$$

is the multiplication by m/n map.

Proposition 9.7. *Let K be a finite extension of $\mathbb{Q}_p$. The following holds*:

(a) *For every smooth, geometrically connected projective curve X of genus at least two over K the map $\iota_{X/K}$ is injective, and it is surjective if the local version of Grothendieck's section conjecture holds for X.*

(b) *For every abelian variety X over K the map $\iota_{X/K}$ is injective, and it is surjective if and only if X is zero-dimensional.*

Proof. First assume that X is a smooth, geometrically connected projective curve of genus at least two over K. Recall that the local version of Grothendieck's section conjecture claims that the map $s_{X/K}$ is a bijection. We also know that in this case $s_{X/K}$ is injective. Therefore claim (a) follows at once from Corollary 5.5 and Lemma 7.7. Assume now that X is an abelian variety over K. By Corollary 5.5 and Lemma 7.7 the map $j_{X/K}$ is a bijection. Moreover, there is a natural bijection

$$\mathrm{Sec}(X/K) \cong H^1\Big(K, \prod_{l \text{ is prime}} T_l(X)\Big), \tag{9.7.1}$$

where $T_l(X)$ denotes the l-th Tate module of X, and under this identification $s_{X/K}$ corresponds to the coboundary map furnished by Kummer theory. In particular $\iota_{X/K}$ is injective, and the cokernel of $s_{X/K}$ is $H^1(K, X)^{\mathrm{tor}}$. By [Milne 1986, Corollary 3.4, p. 53] the groups $H^1(K, X)$ and $X^\vee(K)^\wedge$ are isomorphic, where the latter is the Pontryagin dual of the compact group $X^\vee(K)$ of K-valued points of the dual $X^\vee$ of X. Let $\mathbb{O}_K$ denote the valuation ring of K. The profinite group $X^\vee(K)$ is the direct sum of a finite group and $\dim(X^\vee) = \dim(X)$ copies of $\mathbb{O}_K$

by the inverse function theorem. Therefore the group $H^1(K,X)^{\mathrm{tor}}$ is zero if and only if X is zero-dimensional. Hence claim (b) is true. $\square$

10. Étale-Brauer equivalence versus H-equivalence

Definition 10.1. Assume now that K is a p-adic field, and for every $n \in \mathbb{N}$ let $c_{K,n} : H^2(K,\mu_n) \to \mathbb{Z}/n\mathbb{Z}$ be the isomorphism furnished by local Tate duality. For every geometrically irreducible variety X defined over K let

$$\{\cdot,\cdot\}_n : H^2(K, H_2(\overline{X},\mathbb{Z}/n\mathbb{Z})) \times H^0(K, H^2(\overline{X},\mu_n)) \to \mathbb{Z}/n\mathbb{Z}$$

be the bilinear pairing given by the rule

$$\{x,y\}_n = c_{K,n}(x \cup y)$$

(for every $x \in H^2(K, H_2(\overline{X},\mathbb{Z}/n\mathbb{Z}))$, $y \in H^0(K, H^2(\overline{X},\mu_n))$), where the cup product

$$\cup : H^2(K, H_2(\overline{X},\mathbb{Z}/n\mathbb{Z})) \times H^0(K, H^2(\overline{X},\mu_n)) \to H^2(K,\mu_n)$$

is induced by the evaluation pairing

$$H_2(\overline{X},\mathbb{Z}/n\mathbb{Z}) \times H^2(\overline{X},\mu_n) \to \mu_n.$$

The following lemma is immediate from the functoriality of the constructions involved:

Lemma 10.2. *Let $f : X \to Y$ be a morphism of geometrically irreducible varieties over K. Then the diagram*

$$\begin{array}{ccccc} H^2(K, H_2(\overline{X},\mathbb{Z}/n\mathbb{Z})) & \times & H^0(K, H^2(\overline{X},\mu_n)) & \xrightarrow{\{\cdot,\cdot\}_n} & \mathbb{Z}/n\mathbb{Z} \\ {\scriptstyle H_2(f)_*}\downarrow & & \uparrow{\scriptstyle (f^*)_*} & & \| \\ H^2(K, H_2(\overline{Y},\mathbb{Z}/n\mathbb{Z})) & \times & H^0(K, H^2(\overline{Y},\mu_n)) & \xrightarrow{\{\cdot,\cdot\}_n} & \mathbb{Z}/n\mathbb{Z} \end{array} \qquad (10.2.1)$$

commutes. $\square$

Notation 10.3. Let X be a geometrically irreducible variety over K with $X(K) \neq \varnothing$. For every positive integer n and every $x, y \in X(hK)$ such that $x \sim_1 y$, let $\delta_1^X(x,y)_n$ denote the image of the obstruction class $\delta_1^X(x,y)$ under the composition of the natural map

$$H^2(K,\pi_2(\overline{X})) \to H^2(K,\pi_2(\overline{X})/n\pi_2(\overline{X}))$$

and the homomorphism

$$H_{*,n} : H^2(K,\pi_2(\overline{X})/n\pi_2(\overline{X})) \to H^2(K, H_2(\overline{X},\mathbb{Z}/n\mathbb{Z}))$$

induced by the Hurewicz map

$$H_n : \pi_2(\overline{X})/n\pi_2(\overline{X}) \to H_2(\overline{X}, \mathbb{Z}/n\mathbb{Z}).$$

Let

$$\alpha_n : H^2(X, \mu_n) \to H^2(X, \mathbb{G}_m)$$

denote the map induced by the inclusion $\mu_n \to \mathbb{G}_m$, and finally let

$$\phi_n : H^2(X, \mu_n) \to H^0(K, H^2(\overline{X}, \mu_n))$$

be the map induced by base change.

Lemma 10.4. *Let X be a geometrically irreducible smooth quasiprojective variety over K. For every $n \in \mathbb{N}$, every $c \in H^2(X, \mu_n)$ and every $x, y \in X(K)$ such that $\alpha_n(c) = 0$ and $\iota_{X/K}(x) \sim_1 \iota_{X/K}(y)$, we have*

$$\{\delta_1^X(\iota_{X/K}(x), \iota_{X/K}(y))_n, \phi_n(c)\}_n = 0.$$

Proof. We will need to introduce analogues of the concepts in Definition 10.1 and Notation 10.3 for hypercoverings. For every geometrically irreducible variety Z defined over K and every étale hypercovering H_* of Z let

$$\{\cdot,\cdot\}_n : H^2(K, H_2(\pi_{0/K}(H_*), \mathbb{Z}/n\mathbb{Z})) \times H^0(K, H^2(\pi_{0/K}(H_*), \mu_n)) \to \mathbb{Z}/n\mathbb{Z}$$

also denote the bilinear pairing given by the rule

$$\{a, b\}_n = c_{K,n}(a \cup b)$$

(for every $a \in H^2(K, H_2(\pi_{0/K}(H_*), \mathbb{Z}/n\mathbb{Z})), b \in H^0(K, H^2(\pi_{0/K}(H_*), \mu_n))$), where the cup product

$$\cup : H^2(K, H_2(\pi_{0/K}(H_*), \mathbb{Z}/n\mathbb{Z})) \times H^0(K, H^2(\pi_{0/K}(H_*), \mu_n)) \to H^2(K, \mu_n)$$

is induced by the evaluation pairing

$$H_2(\pi_{0/K}(H_*), \mathbb{Z}/n\mathbb{Z}) \times H^2(\pi_{0/K}(H_*), \mu_n) \to \mu_n.$$

Assume now that $Z(K) \neq \varnothing$ and pick a point $z \in Z(K)$. In Definition 3.5 we introduced a pointed simplicial Γ_K-set $\pi_{0/K}(H_*, z)$ such that there is a natural map $\pi_{0/K}(H_*) \to \pi_{0/K}(H_*, z)$ which is a weak equivalence. Let $a, b \in \pi_0(\pi_{0/K}(H_*)^{\Gamma_K})$ be such that their image is the same under the map

$$\pi_0(\pi_{0/K}(H_*)^{\Gamma_K}) \to \pi_0(P_1(\pi_{0/K}(H_*))^{\Gamma_K})$$

induced by the truncation morphism $\pi_{0/K}(H_*) \to P_1(\pi_{0/K}(H_*))$. By Theorem B on p. 189 of [Goerss 1995] there is an obstruction class $\delta_1^{H_*}(a,b) \in H^2\big(K, \pi_2(\pi_{0/K}(H_*,z))\big)$. For every positive integer n and a,b as above let $\delta_1^{H_*}(a,b)_n$ denote the image of $\delta_1^{H_*}(a,b)$ under the composition of the natural map

$$H^2\big(K, \pi_2(\pi_{0/K}(H_*,z))\big) \to H^2\big(K, \pi_2(\pi_{0/K}(H_*,z))/n\pi_2(\pi_{0/K}(H_*,z))\big),$$

the homomorphism

$$H^2\big(K,\pi_2(\pi_{0/K}(H_*,z))/n\pi_2(\pi_{0/K}(H_*,z))\big)\to H^2(K,H_2(\pi_{0/K}(H_*,z),\mathbb{Z}/n\mathbb{Z}))$$

induced by the Hurewicz map

$$\pi_{0/K}(H_*,z))/n\pi_2(\pi_{0/K}(H_*,z)) \to H_2(\pi_{0/K}(H_*,z),\mathbb{Z}/n\mathbb{Z}),$$

and the inverse of the isomorphism

$$H^2(K, H_2(\pi_{0/K}(H_*),\mathbb{Z}/n\mathbb{Z})) \to H^2(K, H_2(\pi_{0/K}(H_*,z),\mathbb{Z}/n\mathbb{Z}))$$

induced by the weak equivalence $\pi_{0/K}(H_*) \to \pi_{0/K}(H_*,z)$. Finally, for every $a \in X(hK)$ let $a^{H_*} \in \pi_0(\pi_{0/K}(H_*)^{\Gamma_K})$ denote its image under the canonical map $X(hK) \to \pi_0(\pi_{0/K}(H_*)^{\Gamma_K})$.

Now let us start the proof in earnest. Because $\alpha_n(c)=0$ there is a line bundle $\mathscr{L}$ on X such that the image of its isomorphism class under the coboundary map $\mathrm{Pic}(X)=H^1(X,\mathbb{G}_m)\to H^2(X,\mu_n)$ is c. Let $\pi : Y \to X$ denote the total space of $\mathscr{L}$ with the zero section removed; then Y is a $\mathbb{G}_m$-torsor over X whose class in $H^1(X,\mathbb{G}_m)$ is the same as the class of $\mathscr{L}$. It will be enough to show that

$$\{\delta_1^{H_*}\big(\iota_{X/K}(x)^{H_*}, \iota_{X/K}(y)^{H_*}\big)_n, d\}_n$$

is zero for every étale hypercovering H_* of X and every

$$d \in H^0(K, H^2(\pi_{0/K}(H_*),\mu_n))$$

whose image is $\phi_n(c)$ with respect to the homomorphism

$$H^0(K, H^2(\pi_{0/K}(H_*),\mu_n)) \to H^0(K, H^2(\overline{X},\mu_n))$$

induced by the pullback map $H^2(\pi_{0/K}(H_*),\mu_n)\to H^2(\overline{X},\mu_n)$.

Fix such a hypercovering H_*, and let $\pi^*(H_*)$ denote the pullback of H_* onto Y with respect to π. Note that $\phi_n(\pi^*(c))=0$; in fact even $\pi^*(c)=0$. Indeed, the latter follows as the pullback of the torsor $\pi : Y \to X$ onto Y with respect to π is trivial: the diagonal $Y \to Y\times_X Y$ is a section. Let

$$d' \in H^0(K, H^2(\pi_{0/K}(\pi^*(H_*)),\mu_n))$$

be the image of d with respect to the homomorphism

$$H^0(K, H^2(\pi_{0/K}(H_*), \mu_n)) \to H^0(K, H^2(\pi_{0/K}(\pi^*(H_*)), \mu_n))$$

induced by π. As the image of d' is $\phi_n(\pi^*(c))$ under the pullback

$$H^0(K, H^2(\pi_{0/K}(\pi^*(H_*)), \mu_n)) \to H^0(K, H^2(\overline{Y}, \mu_n))$$

by naturality, we get that there is a morphism $f : I_* \to \pi^*(H_*)$ of étale hypercoverings of Y such that $f^*(d') \in H^0(K, H^2(\pi_{0/K}(I_*), \mu_n))$ is zero.

Let $z \in Y(K)$ be arbitrary. (There are such points; for example, the fibre above x contains K-rational points.) Let $\bar{z} \in Y(\overline{K})$ denote the geometric point lying above z, corresponding to the choice of algebraic closure $\overline{K} \supset K$. Note that the image of the canonical map from $\pi_1(\overline{Y}, \bar{z}) \to \pi_{0/K}(I_*, z)$, which is well-defined by Proposition 3.10, is finite. Therefore by Remark 6.9 there are two points x' and y' in $Y(K)$ whose images under π are x and y, respectively, such that $\iota_{X/K}(x')^{I_*}$, $\iota_{X/K}(y')^{I_*} \in \pi_0(\pi_{0/K}(I_*)^{\Gamma_K})$ have the same image under the map

$$\pi_0(\pi_{0/K}(I_*)^{\Gamma_K}) \to \pi_0(P_1(\pi_{0/K}(I_*))^{\Gamma_K})$$

induced by truncation. So by the above the obstruction class

$$\delta_1^{I_*}(\iota_{Y/K}(x')^{I_*}, \iota_{Y/K}(y')^{I_*}) \in H^2\big(K, \pi_2(\pi_{0/K}(I_*, z))\big)$$

is well-defined. By naturality of obstruction classes it will be enough to show that

$$\{\delta_1^{I_*}(\iota_{Y/K}(x')^{I_*}, \iota_{Y/K}(y')^{I_*})_n, f^*(d')\}_n = 0.$$

But this is clear since $f^*(d')$ is zero. □

Lemma 10.5. *Let X be a geometrically irreducible smooth quasiprojective variety over K. For every $n \in \mathbb{N}$, every $c \in H^2(X, \mu_n)$ and every $x, y \in X(K)$ such that $\iota_{X/K}(x) \sim_1 \iota_{X/K}(y)$, we have*

$$(\iota_{X/K}(x), c)_n = (\iota_{X/K}(y), c)_n \implies \{\delta_1^X(\iota_{X/K}(x), \iota_{X/K}(y))_n, \phi_n(c)\}_n = 0.$$

Proof. By a theorem of Gabber (see Theorem 1.1 of [de Jong 2013]) there is an Azumaya algebra A on X which represents $\alpha_n(c) \in H^2(X, \mathbb{G}_m)$. Without loss of generality we may assume that A has rank n by enlarging n if it is necessary, since for every pair of positive integers $n \mid m$ the map $H^2(K, \mu_n) \to H^2(K, \mu_m)$ induced by the inclusion map $\mu_n \subseteq \mu_m$ is injective. Let $\pi : Y \to X$ be the PGL_n-torsor corresponding to A, that is, the torsor whose class in $H^1(X, \mathrm{PGL}_n)$ is the same as the class of A. Let $\sigma \in H^1(K, \mathrm{PGL}_n)$ be the class of the fibre of Y over x and let $\pi^\sigma : Y^\sigma \to X$ be the twist of $\pi : Y \to X$ by σ. Then the fibre of Y^σ over x is a trivial PGL_n-torsor. Because $(x, c)_n = (y, c)_n$ and the natural map $H^1(K, \mathrm{PGL}_n) \to H^2(K, \mu_n)$ is injective, we get that the fibre of Y^σ over y is also a trivial PGL_n-torsor. So there are points x' and y' in $Y^\sigma(K)$ whose images under π^σ are x and y, respectively. By Remark 6.9 we may even assume that $\iota_{Y^\sigma/K}(x') \sim_1 \iota_{Y^\sigma/K}(y')$. So by Lemma 10.2 and the naturality of obstruction classes it will be enough to show that

$$\{\delta_1^{Y^\sigma}(\iota_{Y^\sigma/K}(x'), \iota_{Y^\sigma/K}(y'))_n, \phi_n((\pi^\sigma)^*(c))\}_n = 0.$$

In order to do so, it will be enough to show that

$$\alpha_n((\pi^\sigma)^*(c)) = (\pi^\sigma)^*(\alpha_n(c)) = 0$$

by Lemma 10.4. But this is clear since the pullback of the torsor $\pi : Y^\sigma \to X$ onto Y^σ with respect to π^σ is trivial: the diagonal $Y^\sigma \to Y^\sigma \times_X Y^\sigma$ is a section. □

Notation 10.6. For every quasiprojective variety Y over K consider the Hochschild–Serre spectral sequence

$$E^2_{p,q} = H^p(K, H^q(\overline{Y}, \mu_n)) \Rightarrow H^{p+q}(Y, \mu_n).$$

Because $H^3(K, \mu_n) = 0$ (for example by local Tate duality), the coboundary map

$$d^3_{0,2} : E^3_{0,2} \to E^3_{3,0} \subseteq H^3(K, \mu_n)$$

is zero and therefore

$$E^\infty_{0,2} = E^3_{0,2} = \mathrm{Ker}(d^2_{0,2}).$$

Therefore the spectral sequence furnishes an exact sequence

$$H^2(Y, \mu_n) \xrightarrow{\phi_n} H^0(K, H^2(\overline{Y}, \mu_n)) \xrightarrow{d_n} H^2(K, H^1(\overline{Y}, \mu_n)),$$

which is functorial. Here d_n is the coboundary map $d^2_{0,2} : E^2_{0,2} \to E^2_{2,1}$ and ϕ_n is the quotient map $H^2(Y, \mu_n) \to E^\infty_{0,2} \cong \mathrm{Ker}(d^2_{0,2})$ by the highest step in the filtration on $H^2(Y, \mu_n)$ induced by the spectral sequence.

Lemma 10.7. *Let X be a geometrically irreducible smooth quasiprojective variety over K and let $x, y \in X(K)$ be étale-Brauer equivalent. Then there is a connected finite étale cover $f : Y \to X$ such that*:

(i) *There are $\tilde{x}, \tilde{y} \in Y(K)$ such that $f(\tilde{x}) = x$, $f(\tilde{y}) = y$, and $\tilde{x}, \tilde{y}$ are Brauer-equivalent.*

(ii) *For every $c \in H^0(K, H^2(\overline{X}, \mu_n))$ we have $d_n(f^*(c)) = 0$.*

Proof. Let η be a $\overline{K}$-valued point of X. Let $\mathcal{H}$ be the intersection of the kernels of all continuous homomorphisms $\pi_1(\overline{X}, \eta) \to \mathbb{Z}/n\mathbb{Z}$. Because $\pi_1(\overline{X}, \eta)$ is topologically finitely generated, the subgroup $\mathcal{H}$ is open and characteristic. Let $\mathcal{H}' \subseteq \pi_1(X, \eta)$ be the subgroup generated by $\mathcal{H}$ and by the image of an element of $s_{X/K}(x)$; this is an open subgroup. Let $f : Y \to X$ be the connected finite étale cover such that the image of $\pi_1(Y)$ with respect to $\pi_1(f)$ is $\mathcal{H}'$. By construction there is an $\tilde{x} \in Y(K)$

such that $f(\tilde{x}) = x$. Because x and $y \in X(K)$ are étale-Brauer equivalent, there is a $\tilde{y} \in Y(K)$ such that $f(\tilde{y}) = y$, and $\tilde{x}, \tilde{y}$ are Brauer-equivalent. Recall that by the universal coefficient theorem

$$0 \longrightarrow \mathrm{Ext}^1(H_n(V,\widehat{\mathbb{Z}}),\mathbb{Z}/n\mathbb{Z}) \longrightarrow H^n(V,\mathbb{Z}/n\mathbb{Z}) \xrightarrow{e_n} \mathrm{Hom}(H_n(V,\widehat{\mathbb{Z}}),\mathbb{Z}/n\mathbb{Z})) \to 0$$

for every variety V over $\overline{K}$, where the map e_n is induced by the evaluation pairing and the Ext groups are for the category of pro-groups. In particular there is a natural isomorphism $H^1(V,\mathbb{Z}/n\mathbb{Z}) \cong \mathrm{Hom}(H_1(V,\widehat{\mathbb{Z}}),\mathbb{Z}/n\mathbb{Z}))$. Therefore the pullback map $f^*: H^1(\overline{X},\mathbb{Z}/n\mathbb{Z}) \to H^1(\overline{Y},\mathbb{Z}/n\mathbb{Z})$ is zero. Because the map d_n is functorial, the claim now follows. □

Assume that $X(K) \neq \varnothing$. For every $x, y \in X(hK)$ such that $x \sim_1 y$, let $\epsilon_1^X(x, y)$ denote the image of the obstruction class $\delta_1^X(x, y)$ under the homomorphism

$$H_*: H^2(K, \pi_2(\overline{X})) \to H^2(K, H_2(\overline{X}, \widehat{\mathbb{Z}}))$$

induced by the Hurewicz map

$$H: \pi_2(\overline{X}) \to H_2(\overline{X}, \widehat{\mathbb{Z}}).$$

Proposition 10.8. *Let X be a geometrically irreducible smooth quasiprojective variety over K, and let $x, y \in X(K)$ be étale-Brauer equivalent. Then we have*

$$\epsilon_1^X(\iota_{X/K}(x), \iota_{X/K}(y)) = 0.$$

Note that the claim is meaningful because $\iota_{X/K}(x) \sim_1 \iota_{X/K}(y)$ by Lemma 8.7.

Proof. It will be enough to show that

$$\delta_1^X(\iota_{X/K}(x), \iota_{X/K}(y))_n = 0$$

for every $n \in \mathbb{N}$. By the universal coefficient theorem quoted above and by local Tate duality, every element of $H^2(K, H_2(\overline{X}, \mathbb{Z}/n\mathbb{Z}))$ annihilated by the pairing $\{\cdot,\cdot\}_n$ must be zero. So it will be enough to show that

$$\{\delta_1^X(\iota_{X/K}(x), \iota_{X/K}(y))_n, c\}_n = 0$$

for every $c \in H^0(K, H^2(\overline{X}, \mu_n))$. Let $f: Y \to X$ and $\tilde{x}, \tilde{y} \in Y(K)$ be as in Lemma 10.7. By Lemma 10.2 and the naturality of obstruction classes it will be enough to show that

$$\{\delta_1^Y(\iota_{Y/K}(x'), \iota_{Y/K}(y'))_n, f^*(c)\}_n = 0$$

for every $c \in H^0(K, H^2(\overline{X}, \mu_n))$. Because $d_n(f^*(c)) = 0$, this claim follows from Lemmas 8.6 and 10.5. □

The following result is Theorem 1.1 of the introduction:

Theorem 10.9. *Let K be a finite extension of $\mathbb{Q}_p$, and let X be a smooth quasiprojective variety over K. Then étale-Brauer equivalence and H-equivalence coincide on $X(K)$.*

Proof. Let $x, y \in X(K)$ be étale-Brauer equivalent. We need to show that they are H-equivalent. We may assume without loss of generality that X is geometrically irreducible. We already noted that $\iota_{X/K}(x) \sim_1 \iota_{X/K}(y)$. Also note that by Theorem 7.6 it will be enough to show that $\delta_1^X(\iota_{X/K}(x), \iota_{X/K}(y)) = 0$ since the cohomological dimension of K is 2, so in this case the obstruction classes $\delta_n^X(\iota_{X/K}(x), \iota_{X/K}(y))$ will be zero for every $n \geq 2$, too.

Fix an element s of $s_{X/K}(x)$. For every open characteristic subgroup $\mathcal{H}$ of $\pi_1(\overline{X}, \eta)$ let $\mathcal{H}' \subseteq \pi_1(X, \eta)$ be the subgroup generated by $\mathcal{H}$ and the image of s; this is an open subgroup. Moreover, for every such $\mathcal{H}$ let $f_{\mathcal{H}} : Y_{\mathcal{H}} \to X$ be the connected finite étale cover such that the image of $\pi_1(Y)$ with respect to $\pi_1(f_{\mathcal{H}})$ is $\mathcal{H}'$. By construction there is an $x_{\mathcal{H}} \in Y_{\mathcal{H}}(K)$ such that $f_{\mathcal{H}}(x_{\mathcal{H}}) = x$. Because x and y are étale-Brauer equivalent, there is a $y_{\mathcal{H}} \in Y_{\mathcal{H}}(K)$ such that $f_{\mathcal{H}}(y_{\mathcal{H}}) = y$, and $x_{\mathcal{H}}, y_{\mathcal{H}}$ are étale-Brauer equivalent. By the naturality of the obstruction classes, the cohomology classes

$$\epsilon_1^{Y_{\mathcal{H}}}(x_{\mathcal{H}}, y_{\mathcal{H}}) \in H^2(K, H_2(\overline{Y}_{\mathcal{H}}, \widehat{\mathbb{Z}}))$$

furnish an element of $\lim_{(Y,y')\in \mathrm{Fet}(\overline{X},x')} H^2(K, H_2(Y, \widehat{\mathbb{Z}}))$, where x' is a $\overline{K}$-valued point of X, which is the image of $\delta_1^X(x, y)$ with respect to the map

$$H^2(K, \pi_2(\overline{X})) \cong H^2\Bigl(K, \lim_{(Y,y')\in \mathrm{Fet}(\overline{X},x')} H_2(Y, \widehat{\mathbb{Z}})\Bigr) \longrightarrow \lim_{(Y,y')\in \mathrm{Fet}(\overline{X},x')} H^2(K, H_2(Y, \widehat{\mathbb{Z}}))$$

furnished by the map $b_X \circ a_X$, which is an isomorphism by Theorem 4.3. By Proposition 10.8 the classes $\epsilon_1^{Y_{\mathcal{H}}}(x_{\mathcal{H}}, y_{\mathcal{H}})$ are zero, and hence the theorem holds. □

11. H-equivalence over the field of real numbers

Definition 11.1. Let X be any scheme. Recall that a quadratic space over X is a vector bundle $\mathcal{E}$ over X, that is, a locally free $\mathcal{O}_X$-module of finite rank, and an isomorphism $h : \mathcal{E} \to \mathcal{E}^*$, where $\mathcal{E}^*$ denotes the dual of $\mathcal{E}$, which is symmetric, that is, the composition

$$\mathcal{E} \longrightarrow \mathcal{E}^{**} \xrightarrow{h^*} \mathcal{E}^*$$

is equal to h, where the first map is the natural isomorphism of $\mathcal{E}$ with its bidual, and h^* is the dual of h. In the special case when $X = \mathrm{Spec}(K)$, for K a field, this concept is the same as the concept of a nondegenerate quadratic form over K.

Definition 11.2. Consider the case when $K=\mathbb{R}$. By Sylvester's theorem every nondegenerate quadratic form q over $\mathbb{R}$ is isomorphic to a diagonal form

$$\langle\underbrace{1,1,\ldots,1}_{m},\underbrace{-1,\ldots,-1}_{n}\rangle,$$

and the ordered pair (m,n) only depends on the isomorphism class of q. Let $\rho(q)=m+n$ denote the rank of q and let $\sigma(q)=m-n$ denote the signature of q, respectively. By the above, nondegenerate forms over $\mathbb{R}$ are classified by their rank and signature. Let X be a smooth variety over $\mathbb{R}$, let $U\subset X(\mathbb{R})$ be a connected component, and let $q=(\mathscr{E},h)$ be a quadratic space over X. Then for every $x\in U$ the pullback $x^*(q)$ has the same signature, which we will call the signature of q on U.

Theorem 11.3. *Let X be a smooth variety over $\mathbb{R}$ which is either affine or projective. Let $U,V\subset X(\mathbb{R})$ be two different connected components. Then there is a quadratic space $q=(\mathscr{E},h)$ over X such that the signature of q on U is zero and the signature of q on V is nonzero.*

Proof. See Theorem 1.1.1 in [Mahé 1982] for the affine case, and the main result of [Houdebine and Mahé 1982] for the projective case. □

Definition 11.4. Let K be a field whose characteristic is not 2. Every nondegenerate quadratic form q is isomorphic to a diagonal form

$$\langle a_1,a_2,\ldots,a_n\rangle.$$

The Stiefel–Whitney classes of the form q above are defined as the cup product (see [Milnor 1970])

$$w(q)=1+w_1(q)+\cdots+w_n(q)=(1+\delta(a_1))(1+\delta(a_2))\cdots(1+\delta(a_n)),$$

where $w_i(q)\in H^i(K,\mathbb{Z}/2\mathbb{Z})$ and

$$\delta:K^*\to H^1(K,\mathbb{Z}/2\mathbb{Z})$$

is the boundary map of the Kummer exact sequence

$$0\longrightarrow\mathbb{Z}/2\mathbb{Z}\longrightarrow\mathbb{G}_m\xrightarrow{x\mapsto x^2}\mathbb{G}_m\longrightarrow 0.$$

The Stiefel–Whitney classes are independent of the diagonalisation of q.

Remark 11.5. Assume again that K is the field of real numbers. Then as a graded algebra

$$H^*(\mathbb{R},\mathbb{Z}/2\mathbb{Z})\cong\mathbb{F}_2[t],$$

where t is the generator of the group $H^1(\mathbb{R}, \mathbb{Z}/2\mathbb{Z})$ of order two. Let q be a nondegenerate quadratic form q over $\mathbb{R}$ isomorphic to a diagonal form

$$\langle \underbrace{1, 1, \dots, 1}_{m}, \underbrace{-1, \dots, -1}_{n} \rangle;$$

then by the above

$$w(q) = (1+t)^n = 1 + nt + \cdots + t^n.$$

Esnault, Kahn and Viehweg [Esnault et al. 1993] constructed Stiefel–Whitney classes for any quadratic space $q = (\mathscr{E}, h)$ over a $\mathbb{Z}[1/2]$-scheme X, which lives in mod-2 étale cohomology

$$w_i(q) \in H^i(X, \mathbb{Z}/2\mathbb{Z}),$$

is functorial over the category of $\mathbb{Z}[1/2]$-schemes, and specialises to the construction above when X is the spectrum of a field. We will use these classes to separate connected components of real points of varieties defined over $\mathbb{R}$.

Proposition 11.6. *Let X be a smooth variety over $\mathbb{R}$ which is either affine or projective. Let $U, V \subset X(\mathbb{R})$ be two different connected components. Then there is a natural number i and a cohomology class $c \in H^i(X, \mathbb{Z}/2\mathbb{Z})$ over X such that for every $x \in U$ the pullback $x^*(c) \in H^i(\mathbb{R}, \mathbb{Z}/2\mathbb{Z})$ is zero and for every $x \in V$ the pullback $x^*(c) \in H^i(\mathbb{R}, \mathbb{Z}/2\mathbb{Z})$ is nonzero.*

Proof. By Theorem 2.9 and Proposition 2.10, there is a $c \in H^0(X, \mathbb{Z}/2\mathbb{Z})$ such that $x^*(c) \in H^0(\mathbb{R}, \mathbb{Z}/2\mathbb{Z})$ is zero for every $x \in U$ and $y^*(c) \in H^0(\mathbb{R}, \mathbb{Z}/2\mathbb{Z})$ is nonzero for every $y \in V$ if U and V lie on two different connected components of $X(\mathbb{C})$. Therefore we may assume that X is geometrically connected without loss of generality. Let $q = (\mathscr{E}, h)$ be a quadratic space over X such that the signature of q on U is zero and the signature of q on V is nonzero. We may assume that the signature of q on V is negative by taking $(\mathscr{E}, -h)$ instead, if necessary. Because X is connected, the rank of the vector bundle $\mathscr{E}$ is constant on X. This rank is even, say $2m$, because the signature of q on U is zero. Then the signature of q on V is $2m-2n$, where n is a positive integer bigger than m. By Remark 11.5 above, we have $x^*(w_n(q)) = w_n(x^*(q)) = 0$ for every $x \in U$ and $x^*(w_n(q)) = w_n(x^*(q)) = t^n$ for every $x \in V$. The claim follows. □

Definition 11.7. For every morphism of sites $m : \mathscr{C} \to \mathscr{C}'$, let $m^* : \mathscr{C}' \to \mathscr{C}$ denote the functor underlying m. Let $\mathscr{C}$ be a Grothendieck site. A left action α of a group Γ on $\mathscr{C}$ is a morphism $\alpha(g) : \mathscr{C} \to \mathscr{C}$ of sites for each $g \in \Gamma$ such that $\alpha(1)$ is the identity map of $\mathscr{C}$ and $\alpha(gh) = \alpha(g) \circ \alpha(h)$ for every $g, h \in \Gamma$. When Γ is profinite, we say that the action α is continuous if for every morphism $h : U \to V$ in $\mathscr{C}$ there is an open subgroup Δ of Γ such that $\alpha(g)^*(U) = U, \alpha(g)^*(V) = V$

and $\alpha(g)^*(h) = h$ for every $g \in \Delta$. Assume now that Γ is a profinite group. By a Γ-site $(\mathscr{C}, \alpha)$ we mean a Grothendieck site $\mathscr{C}$ with a continuous left action α of Γ on $\mathscr{C}$. As usual we will drop α from the notation whenever this is convenient.

Example 11.8. A basic example of a Γ-site is the Grothendieck site Γ-Sets, where the coverings are surjective maps, equipped with the left action α such that $\alpha(g)^*(U) = U$ for every object U of Γ-Sets and every $g \in \Gamma$, and for every morphism $h : U \to V$ and $g \in \Gamma$ the map $\alpha(g)^*(h) : U \to V$ is given by the rule $x \mapsto gh(x)$. By a slight abuse of notation we will let Γ-Sets denote this Γ-site, too.

Definition 11.9. Let $(\mathscr{C}, \alpha)$ be a Γ-site. A Γ-invariant object of $(\mathscr{C}, \alpha)$ is an object U of $\mathscr{C}$ such that $\alpha(g)^*(U) = U$ for every $g \in \Gamma$. A Γ-equivariant morphism of $(\mathscr{C}, \alpha)$ is a morphism $h : U \to V$ of $\mathscr{C}$ such that U and V are Γ-invariant objects and $h \circ \alpha(g)^*(\mathrm{id}_U) = \alpha(g)^*(\mathrm{id}_V) \circ h$ for every $g \in \Gamma$. Let $\mathscr{C}^\Gamma$ denote the category whose objects are Γ-invariant objects of $\mathscr{C}$ and whose morphisms are Γ-equivariant maps between these. Since the composition of Γ-equivariant morphisms are Γ-equivariant, with these morphisms $\mathscr{C}^\Gamma$ is indeed a category. Let $\boldsymbol{T}$ denote the Grothendieck topology of the site $\mathscr{C}$; that is, for every object U of $\mathscr{C}$ let $\boldsymbol{T}(U)$ denote the collection of covering sieves of U. We say that a sieve S on $U \in \mathrm{ob}(\mathscr{C}^\Gamma)$ is Γ-invariant if for every $V \in \mathrm{ob}(\mathscr{C})$ and every $h \in S(V)$ there is a $W \in \mathrm{ob}(\mathscr{C}^\Gamma)$, a morphism $h' \in S(W)$ which is Γ-equivariant, and a morphism $h'' : V \to W$ such that $h = h' \circ h''$. For every Γ-invariant S as above, let S^Γ denote the sieve on U in the category $\mathscr{C}^\Gamma$ given by the rule

$$S^\Gamma(V) = S(V) \cap \mathrm{Hom}_{\mathscr{C}^\Gamma}(V, U).$$

For every $U \in \mathrm{ob}(\mathscr{C}^\Gamma)$ let $\boldsymbol{T}^\Gamma(U)$ denote the following collection of sieves S in the category $\mathscr{C}^\Gamma$:

$$\boldsymbol{T}^\Gamma(U) = \{S^\Gamma \mid S \text{ is in } \boldsymbol{T}(U) \text{ and is } \Gamma\text{-invariant}\}.$$

Example 11.10. Let K be a field and let $\Gamma_K = \mathrm{Gal}(\overline{K}|K)$ denote the absolute Galois group of K as above. Let X be a locally Noetherian scheme over K and let $\mathscr{C}$ denote the small étale site of the base change of X to K. Then $\mathscr{C}$ is naturally equipped with the structure of a Γ_K-site, induced by the action of Γ_K on $\overline{K}$. By étale descent the category $\mathscr{C}^{\Gamma_K}$ is equivalent to the small étale category of X and $\boldsymbol{T}^{\Gamma_K}$ is the étale topology of X on it. In particular $\boldsymbol{T}^{\Gamma_K}$ is a Grothendieck topology.

Definition 11.11. Assume now that $\mathscr{C}$ satisfies the conditions in Chapters 8 and 9 of [Artin and Mazur 1969] and $\boldsymbol{T}^\Gamma$ is a Grothendieck topology on $\mathscr{C}^\Gamma$. In particular we suppose that $\mathscr{C}$ is closed under finite coproducts and that it is locally connected in the sense of [Artin and Mazur 1969, Sections 9.1–9.2, pp. 111–112]. Let U be a Γ-invariant object of $(\mathscr{C}, \alpha)$. Then Γ acts on the set $\pi_0(U)$ of connected components of U and this action makes $\pi_0(U)$ into a Γ-set. Let $\pi_\Gamma(U)$ denote this Γ-set. We say that a simplicial object of $\mathscr{C}$ is Γ-invariant if it is a simplicial object of $\mathscr{C}^\Gamma$. If X_* is a Γ-invariant simplicial object of $(\mathscr{C}, \alpha)$, then the face and degeneracy maps of X_* induce Γ-equivariant maps between the Γ-set $\pi_\Gamma(X_n)$ which makes the collection $\{\pi_\Gamma(X_n)\}_{n=1}^\infty$ into an object of Γ-SSets, which we will denote by $\pi_\Gamma(X_*)$. Since the site $(\mathscr{C}^\Gamma, \boldsymbol{T}^\Gamma)$ inherits the good properties of the site $\mathscr{C}$, we may apply [Artin and Mazur 1969, Corollary 8.13(i), p. 105] to conclude that the functor

$$X_* \mapsto \pi_\Gamma(X_*)$$

above induces an object of Pro-Ho(Γ-SSets). We will call the later the Γ-equivariant homotopy type of $\mathscr{C}$ and denote it by $\Pi(\mathscr{C})$.

Remark 11.12. Let K, Γ_K and X be as in Example 11.10, and let $\mathscr{C}$ denote the small étale site of the base change of X to $\bar{K}$, as above. Then the Γ_K-equivariant homotopy type of $\mathscr{C}$ is just the relative étale homotopy type $\mathrm{Et}_{/K}(X)$ of X as defined by Harpaz and Schlank.

Example 11.13. Assume now that Γ is a finite group and let X be a locally connected, Hausdorff, paracompact topological space equipped with a continuous left Γ-action. Let $\mathscr{C}$ be the ordinary site on the coproducts of open subsets of X. Then $\mathscr{C}$ is naturally equipped with the structure of a Γ-site, induced by the action of Γ on X. Moreover, the quotient $\Gamma\backslash X$ of X by the action of Γ is also a locally connected, Hausdorff, paracompact topological space, and the category $\mathscr{C}^\Gamma$ is the category of coproducts of open subsets of $\Gamma\backslash X$ and $\boldsymbol{T}^\Gamma$ is the ordinary site of $\Gamma\backslash X$ on it. In particular $\boldsymbol{T}^\Gamma$ is a Grothendieck topology.

Definition 11.14. Let Γ and X be as above. We say that X is Γ-contractible if there is a subgroup $\Delta \leq \Gamma$ such that the quotient Γ/Δ, equipped with the discrete topology and the natural left Γ-action, is Γ-equivariantly homotopy-equivalent to X. Assume now that every open subset of X is paracompact, and that X is locally Γ-equivariantly contractible; that is, for every (finite) orbit $O \in X$ and every Γ-invariant open $U \subseteq X$ containing O there an open Γ-invariant and Γ-contractible $V \subseteq X$ such that $V \subseteq U$ and $O \subseteq V$.

Note that when Γ is finite Γ-SSets is just the usual category of simplicial sets with a Γ-action, and moreover Goerss' notion of homotopy fixed point spaces coincides with the usual one. Also note that the singular complex S_*X of X is equipped with an action of Γ which makes it an object of Γ-SSets.

Theorem 11.15. *Let U_* be a Γ-invariant hypercovering of $\mathscr{C}$ such that every Γ-orbit of connected components of every U_n is Γ-contractible. Then the simplicial Γ-set $\pi_\Gamma(U_*)$ is isomorphic to the simplicial Γ-set S_*X in* Ho(Γ-SSets).

Proof. Let S_*U_n denote the singular complex of U_n. Then S_*U_* is a bisimplicial object in Γ-Sets. We denote by $(DU)_*$ its diagonal simplicial Γ-set $(DU)_n = S_nU_n$. Then we have obvious maps of simplicial Γ-sets

$$\begin{tikzcd} & (DU)_* \arrow[dl, "\alpha"'] \arrow[dr, "\beta"] & \\ \pi_\Gamma(U_*) & & S_*X \end{tikzcd}$$

and we claim that these two maps are homotopy equivalences in Ho(Γ-SSets), which will prove the theorem.

For every subgroup $\Delta \leq \Gamma$ and every Γ-set Y, let Y^Δ denote the subset of Y fixed by Δ. Similarly, for every simplicial Γ-set Y_* let Y_*^Δ denote the simplicial set $\{Y_n^\Delta\}_{n=1}^\infty$ such that the face and degeneracy maps are the restrictions of such maps of the simplicial set Y_*. By the definition of the strict model structure we need to show that the maps

$$\begin{tikzcd} & (DU)_*^\Delta \arrow[dl, "\alpha|_{(DU)_*^\Delta}"'] \arrow[dr, "\beta|_{(DU)_*^\Delta}"] & \\ \pi_\Gamma(U_*)^\Delta & & (S_*X)^\Delta \end{tikzcd}$$

of simplicial sets are homotopy equivalences in Ho(SSets) for every Δ as above.

Note that for every Γ-invariant hypercovering V_* of X the simplicial object V_*^Γ in the category of disjoint unions of open sets of the closed subspace X^Δ of X is a hypercovering, too, since it is the pullback of the hypercovering V_* onto X^Δ with respect to the inclusion map. Because the U_n are Γ-contractible we have $\pi_\Gamma(U_*)^\Delta = \pi_0(U_*^\Delta)$. Moreover, every Δ-invariant singular simplex of X must lie in X^Δ, so $(S_*X)^\Delta$ is the singular complex S_*X^Δ of X^Δ. Similarly let $S_*U_n^\Delta$ denote the singular complex of U_n^Δ. Then $S_*U_*^\Delta$ is a bisimplicial set. We denote by $(DU^\Delta)_*$ its diagonal simplicial set $(DU^\Delta)_n = S_nU_n^\Delta$. Then we have $(DU)_*^\Delta = (DU^\Delta)_*$, so we only need to show that the analogues

$$\begin{tikzcd} & (DU^\Delta)_* \arrow[dl, "\alpha^\Delta"'] \arrow[dr, "\beta^\Delta"] & \\ \pi_0(U_*^\Delta) & & S_*X^\Delta \end{tikzcd}$$

of the maps α and β for the topological space X^Δ are homotopy equivalences in Ho(SSets) for every Δ as above. Since every Γ-orbit of connected components of every U_n is Γ-contractible, the connected components of U_n^Δ are contractible. Similarly, X^Δ is locally contractible, since X is locally Γ-contractible. Because

X^Δ is a closed subspace of a paracompact topological space, it is also paracompact. The claim now follows from [Artin and Mazur 1969, Theorem 12.1, p. 129]. □

Corollary 11.16. *Let $\mathscr{C}$ be the ordinary site on the coproducts of open subsets of a Hausdorff topological space X equipped with a continuous left Γ-action. Assume that every open subset of X is paracompact and that X is locally Γ-equivariantly contractible. Then the pro-object $\Pi(\mathscr{C})$ is canonically isomorphic to the element S_*X in* Pro-Ho(Γ-SSets).

Proof. Because those Γ-invariant hypercoverings U_* of $\mathscr{C}$ such that every connected component of every U_n is Γ-contractible are cofinal by assumption, the claim follows immediately. □

Definition 11.17. Let again Γ be an arbitrary profinite group. A morphism $m : (\mathscr{C}, \alpha) \to (\mathscr{C}', \alpha')$ of Γ-sites is a morphism of sites $m : \mathscr{C} \to \mathscr{C}'$ such that for every $g \in \Gamma$ and every morphism $h : U \to V$ of $\mathscr{C}'$ we have

$$\begin{aligned} \alpha(g)^*(m^*(U)) &= m(\alpha'(g)^*(U)), \\ \alpha(g)^*(m^*(V)) &= m^*(\alpha'(g)^*(V)), \\ \alpha(g)^*(m^*(h)) &= m^*(\alpha'(g)^*(h)). \end{aligned}$$

For every such m, the underlying functor m^* carries Γ-invariant hypercoverings to Γ-invariant hypercoverings, and so it furnishes a map

$$\Pi(m) : \Pi(\mathscr{C}) \to \Pi(\mathscr{C}')$$

in Pro-Ho(Γ-SSets) (when these are defined). The map $\Pi(m)$ in turn induces a map

$$\Pi(m)(E\Gamma) : \Pi(\mathscr{C})(E\Gamma) \to \Pi(\mathscr{C}')(E\Gamma)$$

of homotopy fixed points.

Definition 11.18. A Γ-invariant point (or more conveniently Γ-point) of a Γ-site $(\mathscr{C}, \alpha)$ is a morphism $p : \Gamma\text{-Sets} \to \mathscr{C}^\Gamma$ of Γ-sites. Note that the composition of p^* and the functor Γ-Sets $\to$ Sets forgetting the Γ-action is a point of the site $\mathscr{C}$ in the sense of [Artin and Mazur 1969, Sections 8.3-8.4, pp. 95–96], which perhaps justifies our terminology. We will let $\mathscr{C}(\Gamma)$ denote the set of Γ-points of $(\mathscr{C}, \alpha)$. Note that the homotopy type $\Pi(\Gamma\text{-Sets})$ is contractible (this is clear from Corollary 11.16, too), so the set $\Pi(\Gamma\text{-Sets})(E\Gamma)$ has one element. For every $p \in \mathscr{C}(\Gamma)$ let $\iota_{\mathscr{C}}(p) \in \Pi(\mathscr{C})(E\Gamma)$ denote the image of $\Pi(\Gamma\text{-Sets})(E\Gamma)$ with respect to $\Pi(p)$. Clearly

$$\iota_{\mathscr{C}} : \mathscr{C}(\Gamma) \to \Pi(\Gamma\text{-Sets})(E\Gamma)$$

is a natural transformation.

Example 11.19. Let K, $\Gamma_K = \mathrm{Gal}(\overline{K}|K)$, X and $\mathscr{C}$ be as in Example 11.10. Since the small étale site of $\mathrm{Spec}(\overline{K})$ is isomorphic to Γ_K-Sets as a Γ_K-site, every K-valued point of X supplies a Γ_K-point of the site $\mathscr{C}$. Therefore the map $\iota_{\mathscr{C}}$ introduced above is a generalisation of the map $\iota_{X/K}$. Similarly, if Γ is a finite group, X is a locally connected, Hausdorff, paracompact topological space equipped with a continuous left Γ-action and $\mathscr{C}$ is as in Example 11.13, then every point of X fixed by Γ furnishes a Γ-point of the site $\mathscr{C}$, and hence the restriction of $\iota_{\mathscr{C}}$ onto X^Γ is a map $X^\Gamma \to \Pi(\mathscr{C})(E\Gamma)$.

Proposition 11.20. *Let K be the field $\mathbb{R}$ of real numbers, and let X be a variety over K. Then two K-rational points of X are H-equivalent if they are in the same connected component of the topological space $X(K)$.*

Proof. Let $\Gamma = \Gamma_K = \mathrm{Gal}(\overline{K}|K)$ be the group of two elements, and let $\mathscr{C}'$ denote the small étale site of the base change of X to $\overline{K}$. Moreover, let $\mathscr{C}$ be the ordinary site on the coproducts of open subsets of $X(\overline{K})$ with respect to its usual topology. In addition to these Γ-sites we also introduce the Γ-site $\mathscr{C}''$ whose objects are topological spaces X' lying over the topological space of $X(\overline{K})$ such that the map $X' \to X$ is a local isomorphism, i.e., such that every point $x \in X'$ has a neighbourhood which is isomorphic onto its image. Since any étale map of schemes $X' \to X$ over $\overline{K}$ is a local isomorphism on the underlying topological spaces, and since an open set is in $\mathscr{C}''$, we have morphisms of Γ-sites

$$\begin{tikzcd} & \mathscr{C}'' \arrow[dl, "m"'] \arrow[dr, "m'"] & \\ \mathscr{C} & & \mathscr{C}' \end{tikzcd}$$

Now it is clear from the definition of local isomorphisms that every Γ-invariant hypercovering of $\mathscr{C}''$ is dominated by a Γ-invariant hypercovering of $\mathscr{C}$. Thus the map $\Pi(m) : \Pi(\mathscr{C}'') \to \Pi(\mathscr{C})$ is a homotopy equivalence in Pro-Ho(Γ-SSets), and so

$$\Pi(m)(E\Gamma) : \Pi(\mathscr{C}'')(E\Gamma) \to \Pi(\mathscr{C})(E\Gamma)$$

is a bijection. Therefore by the naturality of the maps $\iota_{\mathscr{C}}$, $\iota_{\mathscr{C}'}$ and $\iota_{\mathscr{C}''}$ it will be enough to show that for every pair x, y of K-rational points of X lying in the same connected component of the topological space $X(K)$ we have $\iota_{\mathscr{C}}(x) = \iota_{\mathscr{C}}(y)$. Let $f : [0, 1] \to X(K)$ be a continuous path connecting x with y; that is, we have $f(0) = x$ and $f(1) = y$. By naturality again it will be enough to show that $\iota_{\mathscr{D}}(0) = \iota_{\mathscr{D}}(1)$, where $\mathscr{D}$ is the ordinary site on the coproducts of open subsets of $[0, 1]$ with respect to its usual topology, equipped with the trivial Γ-action. But the interval $[0, 1]$ is contractible to a point Γ-equivariantly, so $\Pi(\mathscr{D})(E\Gamma)$ is a one-element set by Corollary 11.16. $\square$

Remark 11.21. It is not difficult to push the arguments of this proof a little bit further to prove an equivariant version of [Artin and Mazur 1969, Theorem 12.9, p. 142], using an equivariant analogue of the profinite completion functor, but we will not pursue this further since it would take us too far away from our original project. However in a forthcoming publication we will in fact prove such a claim in a much more general context.

The following result is Theorem 1.2 of the introduction:

Theorem 11.22. *Let K be the field $\mathbb{R}$ of real numbers, and let X be a smooth affine or projective variety over K. Then two K-rational points of X are H-equivalent if and only if they are in the same connected component of the topological space $X(K)$.*

Proof. By Lemma 8.5 and Proposition 11.6, two H-equivalent real points of X must be in the same component. On the other hand, by Proposition 11.20 two real points of X in the same connected component must be H-equivalent. □

12. The homotopy section principle

Notation 12.1. Recall that for every field of characteristic zero the topological $\mathrm{Gal}(\overline{K}|K)$-module $\widehat{\mathbb{Z}}(1)$ is defined as the projective limit $\varprojlim_{n\in\mathbb{N}} \mu_n$, where the directed set structure on $\mathbb{N}$ is furnished by divisibility and for every $m, n \in \mathbb{N}$ such that $m \mid n$ the transition map $\mu_n \to \mu_m$ is multiplication by n/m. For every number field K let $|K|$ denote the set of places of K, and for every $v \in |K|$ let K_v denote the completion of K with respect to v. For every $v \in |K|$ fix an embedding $j_v : \overline{K} \to \overline{K}_v$ of K-extensions. For every $k \in \mathbb{N}$ and every discrete $\mathrm{Gal}(\overline{K}|K)$-module M let $Ш^k(K, M)$ denote the subgroup

$$Ш^k(K, M) = \mathrm{Ker}\Bigg(\prod_{v\in|K|} j_{v*} : H^k(K, M) \to \prod_{v\in|K|} H^k(K_v, M)\Bigg)$$

of $H^k(K, M)$, where j_{v*} denotes the restriction map induced by j_v for every $v \in |K|$. For every topological $\mathrm{Gal}(\overline{K}|K)$-module M which is a projective limit $\varprojlim_{i\in I} M_i$ of discrete $\mathrm{Gal}(\overline{K}|K)$-modules let $Ш^k(K, M)$ denote $\varprojlim_{i\in I} Ш^k(K, M_i)$.

Lemma 12.2. *We have $Ш^2(K, \widehat{\mathbb{Z}}(1)) = 0$ for every number field K.*

Proof. By [Milne 1986, Theorem 4.10(a), p. 70] the group $Ш^2(K, \widehat{\mathbb{Z}}(1))$ is zero if and only if $Ш^1(K, \mathbb{Q}/\mathbb{Z})$ is, where we equip $\mathbb{Q}/\mathbb{Z}$ with the discrete topology and the trivial $\mathrm{Gal}(\overline{K}|K)$-action. We may identify $Ш^1(K, \mathbb{Q}/\mathbb{Z})$ with the kernel of the map

$$\mathrm{Hom}(\Gamma_K, \mathbb{Q}/\mathbb{Z}) \to \prod_{v\in|K|} \mathrm{Hom}(\Gamma_{K_v}, \mathbb{Q}/\mathbb{Z})$$

furnished by restriction onto the family of subgroups Γ_{K_v} of Γ_K for all $v \in |K|$. The claim now follows from the Chebotarev density theorem. □

Proposition 12.3. *Assume that K is a number field and $X(K)$ is nonempty. For every $x, y \in X(hK)$ such that $x \sim_2 y$ we have*

$$x = y \iff r_{v*}(x) = r_{v*}(y) \quad (v \in |K|, v \text{ is real}).$$

Proof. The product of the restriction maps

$$\prod_{\substack{v \in |K| \\ v \text{ is real}}} j_{v*} : H^n(K, M) \to \prod_{\substack{v \in |K| \\ v \text{ is real}}} H^n(K_v, M)$$

is injective for every integer $n \geq 3$ and every discrete Galois module M over K by [Milne 1986, Theorem 4.10(c), p. 70]. Hence for every $n \geq 2$ and every pair of sections $x, y \in X(hK)$ such that $x \sim_n y$ we get that

$$\delta_n^X(x, y) = 0 \iff \delta_n^{X_v}(r_{v*}(x), r_{v*}(y)) = 0 \quad (v \in |K|, v \text{ is real})$$

from naturality of the obstruction classes. The claim now follows from Theorem 7.6. □

Definition 12.4. Let K be for a moment any field of characteristic zero. Recall that two points $x, y \in X(K)$ are called directly R-equivalent if there is a rational map $f : \mathbb{P}^1_K \dashrightarrow X$ of K-varieties such that $f(0) = x$ and $f(\infty) = y$. The R-equivalence on $X(K)$ is the equivalence relation generated by direct R-equivalence. Let $X(K)/R$ denote the equivalence classes of this relation. Note that A-equivalence coincides with R-equivalence when X is projective by the valuative criterion of properness. In this case let

$$\iota_{X/K,R} : X(K)/R \to X(hK)$$

be the map furnished by Proposition 7.8.

Notation 12.5. It is particularly interesting to study $X(K)/R$ through the map $\iota_{X/K,R}$ when K is a number field. For every variety X defined over K and every $v \in |K|$ let X_v denote the base change of X to $\mathrm{Spec}(K_v)$. For every $v \in |K|$ the embedding j_v furnishes a map

$$r_{v*} : X(hK) \to X_v(hK_v)$$

by functoriality. Let $\mathbb{A}_K = \prod'_{v \in |K|} K_v$ denote the ring of adèles of K. Let $X(h\mathbb{A}_K)$ denote the image of $X(\mathbb{A}_K)$ with respect to the map

$$\prod_{v \in |K|} \iota_{X_v/K_v} : X_v(K_v) \to \prod_{v \in |K|} X_v(hK_v).$$

We define the set $\mathrm{Sel}(X/K)$ of Selmer homotopy fixed points of X to be

$$\mathrm{Sel}(X/K) = \Big(\prod_{v\in|K|} r_{v*}\Big)^{-1}(X(h\mathbb{A}_k)) \subseteq X(hK).$$

We are interested in the following natural generalisation of the Shafarevich–Tate conjecture:

Homotopy section principle (HSP). *Assume that X is smooth and projective. Then the map*

$$X(K)/R \xrightarrow{\iota_{X/K,R}} \mathrm{Sel}(X/K) \tag{12.5.1}$$

is injective and its image is dense with respect to the pro-discrete topology of $\mathrm{Sel}(X/K)$.

The claim above is obviously true if X does not have local points everywhere, and hence HSP should be considered as a new form of the local-global principle. The next proposition shows that HSP is indeed a generalisation of standard conjectures of this sort:

Proposition 12.6. *Let K be a number field.*

(a) *HSP holds for Brauer–Severi varieties and for nonsingular quadratic hypersurfaces $H \subset \mathbb{P}^n_K$ of positive dimension.*

(b) *Let X be a smooth, geometrically connected projective curve X of genus at least two over K. Then HSP holds for X if and only if a weak* (*local-global*) *form of Grothendieck's section conjecture holds for X.*

(c) *Let X be an abelian variety over K. Then HSP holds for X if and only if the Shafarevich–Tate conjecture holds for X.*

Proof. First assume that X is either a Brauer–Severi variety or a nonsingular quadratic hypersurface of positive dimension. When $\mathrm{Sel}(X/K)$ is empty there is nothing to prove. Assume now that $\mathrm{Sel}(X/K)$ is nonempty: then $X(\mathbb{A}_K)$ is nonempty, too. Because the local-global principle holds for X we get that $X(K)$ is also nonempty. In this case $X(K)/R$ consists of one element, and hence it will be enough to show that $\mathrm{Sel}(X/K)$ also has one element. Let $x, y \in \mathrm{Sel}(X/K)$. Because for every $v \in |K|$ the set $X_v(K_v)/R$ has one element, we get that $r_{v*}(x) = r_{v*}(y)$ for every such v.

By [Artin and Mazur 1969, Corollary 12.13, p. 144] we know that $X_{\overline{K}}$ is weakly homotopy equivalent to $X_{\mathbb{C}}$. Because $\pi_1(X(\mathbb{C})) = \{1\}$ (either because $X_{\mathbb{C}}$ is isomorphic to $\mathbb{P}^n_{\mathbb{C}}$ for $n = \dim(X)$ or by the Lefschetz hyperplane section theorem), we get $\pi_1(X_{\overline{K}}) = \{1\}$ by Theorem 2.9. Therefore $x \sim_1 y$. Moreover,

$$\pi_2(X_{\overline{K}}) = H_2(X_{\overline{K}}, \widehat{\mathbb{Z}}) = \mathrm{Hom}(H^2(X_{\overline{K}}, \widehat{\mathbb{Z}}), \widehat{\mathbb{Z}})$$

by [Artin and Mazur 1969, Corollary 6.2, p. 70]. When X is a Brauer–Severi variety, since it has a rational point it is isomorphic to $\mathbb{P}^n_K$. Therefore $H^2(X_{\overline{K}}, \widehat{\mathbb{Z}}) \cong \widehat{\mathbb{Z}}(-1)$, and hence $\pi_2(X_{\overline{K}}) \cong \widehat{\mathbb{Z}}(1)$. Because $r_{v*}(x) = r_{v*}(y)$ for every $v \in |K|$, we have $\delta_2^X(x, y) \in Ш^2(K, \pi_2(X_{\overline{K}}))$, so this obstruction class vanishes by Lemma 12.2. So we get that $x \sim_2 y$.

When X is a quadratic hypersurface of dimension at least 3, its embedding $X \to \mathbb{P}^{n+1}_K$ as such a hypersurface induces an isomorphism

$$H^2(X_{\overline{K}}, \widehat{\mathbb{Z}}) \cong H^2(\mathbb{P}^{n+1}_{\overline{K}}, \widehat{\mathbb{Z}}) \cong \widehat{\mathbb{Z}}(-1)$$

by the Lefschetz hyperplane section theorem, and hence we may conclude as above that $x \sim_2 y$. The only remaining case is of a quadratic surface. In this case either

$$H^2(X_{\overline{K}}, \widehat{\mathbb{Z}}) \cong \widehat{\mathbb{Z}}(-1) \oplus \widehat{\mathbb{Z}}(-1),$$

when both pencils of lines on X are defined over K, or it is the induction of $\widehat{\mathbb{Z}}(-1)$ from a quadratic extension of K. Clearly in the first case the group $Ш^2(K, \pi_2(X_{\overline{K}}))$ still vanishes by Lemma 12.2, while in the second case this claim follows from Shapiro's lemma and Lemma 12.2. Again we get that $x \sim_2 y$. Claim (a) now follows from Proposition 12.3.

Assume now that X is a smooth, geometrically connected projective curve of genus at least two over K. Then there is a commutative diagram

$$\begin{array}{ccc} X(K) & \xrightarrow{s_{X/K}} & \mathrm{Sec}(X/K) \\ \downarrow & & \downarrow \\ \prod\limits_{v\in|K|} X_v(K_v) & \xrightarrow{\prod_{v\in|K|} s_{X_v/K_v}} & \prod\limits_{v\in|K|} \mathrm{Sec}(X_v/K_v) \end{array}$$

where the vertical maps are the obvious maps. The weak local-global form of Grothendieck's section conjecture asserts that the diagram above is cartesian. We also know that in this case $s_{X/K}$ is injective and by Faltings' theorem $X(K)$ is finite. In particular, in this case the map in (12.5.1) has dense image if and only if it is surjective.

Let's assume first that the weak local-global form of Grothendieck's section conjecture holds for X, and show that HSP holds for X. This is trivial when $\mathrm{Sel}(X/K)$ is empty, so we may assume that $\mathrm{Sel}(X/K)$ is nonempty. Then $\mathrm{Sec}(X/K)$ has an element whose image in $\prod_{v\in|K|} \mathrm{Sec}(X_v/K_v)$ lies in the image of $\prod_{v\in|K|} s_{X_v/K_v}$, and hence $X(K)$ is nonempty, by our assumption. The claim now follows from Corollary 5.5 and Lemma 7.7. Let us prove the converse. We may assume that $\mathrm{Sec}(X/K)$ has an element whose image in $\prod_{v\in|K|} \mathrm{Sec}(X_v/K_v)$ lies in the image of $\prod_{v\in|K|} s_{X_v/K_v}$ without loss of generality. By the main theorem of [Harpaz and

Schlank 2013] for smooth projective varieties, the set $\mathrm{Sel}(X/K)$ is nonempty if and only if the étale-Brauer set of X is. In our case the latter is nonempty (see [Stix 2011]), so we get that $\mathrm{Sel}(X/K)$ is nonempty, too. So by our hypothesis $X(K)$ is nonempty, so the claim follows from Corollary 5.5 and Lemma 7.7. Claim (b) is settled.

Finally, consider the case when X is an abelian variety over K. By Corollary 5.5 and Lemma 7.7 the map $j_{X/K}$ is a bijection. Moreover, there is a natural bijection

$$\mathrm{Sec}(X/K) \cong H^1\Bigl(K, \prod_{l \text{ is prime}} T_l(X)\Bigr), \tag{12.6.1}$$

where $T_l(X)$ denotes the l-th Tate module of X, and under this identification $s_{X/K}$ corresponds to the coboundary map furnished by Kummer theory. In particular $\iota_{X/K}$ is injective.

The image of $\mathrm{Sel}(X/K) \subseteq X(hK)$ with respect to the composition of $j_{X/K}$ and the isomorphism of (12.6.1) is $\mathrm{Sel}(K, X)^{\mathrm{tor}}$, where $\mathrm{Sel}(K, X)$ is the Selmer group of X over K. The quotient of $\mathrm{Sel}(K, X)^{\mathrm{tor}}$ by the closure of the image of $X(K)$ under the coboundary map is $Ш(K, X)^{\mathrm{tor}}$, where $Ш(K, X)$ is the Tate–Shafarevich group of X over K. Since the group $Ш(K, X)[n]$ is finite for every positive integer n, the group $Ш(K, X)^{\mathrm{tor}}$ is trivial if and only if $Ш(K, X)$ is finite. So claim (c) holds. □

Remark 12.7. It is tempting to believe that HSP should hold for every smooth projective variety because of its very general form, but this is not true. The fundamental reason is the Harpaz–Schlank theorem quoted above, which implies that if HSP holds for X then the Brauer–Manin obstruction applied to étale covers is the only obstruction for the Hasse principle. Since now there are many counterexamples to this claim (see [Poonen 2010; Harpaz and Skorobogatov 2014; Colliot-Thélène et al. 2013]) we get that there are two- and three-dimensional counterexamples to HSP. However we can offer some positive results; see Theorems 13.3, 13.7 and 14.8 in the next two sections.

13. Geometrically rational and birational surfaces

Lemma 13.1. *Let X be a smooth, projective, geometrically rational surface defined over a number field K. Then the group $Ш^2(K, \pi_2(\overline{X}))$ is finite.*

Proof. There is a finite Galois extension $L|K$ such that the action of $\mathrm{Gal}(\overline{L}|L)$ on $\mathrm{Pic}(\overline{X})$ is trivial. Hence $Ш^1(L, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z}) = 0$ by the Chebotarev density theorem (see the proof of Lemma 12.2), where we equip $\mathbb{Q}/\mathbb{Z}$ with the discrete topology and the trivial $\mathrm{Gal}(\overline{K}|K)$-action. As the image of $Ш^1(K, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z})$ lies in $Ш^1(L, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z})$ with respect to the restriction map

$$H^1(K, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z}) \to H^1(L, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z}),$$

the group $Ш^1(K, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z})$ lies in the image of the inflation map

$$H^1(\mathrm{Gal}(L|K), \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z}) \to H^1(K, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z}).$$

Since $H^1(\mathrm{Gal}(L|K), \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z})$ is finite, we get that $Ш^1(K, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z})$ is finite too. We have $\pi_1(\overline{X}) = \pi_1(\mathbb{P}^2_{\overline{K}}) = \{1\}$ because the étale fundamental group is a birational invariant. Hence

$$\begin{aligned}\pi_2(\overline{X}) = H_2(\overline{X}, \widehat{\mathbb{Z}}) &= \mathrm{Hom}(H^2(\overline{X}, \mathbb{Q}/\mathbb{Z}), \mathbb{Q}/\mathbb{Z}) \\ &= \mathrm{Hom}(\mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z}, \mathbb{Q}/\mathbb{Z} \otimes_{\widehat{\mathbb{Z}}} \widehat{\mathbb{Z}}(1))\end{aligned}$$

by [Artin and Mazur 1969, Corollary 6.2, p. 70] and the fact that for a geometrically rational surface X the étale Chern class map

$$c_1 : \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z} \to H^2(\overline{X}, \mathbb{Q}/\mathbb{Z} \otimes_{\widehat{\mathbb{Z}}} \widehat{\mathbb{Z}}(1))$$

is an isomorphism. Hence by [Milne 1986, Theorem 4.10(a), p. 70] there is a perfect duality between $Ш^2(K, \pi_2(\overline{X}))$ and $Ш^1(K, \mathrm{Pic}(\overline{X}) \otimes \mathbb{Q}/\mathbb{Z})$. The claim is now clear. □

For every $s \in \mathrm{Sel}(X/K)$, let $\mathrm{Sel}_0(s) \subseteq \mathrm{Sel}(X/K)$ denote the preimage of $\prod_{v \in |K|} r_{v*}(s)$ with respect to the map $\prod_{v \in |K|} r_{v*}$.

Lemma 13.2. *Let X be a smooth, projective, geometrically rational surface defined over a number field K such that $X(K) \neq \varnothing$, and let $s \in \mathrm{Sel}(X/K)$. Then the cardinality of $\mathrm{Sel}_0(s)$ is at most the order of $Ш^2(K, \pi_2(\overline{X}))$.*

Proof. By Proposition 12.3 it will be enough to show that the number of equivalence classes of the relation $\sim_2$ in $\mathrm{Sel}_0(s)$ is at most the order of $Ш^2(K, \pi_2(\overline{X}))$. For every $x, y \in \mathrm{Sel}_0(s)$ at least we have $x \sim_1 y$ by Theorem 7.6, since $\pi_1(\overline{X}) = \{1\}$, and so $|\mathrm{Sec}(X/K)| = 1$. Moreover, $\delta_1^X(x, y) \in Ш^2(K, \pi_2(\overline{X}))$ by the naturality of obstruction classes. But

$$\delta_1^X(x, y) = \delta_1^X(x, s) - \delta_1^X(y, s),$$

so the claim follows from the pigeonhole principle. □

Theorem 13.3. *Let X be a smooth, projective, geometrically rational surface defined over a number field K such that $X(K) \neq \varnothing$. Then $\mathrm{Sel}(X/K)$ is finite.*

Because we expect that $X(K)/R$ is finite for such an X and K, this result should be also expected, assuming that HSP holds for X.

Proof. Because for every archimedean $v \in |K|$ the topological space $X_v(K_v)$ has only finitely many connected components, the set $X_v(K_v)/H$ is finite for such v by Theorem 1.2. Since $\pi_1(\overline{X}) = \{1\}$ we get that Brauer equivalence and H-equivalence coincide on $X_v(K_v)$ for every nonarchimedean $v \in |K|$ by Theorem 1.1. Hence by [Bloch 1981, Corollary A.2, p. 55] the set $X_v(K_v)/H$ is finite for every $v \in |K|$.

Moreover, $X_v(K_v)/H$ has at most one element when v is nonarchimedean and X has good reduction at v by [Bloch 1981, Corollary A.3, p. 55]. Hence we may conclude that the set $\prod_{v\in|K|} X_v(K_v)/H$ is finite. So we only need to show that the map

$$\prod_{v\in|K|} r_{v*} : \mathrm{Sel}(X/K) \to \prod_{v\in|K|} X_v(K_v)/H$$

is finite-to-one. This follows from Lemmas 13.1 and 13.2. □

Lemma 13.4. *Let X and Y be smooth projective surfaces over a field K, and let $\pi : X \to Y$ be a composition of monoidal transformations over K. Then the map $\pi_* : X(K)/R \to Y(K)/R$ induced by π is a bijection.*

Proof. We may immediately reduce to the case when π is the blow-up of an irreducible subvariety $S \subset Y$ of dimension zero by induction on the number of blow-ups in some sequence of contractions $X \to X_1 \to \cdots \to X_n = Y$ whose composition is π. If S has no K-valued points then $\pi^{-1}(S)$ has no K-valued points either and the map π_* is obviously a bijection. Otherwise S consists of one K-valued point. In this case $\pi^{-1}(S)$ is isomorphic to $\mathbb{P}^1_K$ and the claim is clear. □

Lemma 13.5. *Let X and Y be smooth projective surfaces over the field $\mathbb{R}$, and let $\pi : X \to Y$ be a composition of monoidal transformations over $\mathbb{R}$. Then the map $\pi_* : \pi_0(X(\mathbb{R})) \to \pi_0(Y(\mathbb{R}))$ induced by π is a bijection.*

Proof. The argument is the same as above. □

Proposition 13.6. *Let X and Y be smooth geometrically irreducible projective surfaces over a number field K and let $\pi : X \to Y$ be a composition of monoidal transformations over K. Assume that Y is simply connected and $Y(K) \neq \varnothing$. Then the map $\pi_* : \mathrm{Sel}(X/K) \to \mathrm{Sel}(Y/K)$ induced by π is injective.*

Proof. Note that $X(K) \neq \varnothing$, since X is birational to Y. Again we may assume without loss of generality that π is the blow-up of an irreducible subvariety of dimension zero. Let $x, y \in \mathrm{Sel}(X/K)$ be such that $\pi_*(x) = \pi_*(y)$. Because the map induced by π between the fundamental groups is an isomorphism, we get that $x \sim_1 y$. Also note that by Theorem 7.6 it will be enough to show that $\delta_1^X(x, y) = 0$, since by Lemma 13.5 and Theorem 1.2 we have $r_{v*}(x) = r_{v*}(y)$ for every real place v of K, so the higher obstruction classes $\delta_n^X(x, y)$ will vanish for every $n \geq 2$ by Proposition 12.3.

By the Hurewicz theorem

$$\pi_2(\overline{X}) \cong H_2(\overline{X}, \hat{\mathbb{Z}}) \cong H_2(\overline{Y}, \hat{\mathbb{Z}}) \oplus \mathrm{Ker}(H_2(\pi)) \cong \pi_2(\overline{Y}) \oplus \mathrm{Ker}(H_2(\pi)),$$

where

$$H_2(\pi) : H_2(\overline{X}, \hat{\mathbb{Z}}) \to H_2(\overline{Y}, \hat{\mathbb{Z}})$$

is induced by π. Because π is the blow-up of a single closed point, the module $\mathrm{Ker}(H_2(\pi))$ is the induction of $\widehat{\mathbb{Z}}(1)$ from a finite extension of K. Therefore $Ш^2(K, \mathrm{Ker}(H_2(\pi))$ vanishes by Lemma 12.2 and Shapiro's lemma. Therefore $\delta_1^X(x,y)$ is zero by the naturality of obstruction classes. □

Theorem 13.7. *Let $\pi : X \to Y$ be a composition of monoidal transformations between geometrically irreducible smooth projective surfaces over K. Assume that* $\mathrm{Sel}(Y/K)$ *is finite, Y is simply connected, the set $Y(K)$ is nonempty, and HSP holds for Y over K. Then* $\mathrm{Sel}(X/K)$ *is finite and HSP holds for X over K, too.*

This result can be used to supply many examples of surfaces satisfying HSP, for example blow-ups of Châtelet surfaces; see Theorem 14.8 below.

Proof. By Proposition 13.6 the set $\mathrm{Sel}(X/K)$ injects into $\mathrm{Sel}(Y/K)$, so it is finite. Let $x, y \in X(K)$ be H-equivalent. Then $\pi(x)$ and $\pi(y)$ are H-equivalent elements of $Y(K)$, so they are R-equivalent, since HSP holds for Y. Therefore x and y are also R-equivalent by Lemma 13.4. We get that the map

$$\iota_{X/K} : X(K)/R \to \mathrm{Sel}(X/K) \tag{13.7.1}$$

is injective. Let s be an element of $\mathrm{Sel}(X/K)$. Because $\mathrm{Sel}(Y/K)$ is finite, the topology on it is discrete. Therefore there is a $y \in Y(K)$ such that $\iota_{Y/K}(y) = \pi_*(s)$. Let $x \in X(K)$ be such that $\pi(x) = y$. Then $\pi_*(\iota_{X/K}(x)) = \iota_{Y/K}(y) = \pi_*(s)$, so by Proposition 13.6 we get that $\iota_{X/K}(x) = s$, so the map in (13.7.1) is also surjective. □

14. Generalised Châtelet surfaces

Notation 14.1. For every torus S defined over a field K of characteristic zero let $C(S)$ denote the $\mathrm{Gal}(\overline{K}|K)$-module of its cocharacters, and for every scheme X over K let

$$\delta : H^1(X, S) \to H^2(X, C(S) \otimes \widehat{\mathbb{Z}}(1))$$

be the projective limit of the coboundary maps

$$\delta_n : H^1(X, S) \to H^2(X, S[n])$$

furnished by the corresponding Kummer exact sequences, where $S[n]$ denotes the n-torsion subgroup scheme of S. Note that, when K is a number field, δ maps $Ш^1(K, S)$ into $Ш^2(K, C(S) \otimes \widehat{\mathbb{Z}}(1))$. Let

$$\delta_0 : Ш^1(K, S) \to Ш^2(K, C(S) \otimes \widehat{\mathbb{Z}}(1))$$

be the restriction of δ onto $Ш^1(K, S)$.

Lemma 14.2. *Let S be a torus defined over a field K of characteristic zero.*

(i) *The map $\delta : H^1(K, S) \to H^2(K, C(S) \otimes \widehat{\mathbb{Z}}(1))$ is injective.*

(ii) *The map $\delta_0 : Ш^1(K, S) \to Ш^2(K, C(S) \otimes \widehat{\mathbb{Z}}(1))$ is bijective when K is a number field.*

Proof. First note that for every field K of characteristic zero and for every torus S defined over K the cohomology group $H^1(K, S)$ has finite exponent. In fact there is a finite Galois extension $L|K$ such that the action of $\mathrm{Gal}(\overline{L}|L)$ on $C(S)$ is trivial. Hence $H^1(L, S) = 0$ by Hilbert's theorem 90. Therefore $H^1(K, S)$ is the image of $H^1(\mathrm{Gal}(L|K), S)$ under the inflation map. But the group $H^1(\mathrm{Gal}(L|K), S)$ is annihilated by the order of $\mathrm{Gal}(L|K)$. The first claim also follows.

Now we prove that δ_0 is also surjective when K is a number field. Let $L|K$ be a finite Galois extension of the type above and assume that the degree of this extension is m. By the Albert–Brauer–Hasse–Noether theorem the group $Ш^2(L, S)$ is trivial, and hence $Ш^2(K, S)$ is annihilated by multiplication by m, since it is a subquotient of $H^2(\mathrm{Gal}(L|K), S)$. The cokernel of the restriction

$$\delta_n|_{Ш^1(X,S)} : Ш^1(X, S) \to Ш^2(X, S[n]) \tag{14.2.1}$$

of δ_n onto $Ш^1(X, S)$ is a subgroup of $Ш^2(X, S)$. Therefore the map in (14.2.1) surjects onto $mШ^2(X, S[n])$ by the above. Claim (ii) now follows by taking the limit. □

Definition 14.3. We say that a smooth geometrically irreducible projective surface X defined over a field K of characteristic zero is a generalised Châtelet surface if there is an $a \in K^*$ and a separable polynomial $P \in K[x]$ of degree 4 such that X is a smooth compactification of the affine surface given by the equation

$$y^2 - az^2 = P(x)$$

over K. For the sake of brevity we will frequently drop the adjective 'generalised' when we talk about generalised Châtelet surfaces.

Proposition 14.4. *Let X be a Châtelet surface defined over a number field K. Then for every $s \in \mathrm{Sel}(X/K)$ there is an $x \in X(K)$ such that $\iota_{X_v/K_v}(i_v(x)) = r_{v*}(s)$ for every $v \in |K|$.*

Proof. Let S be a finite subset of $|K|$ which contains every archimedean place of K and every nonarchimedean place of K where X does not have good reduction. For every $v \in |K|$ choose an $x_v \in X_v(K_v)$ such that $\iota_{X_v/K_v}(x_v) = r_{v*}(s)$. By [Colliot-Thélène et al. 1987b, Theorem 8.6(b), p. 87] for every $v \in |K|$ there is an open neighbourhood $U_v \subseteq X_v(K_v)$ of x_v contained by the R-equivalence class of x_v in $X_v(K_v)$. Because $\prod_{v\in|K|} x_v \in X(\mathbb{A}_K)^{\mathrm{Br}}$ by the easy direction of the Harpaz–Schlank theorem, there is an $x \in X(K)$ such that $i_v(x) \in U_v$ for every

$v \in S$ by [ibid., Theorem 8.11(c), p. 92]. Clearly $\iota_{X_v/K_v}(i_v(x)) = r_{v*}(s)$ for every $v \in S$. Because for every $v \in |K| - S$ the set $X(K)/R$ consists of one element by [ibid., Theorem 8.6(c), p. 87], we get that $\iota_{X_v/K_v}(i_v(x)) = r_{v*}(s)$ for every $v \in |K| - S$ as well. □

Definition 14.5. For every smooth projective geometrically irreducible variety X defined over a field K let $\mathrm{CH}_0(X)$ denote the Chow group of zero-dimensional cycles on X, and let $A_0(X)$ denote the kernel of the degree map $\deg : \mathrm{CH}_0(X) \to \mathbb{Z}$. Fix a point $x \in X(K)$. Then there is a map $\Psi_x : X(K)/R \to A_0(X)$ which for every $y \in X(K)$ maps the R-equivalence class of y to $[y] - [x]$. When K is a number field let $\text{Ш}^1 A_0(X)$ denote the subgroup of those elements c of $A_0(X)$ such that the base change of c to K_v is the zero element of $A_0(X_v)$ for every $v \in |K|$. Now let X be a Châtelet surface, let S be a torus over K whose group of characters is isomorphic to $\mathrm{Pic}(\overline{X})$ as a $\mathrm{Gal}(\overline{K}|K)$-module and let $\mathcal{T}$ be a universal torsor over X. Then we have a map

$$\rho_{\mathcal{T}} : X(K) \to H^1(K, S)$$

which associates to every $P \in X(K)$ the class of the fibre of $\mathcal{T}$ at P. Moreover, there is a unique homomorphism $\Phi_{\mathcal{T}} : A_0(X) \to H^1(K, S)$ such that for every $y \in X(K)$ we have $\Phi_{\mathcal{T}}([y]-[x]) = \rho_{\mathcal{T}}(y) - \rho_{\mathcal{T}}(x)$ (see [ibid., p. 88]). In particular the map $\rho_{\mathcal{T}}$ factors through R-equivalence; let $\rho_{\mathcal{T},R} : X(K)/R \to H^1(K, S)$ be the map which sends the R-equivalence class of every $y \in X(K)$ to $\rho_{\mathcal{T}}(y)$. When K is a number field, $\Phi_{\mathcal{T}}$ maps $\text{Ш}^1 A_0(X)$ into $\text{Ш}^1(K, S)$. Let $\Phi_0 : \text{Ш}^1 A_0(X) \to \text{Ш}^1(K, S)$ be the restriction of $\Phi_{\mathcal{T}}$ onto $\text{Ш}^1 A_0(X)$.

Theorem 14.6 (Colliot-Thélène, Sansuc, Swinnerton-Dyer). *Let X be a Châtelet surface, let $\mathcal{T}$ be a universal torsor over X and let $x \in X(K)$.*

(i) *The map $\rho_{\mathcal{T},R}$ is an injection.*

(ii) *The map Ψ_x is a bijection.*

(iii) *When K is a number field the map Φ_0 is a bijection.*

Proof. The map $\rho_{\mathcal{T},R}$ is injective by [Colliot-Thélène et al. 1987b, Theorem 8.5(a), p. 86]. Claim (ii) is true by [ibid., Theorem 8.8, p. 89] while claim (iii) holds by [ibid., Theorem 8.10, p. 91]. □

Let X be a Châtelet surface over a number field K. Then for every $x \in X(K)$ let $R_0(x) \subseteq X(K)/R$ denote the set of those R-equivalences classes s such that for every $y \in s$ the points $x, y \in X_v(K_v)$ are R-equivalent for every $v \in |K|$.

Corollary 14.7. *For every X, K and x as above, the set $R_0(x)$ has the same cardinality as $\text{Ш}^1(K, S)$.*

Proof. Note that Ψ_x maps $R_0(x)$ into $Ш^1 A_0(X)$. Therefore it will be enough to show that the induced map

$$\Psi_x|_{R_0(x)} : R_0(x) \to Ш^1 A_0(X) \tag{14.7.1}$$

is a bijection by part (iii) of Theorem 14.6. It is injective by part (ii) of Theorem 14.6. Let $s \in Ш^1 A_0(X)$ be arbitrary; by part (ii) of Theorem 14.6 there is a $y \in X(K)$ such that $[y] - [x]$ is s. Because Ψ_x is a bijection over K_v for every $v \in |K|$ by part (ii) of Theorem 14.6, we get that the points $x, y \in X_v(K_v)$ are R-equivalent for every $v \in |K|$. Therefore the map in (14.7.1) is surjective, too. □

Theorem 14.8. *Let K be a number field and let X be a generalised Châtelet surface over K. Then HSP holds for X.*

Remark 14.9. According to [Colliot-Thélène et al. 1987a, Remark 8.10.2, p. 91] there is a Châtelet surface X defined over a number field K on which R-equivalence is strictly finer than Brauer equivalence. Hence by Theorem 14.8 we get that H-equivalence is finer than étale-Brauer equivalence over number fields. Moreover, there are two rational points $x, y \in X(K)$ such that $x, y \in X_v(K_v)$ are H-equivalent for every $v \in |K|$, but x and y are not H-equivalent over K. The theorem above is also interesting because it covers a whole class of varieties X for which HSP holds, but which are not homogeneous spaces; moreover, every R-equivalence class of $X(K)$ is Zariski-dense (see [Colliot-Thélène et al. 1987b, Theorem 8.5(b), p. 86]) and, by [Colliot-Thélène et al. 1987b, Theorem 8.13, p. 95], the set $X(K)/R$ could be arbitrarily large for X defined over $\mathbb{Q}$.

Proof of Theorem 14.8. Because Châtelet surfaces are geometrically rational it will be both necessary and sufficient to show that the map

$$\iota_{X/K,R} : X(K)/R \to \mathrm{Sel}(X/K)$$

is a bijection by Theorem 13.3. Let $c \in H^1(X, S)$ be the cohomology class corresponding to the universal torsor $\pi : \mathcal{T} \to X$. Let $x, y \in X(K)$ be in two different R-equivalence classes. By part (i) of Theorem 14.6 the pullback classes $x^*(c), y^*(c) \in H^1(K, S)$ are different. Therefore the classes

$$\delta(x^*(c)) = x^*(\delta(c)),\ \delta(y^*(c)) = y^*(\delta(c)) \in H^2(K, C(S) \otimes \widehat{\mathbb{Z}}(1))$$

are also different by part (i) of Lemma 14.2. Therefore x and y are not H-equivalent by Lemma 8.5. The injectivity of $\iota_{X/K,R}$ follows.

Now we prove that it is surjective, too. Let $s \in \mathrm{Sel}(X/K)$. By Proposition 14.4 there is an $x \in X(K)$ such that $\iota_{X_v/K_v}(i_v(x)) = r_{v*}(s)$ for every $v \in |K|$. Note that $\iota_{X/K,R}$ maps $R_0(x)$ into $\mathrm{Sel}_0(s)$. Therefore it will be enough to show that the induced map

$$\iota_{X/K,R}|_{R_0(x)} : R_0(x) \to \mathrm{Sel}_0(s) \tag{14.9.1}$$

is a bijection. By the above this map is injective, so it will be enough to show that the cardinality of $\mathrm{Sel}_0(s)$ is at most $Ш^2(K, C(S) \otimes \widehat{\mathbb{Z}}(1))$ by part (ii) of Lemma 14.2 and Corollary 14.7. By the definition of S we have that

$$C(S) \otimes \widehat{\mathbb{Z}}(1) = \mathrm{Hom}(\mathrm{Pic}(\overline{X}), \mathbb{Z}) \otimes \widehat{\mathbb{Z}}(1) = \mathrm{Hom}(\mathrm{Pic}(\overline{X}) \otimes \widehat{\mathbb{Z}}, \widehat{\mathbb{Z}}(1)).$$

We already noted in the proof of Lemma 13.1 that

$$\pi_2(\overline{X}) = \mathrm{Hom}(\mathrm{Pic}(\overline{X}) \otimes \widehat{\mathbb{Z}}, \widehat{\mathbb{Z}}(1)), \quad \text{so } \pi_2(\overline{X}) = C(S) \otimes \widehat{\mathbb{Z}}(1).$$

Therefore it will be enough to show that the cardinality of $\mathrm{Sel}_0(s)$ is at most $Ш^2(K, \pi_2(\overline{X}))$. But this is the content of Lemma 13.2. □

Acknowledgements

This article was inspired by the work of Bertrand Toën [2004], who, according to my best knowledge, was first to suggest the study of rational points via maps like $\iota_{X/K}$. I also wish to thank Jean-Louis Colliot-Thélène, Yonathan Harpaz, Tomer Schlank, Alexei Skorobogatov and Akio Tamagawa either for useful discussions or for useful correspondence concerning the contents of this paper, and the referee for their many comments and corrections. The author was partially supported by the EPSRC grants P19164 and P36794.

References

[Artin and Mazur 1969] M. Artin and B. Mazur, *Étale homotopy*, Lecture Notes in Mathematics **100**, Springer, Berlin-New York, 1969. MR 39 #6883 Zbl 0182.26001

[Bloch 1981] S. Bloch, "On the Chow groups of certain rational surfaces", *Ann. Sci. École Norm. Sup.* (4) **14**:1 (1981), 41–59. MR 83e:14006 Zbl 0524.14006

[Colliot-Thélène et al. 1987a] J.-L. Colliot-Thélène, J.-J. Sansuc, and P. Swinnerton-Dyer, "Intersections of two quadrics and Châtelet surfaces, I", *J. Reine Angew. Math.* **373** (1987), 37–107. MR 88m:11045a Zbl 0622.14029

[Colliot-Thélène et al. 1987b] J.-L. Colliot-Thélène, J.-J. Sansuc, and P. Swinnerton-Dyer, "Intersections of two quadrics and Châtelet surfaces, II", *J. Reine Angew. Math.* **374** (1987), 72–168. MR 88m:11045b Zbl 0622.14030

[Colliot-Thélène et al. 2013] J.-L. Colliot-Thélène, A. Pál, and A. Skorobogatov, "Pathologies of the Brauer–Manin obstruction", preprint, 2013. To appear in *Math. Z.* arXiv 1310.5055

[Esnault et al. 1993] H. Esnault, B. Kahn, and E. Viehweg, "Coverings with odd ramification and Stiefel–Whitney classes", *J. Reine Angew. Math.* **441** (1993), 145–188. MR 94m:14017 Zbl 0772.57028

[Friedlander 1982] E. M. Friedlander, *Étale homotopy of simplicial schemes*, Annals of Mathematics Studies **104**, Princeton University Press and University of Tokyo Press, 1982. MR 84h:55012 Zbl 0538.55001

[Goerss 1995] P. G. Goerss, "Homotopy fixed points for Galois groups", pp. 187–224 in *The Čech centennial* (Boston, 1993), edited by M. Cenkl and H. Miller, Contemp. Math. **181**, Amer. Math. Soc., Providence, RI, 1995. MR 96a:55008 Zbl 0830.55002

[Grothendieck 1971] A. Grothendieck, *Revêtements étales et groupe fondamental* = SGA1 (Séminaire de Géométrie Algébrique du Bois Marie 1960–1961), Lecture Notes in Math. **224**, Springer, Berlin, 1971. MR 50 #7129 Zbl 0234.14002

[Harari 2000] D. Harari, "Weak approximation and non-abelian fundamental groups", *Ann. Sci. École Norm. Sup.* (4) **33**:4 (2000), 467–484. MR 2002e:14034 Zbl 1073.14522

[Harpaz and Schlank 2013] Y. Harpaz and T. M. Schlank, "Homotopy obstructions to rational points", pp. 280–413 in *Torsors, étale homotopy and applications to rational points*, edited by A. N. Skorobogatov, London Math. Soc. Lecture Note Ser. **405**, Cambridge Univ. Press, 2013. MR 3077173 Zbl 1298.14025

[Harpaz and Skorobogatov 2014] Y. Harpaz and A. N. Skorobogatov, "Singular curves and the étale Brauer–Manin obstruction for surfaces", *Ann. Sci. Éc. Norm. Supér.* (4) **47**:4 (2014), 765–778. MR 3250063 Zbl 06380119

[Houdebine and Mahé 1982] J. Houdebine and L. Mahé, "Séparation des composantes connexes réelles dans le cas des variétés projectives", pp. 358–370 in *Real algebraic geometry and quadratic forms* (Rennes, 1981), edited by J.-L. Colliot-Thélène et al., Lecture Notes in Math. **959**, Springer, Berlin-New York, 1982. MR 84m:14030 Zbl 0521.14008

[de Jong 2013] J. de Jong, "A result of Gabber", preprint, 2013, http://www.math.columbia.edu/~dejong/papers/2-gabber.pdf.

[Mahé 1982] L. Mahé, "Signatures et composantes connexes", *Math. Ann.* **260**:2 (1982), 191–210. MR 84a:14011 Zbl 0507.14019

[Milne 1986] J. S. Milne, *Arithmetic duality theorems*, Perspectives in Mathematics **1**, Academic Press, Boston, 1986. MR 88e:14028 Zbl 0613.14019

[Milnor 1970] J. Milnor, "Algebraic K-theory and quadratic forms", *Invent. Math.* **9** (1970), 318–344. MR 41 #5465 Zbl 0199.55501

[Mochizuki 2003] S. Mochizuki, "Topics surrounding the anabelian geometry of hyperbolic curves", pp. 119–165 in *Galois groups and fundamental groups*, edited by L. Schneps, Math. Sci. Res. Inst. Publ. **41**, Cambridge Univ. Press, 2003. MR 2004m:14052 Zbl 1053.14029

[Poonen 2010] B. Poonen, "Insufficiency of the Brauer–Manin obstruction applied to étale covers", *Ann. of Math.* (2) **171**:3 (2010), 2157–2169. MR 2011g:14061 Zbl 1284.11096

[Quick 2011] G. Quick, "Continuous group actions on profinite spaces", *J. Pure Appl. Algebra* **215**:5 (2011), 1024–1039. MR 2012k:55018 Zbl 1227.55013

[Serre 1997] J.-P. Serre, *Galois cohomology*, 5th ed., Springer, Berlin, 1997. MR 98g:12007 Zbl 0902.12004

[Skorobogatov 1999] A. N. Skorobogatov, "Beyond the Manin obstruction", *Invent. Math.* **135**:2 (1999), 399–424. MR 2000c:14022 Zbl 0951.14013

[Stix 2011] J. Stix, "The Brauer–Manin obstruction for sections of the fundamental group", *J. Pure Appl. Algebra* **215**:6 (2011), 1371–1397. MR 2012c:14048 Zbl 1227.14032

[Tamagawa 1997] A. Tamagawa, "The Grothendieck conjecture for affine curves", *Compositio Math.* **109**:2 (1997), 135–194. MR 99a:14035 Zbl 0899.14007

[Toën 2004] B. Toën, "Homotopy types of algebraic varieties", preprint, 2004.

Communicated by Barry Mazur
Received 2013-09-09 Revised 2015-02-11 Accepted 2015-03-11

a.pal@imperial.ac.uk *Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2AZ, United Kingdom*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.875

Fermat's last theorem over some small real quadratic fields

Nuno Freitas and Samir Siksek

Using modularity, level lowering, and explicit computations with Hilbert modular forms, Galois representations, and ray class groups, we show that for $3 \leq d \leq 23$, where $d \neq 5, 17$ and is squarefree, the Fermat equation $x^n + y^n = z^n$ has no nontrivial solutions over the quadratic field $\mathbb{Q}(\sqrt{d})$ for $n \geq 4$. Furthermore, we show that for $d = 17$, the same holds for prime exponents $n \equiv 3, 5 \pmod 8$.

1. Introduction

Interest in the Fermat equation

$$x^n + y^n = z^n \tag{1}$$

over various number fields goes back to the 19th and early 20th centuries, with the work of Maillet (1897) and Furtwängler (1910) [Dickson 1920, pages 758 and 768]. However, until the current work, the only number fields for which Fermat's last theorem has been proved are $\mathbb{Q}$ and $\mathbb{Q}(\sqrt{2})$. These proofs are respectively due to Wiles [1995] (as consequence of his ground-breaking proof of modularity of semistable elliptic curves over $\mathbb{Q}$) and to Jarvis and Meekin [2004]. The precise statements are that if $K = \mathbb{Q}$ and $n \geq 3$ or $K = \mathbb{Q}(\sqrt{2})$ and $n \geq 4$, then the only solutions to (1) in K are the trivial ones satisfying $xyz = 0$. In [Freitas and Siksek 2015], it is shown that for five-sixths of real quadratic fields K, there is a bound B_K such that for prime exponents $n \geq B_K$, the only solutions to the Fermat equation (1) over K are the trivial ones. This paper is concerned with proving Fermat's last theorem for several other real quadratic fields. Our main results are the following two theorems.

Theorem 1. *Let $3 \leq d \leq 23$ be squarefree with $d \neq 5, 17$. Then the Fermat equation (1) does not have any nontrivial solutions over $\mathbb{Q}(\sqrt{d})$ with exponent $n \geq 4$.*

Freitas is supported through a grant within the framework of the DFG Priority Programme 1489 *Algorithmic and Experimental Methods in Algebra, Geometry and Number Theory*. Siksek is supported by an EPSRC Leadership Fellowship EP/G007268/1, and EPSRC *LMF: L-Functions and Modular Forms* Programme Grant EP/K034383/1.
MSC2010: primary 11D41; secondary 11F80, 11F03.
Keywords: Fermat, modularity, Galois representation, level lowering.

Theorem 2. *The Fermat equation (1) has no nontrivial solutions over $\mathbb{Q}(\sqrt{17})$ for prime exponents $n \geq 5$ satisfying $n \equiv 3, 5 \pmod 8$.*

Remark. For $n = 3$, equation (1) defines an elliptic curve having rank 0 over $\mathbb{Q}$; it does, however, have positive rank over some of the quadratic fields in the statement of Theorem 1. We therefore impose $n \geq 4$.

It is sufficient to prove Theorem 1 for exponents $n = 4, 6, 9$, and for prime exponents $n = p \geq 5$. In fact, all solutions to the Fermat equation in quadratic fields for $n = 4, 6, 9$ have been determined by Aigner [1934; 1957]. These are all defined over imaginary quadratic fields except for the trivial solutions. We may therefore restrict our attention to prime exponents $n = p \geq 5$.

Let $d \geq 2$ be a squarefree positive integer, and let $K = \mathbb{Q}(\sqrt{d})$, and write $\mathcal{O}_K$ for its ring of integers. By the *Fermat equation with exponent p over K*, we mean the equation

$$a^p + b^p + c^p = 0, \quad a, b, c \in \mathcal{O}_K. \tag{2}$$

A solution (a, b, c) is called *trivial* if $abc = 0$, otherwise it is *nontrivial*. For $p = 5, 7, 11$, all solutions of degree at most $(p-1)/2$ have been determined by Gross and Rohrlich [1978, Theorem 5]. It turns out that the only nontrivial quadratic solutions are permutations of $(1, \omega, \omega^2)$, where ω is a primitive cube root of unity. For $p = 13$, the same was shown to be true by Tzermias [2004]. We shall therefore henceforth assume that $p \geq 17$.

A brief overview of the method and difficulties. As in the proof of Fermat's last theorem over $\mathbb{Q}$ and $\mathbb{Q}(\sqrt{2})$, let (a, b, c) be a nontrivial solution to the Fermat equation (2), and consider the Frey elliptic curve

$$E_{a,b,c} : Y^2 = X(X - a^p)(X + b^p). \tag{3}$$

Write $E = E_{a,b,c}$ and denote by $\bar{\rho}_{E,p}$ its mod p Galois representation. An essential fact to the proof of Fermat's last theorem over $\mathbb{Q}$ and $\mathbb{Q}(\sqrt{2})$ is the modularity of the Frey curve. Modularity of elliptic curves over all real quadratic fields is now known (see [Freitas et al. 2014]). In particular, our Frey curve $E_{a,b,c}$ is modular over K. The proof of Fermat's last theorem over $\mathbb{Q}$ and $\mathbb{Q}(\sqrt{2})$ makes essential use of the fact that it is always possible to scale and permute the hypothetical nontrivial solution so that not only are a, b, c algebraic integers, but they are also coprime, and they satisfy additional 2-adic restrictions; over $\mathbb{Q}$, these are $a \equiv -1 \pmod 4$ and $2 \mid b$. In both cases, a suitable choice of scaling produces a semistable Frey curve $E_{a,b,c}$. Applying suitable level-lowering results to the modular Galois representation $\bar{\rho}_{E,p}$ shows that it arises from an eigenform of level 2 for $\mathbb{Q}$, and a Hilbert eigenform of level $\sqrt{2}$ for $\mathbb{Q}(\sqrt{2})$. There are no eigenforms at these levels, giving a contradiction and completing the proof for both fields.

It should be possible to carry out the same level lowering strategy over any real quadratic field K. To build on this and prove Fermat's last theorem over K there are, however, three principal difficulties:

(a) Verifying the irreducibility of $\bar{\rho}_{E,p}$; this is required for applying level lowering theorems.

(b) Computing the newforms at the levels predicted by conductor computations and level lowering; in general, these levels depend on the class and unit groups of K and are not of small norm.

(c) Dealing with the newforms found at these levels; in general, these spaces will be nonzero.

In [Freitas and Siksek 2015], it is shown that these difficulties disappear for $p > C_K$, where C_K is some inexplicit constant, for five-sixths of real quadratic fields K. In this paper, we show how to deal with (a), (b), (c) for the fields in the statement of Theorem 1. For $K = \mathbb{Q}(\sqrt{d})$ with $d = 5, 17$, our method for (c) fails. However, for $d = 17$, we are still able to prove Fermat's last theorem for half of exponents p using an argument of Halberstadt and Kraus [2002], yielding Theorem 2.

For $K = \mathbb{Q}$, it follows from Mazur's celebrated theorem on isogenies [1978] that $\bar{\rho}_{E,p}$ is irreducible for the Frey curve E and $p \geq 5$. The analogue of Mazur's theorem is not known for any other number field. However, Kraus shows that for K, a real quadratic field of class number 1, and E, a semistable elliptic curve over K, if $\bar{\rho}_{E,p}$ is reducible, then $p \leq 13$ or ramifies in K or $p \mid \mathrm{Norm}_{K/\mathbb{Q}}(u^2 - 1)$, where u is a fundamental unit of K. It is possible to give corresponding bounds for p when the class group is nontrivial and E has some fixed set of additive primes (see for example [David 2012]), although these bounds are considerably worse. In Section 6, we overcome these difficulties for the fields in Theorem 1 through explicit class field theory computations. Without this, we would not have been able to deal with small exponents p.

Assuming modularity of the Frey curve, it should be possible to apply the same strategy to the Fermat equation over many totally real fields, although the computation of newforms for totally real fields with either large degree or large discriminant is likely to be impractical. In Section 9, we illustrate this by looking at $\mathbb{Q}(\sqrt{30})$ and $\mathbb{Q}(\sqrt{79})$. At the end of that section, we also include a comparison between the recipes of the current paper and those of [Freitas and Siksek 2015], illustrating the need for the improvements in the current work.

The computations in this paper were carried out on the computer algebra system Magma [Bosma et al. 1997]. In particular, we have used Magma's Hilbert modular forms package (for the theory see [Dembélé and Donnelly 2008] and [Dembélé and Voight 2013]) and class field theory package (due to C. Fieker).

Notational conventions. Throughout, p denotes an odd rational prime, and K a totally real number field, with ring of integers $\mathcal{O}_K$. We write S for the set of prime ideals in $\mathcal{O}_K$ dividing 2. For a nonzero ideal $\mathfrak{a}$ of $\mathcal{O}_K$, we denote by $[\mathfrak{a}]$ the class of $\mathfrak{a}$ in the class group $\mathrm{Cl}(K)$. For a nontrivial solution (a, b, c) to the Fermat equation (2), let

$$\mathcal{G}_{a,b,c} := a\mathcal{O}_K + b\mathcal{O}_K + c\mathcal{O}_K, \tag{4}$$

and let $[a, b, c]$ denote the class of $\mathcal{G}_{a,b,c}$ in $\mathrm{Cl}(K)$. We exploit the well-known fact (e.g., [Cassels and Fröhlich 1967, Theorem VIII.4]) that every ideal class contains infinitely many prime ideals. Let $r = \#(\mathrm{Cl}(K)/\mathrm{Cl}(K)^2)$. Let $\mathfrak{m}_1 = 1 \cdot \mathcal{O}_K$, and fix, once and for all, odd prime ideals $\mathfrak{m}_2, \dots, \mathfrak{m}_r$ such that $[\mathfrak{m}_1], \dots, [\mathfrak{m}_r]$ represent the cosets of $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$. Let

$$\mathcal{H} = \{\mathfrak{m}_1, \dots, \mathfrak{m}_r\}.$$

Let $G_K = \mathrm{Gal}(\overline{K}/K)$. For an elliptic curve E/K, we write

$$\bar{\rho}_{E,p} : G_K \to \mathrm{Aut}(E[p]) \cong \mathrm{GL}_2(\mathbb{F}_p) \tag{5}$$

for the representation of G_K on the p-torsion of E. For a Hilbert eigenform $\mathfrak{f}$ of parallel weight 2 over K, we let $\mathbb{Q}_\mathfrak{f}$ denote the field generated by its eigenvalues. In this situation, ϖ will denote a prime of $\mathbb{Q}_\mathfrak{f}$ above p; of course, if $\mathbb{Q}_\mathfrak{f} = \mathbb{Q}$, we write p instead of ϖ. All other primes we consider are primes of K. We reserve the symbol $\mathfrak{P}$ for primes belonging to S, and $\mathfrak{m}$ for primes belonging to $\mathcal{H}$. An arbitrary prime of K is denoted by $\mathfrak{q}$, and $G_\mathfrak{q}$ and $I_\mathfrak{q}$ are the decomposition and inertia subgroups of G_K at $\mathfrak{q}$.

2. Level lowering

We need a level lowering result that plays the role of the Ribet step [1990] in the proof of Fermat's last theorem. The following theorem is deduced in [Freitas and Siksek 2015] from the work of Fujiwara [2006], Jarvis [2004] and Rajaei [2001].

Theorem 3. *Let K be a real quadratic field, and E/K an elliptic curve of conductor $\mathcal{N}$. Let p be a rational prime. For a prime ideal $\mathfrak{q}$ of K denote by $\Delta_\mathfrak{q}$ the discriminant of a local minimal model for E at $\mathfrak{q}$. Let*

$$\mathcal{M}_p := \prod_{\substack{\mathfrak{q} \mid \mathcal{N}, \\ p \mid \upsilon_\mathfrak{q}(\Delta_\mathfrak{q})}} \mathfrak{q} \quad \text{and} \quad \mathcal{N}_p := \frac{\mathcal{N}}{\mathcal{M}_p}. \tag{6}$$

Suppose the following:

(i) *Either $p \geq 5$, or $K = \mathbb{Q}(\sqrt{5})$ and $p \geq 7$.*

(ii) *E is modular.*

(iii) $\bar{\rho}_{E,p}$ *is irreducible.*

(iv) E *is semistable at all* $\mathfrak{q} \mid p$.

(v) $p \mid \upsilon_{\mathfrak{q}}(\Delta_{\mathfrak{q}})$ *for all* $\mathfrak{q} \mid p$.

Then, there is a Hilbert eigenform $\mathfrak{f}$ *of parallel weight* 2 *that is new at level* $\mathcal{N}_p$ *and some prime* ϖ *of* $\mathbb{Q}_{\mathfrak{f}}$ *such that* $\varpi \mid p$ *and* $\bar{\rho}_{E,p} \sim \bar{\rho}_{\mathfrak{f},\varpi}$.

3. Scaling and the odd part of the level

Let (a, b, c) be a nontrivial solution to the Fermat equation (2). Let $\mathcal{G}_{a,b,c}$ be as given in (4), which we think of as the greatest common divisor of a, b, c. An odd prime not dividing $\mathcal{G}_{a,b,c}$ is a prime of good or multiplicative reduction for $E_{a,b,c}$ and does not appear in the final level $\mathcal{N}_p$, as we see in due course. An odd prime dividing $\mathcal{G}_{a,b,c}$ exactly once is an additive prime and does appear in $\mathcal{N}_p$. To control $\mathcal{N}_p$, we need to control $\mathcal{G}_{a,b,c}$.

Scaling. We refer to page 878 for the notation.

Lemma 3.1. *Let* (a, b, c) *be a nontrivial solution to* (2). *There is a nontrivial integral solution* (a', b', c') *to* (2) *and some* $\mathfrak{m} \in \mathcal{H}$ *such that the following hold*:

(i) *For some* $\xi \in K^*$, *we have* $a' = \xi a$, $b' = \xi b$, $c' = \xi c$.

(ii) $\mathcal{G}_{a',b',c'} = \mathfrak{m} \cdot \mathfrak{r}^2$, *where* $\mathfrak{r}$ *is an odd prime ideal with* $\mathfrak{r} \neq \mathfrak{m}$.

(iii) $[a', b', c'] = [a, b, c]$.

Proof. Recall that $\mathcal{H} = \{\mathfrak{m}_1, \dots, \mathfrak{m}_r\}$ and that $[\mathfrak{m}_1], \dots, [\mathfrak{m}_r]$ represent the cosets of $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$. Thus for some $\mathfrak{m} \in \mathcal{H}$, we have $[a, b, c] = [\mathfrak{m}] \cdot [\mathfrak{b}]^2$, where $\mathfrak{b}$ is a fractional ideal. Now every ideal class is represented by infinitely many prime ideals. Thus there is an odd prime ideal $\mathfrak{r} \neq \mathfrak{m}$ such that $[a, b, c] = [\mathfrak{m}] \cdot [\mathfrak{r}]^2$. It follows that $\mathfrak{m} \cdot \mathfrak{r}^2 = (\xi) \cdot \mathcal{G}_{a,b,c}$ for some $\xi \in K^*$. Let a', b', c' be as in (i). Note that

$$(a') = (\xi) \cdot (a) = \mathfrak{m} \cdot \mathfrak{r}^2 \cdot \mathcal{G}_{a,b,c}^{-1} \cdot (a),$$

which is an integral ideal since $\mathcal{G}_{a,b,c}$ (by its definition) divides a. Thus a' is in $\mathcal{O}_K$ and similarly so are b' and c'. For (ii) and (iii), note that

$$\mathcal{G}_{a',b',c'} = a'\mathcal{O}_K + b'\mathcal{O}_K + c'\mathcal{O}_K = (\xi) \cdot (a\mathcal{O}_K + b\mathcal{O}_K + c\mathcal{O}_K) = (\xi) \cdot \mathcal{G}_{a,b,c} = \mathfrak{m} \cdot \mathfrak{r}^2. \quad \square$$

Behaviour at odd primes. For $u, v, w \in \mathcal{O}_K$ such that $uvw \neq 0$ and $u + v + w = 0$, let

$$E : y^2 = x(x - u)(x + v). \tag{7}$$

The invariants c_4, c_6, Δ, j have their usual meanings and are given by

$$\begin{aligned} c_4 &= 16(u^2 - vw) = 16(v^2 - wu) = 16(w^2 - uv), \\ c_6 &= -32(u - v)(v - w)(w - u), \quad \Delta = 16u^2v^2w^2, \quad j = c_4^3/\Delta. \end{aligned} \tag{8}$$

The following elementary lemma is a straightforward consequence of the properties of elliptic curves over local fields (e.g., [Silverman 1986, §VII.1 and §VII.5]).

Lemma 3.2. *With the above notation, let $\mathfrak{q} \nmid 2$ be a prime and let*

$$s = \min\{\upsilon_{\mathfrak{q}}(u), \upsilon_{\mathfrak{q}}(v), \upsilon_{\mathfrak{q}}(w)\}.$$

Write $E_{\min}$ for a local minimal model at $\mathfrak{q}$.

(i) *$E_{\min}$ has good reduction at $\mathfrak{q}$ if and only if s is even and*

$$\upsilon_{\mathfrak{q}}(u) = \upsilon_{\mathfrak{q}}(v) = \upsilon_{\mathfrak{q}}(w). \tag{9}$$

(ii) *$E_{\min}$ has multiplicative reduction at $\mathfrak{q}$ if and only if s is even and (9) fails to hold. In this case, the minimal discriminant $\Delta_{\mathfrak{q}}$ at $\mathfrak{q}$ satisfies*

$$\upsilon_{\mathfrak{q}}(\Delta_{\mathfrak{q}}) = 2\upsilon_{\mathfrak{q}}(u) + 2\upsilon_{\mathfrak{q}}(v) + 2\upsilon_{\mathfrak{q}}(w) - 6s.$$

(iii) *$E_{\min}$ has additive potentially multiplicative reduction at $\mathfrak{q}$ if and only if s is odd and (9) fails to hold.*

(iv) *$E_{\min}$ has additive potentially good reduction at $\mathfrak{q}$ if and only if s is odd and (9) holds. Moreover, E acquires good reduction over a quadratic extension of $K_{\mathfrak{q}}$.*

The odd part of the level. Let (a, b, c) be a nontrivial solution to the Fermat equation (2) with odd prime exponent p. Write E for the Frey curve in (3). Let $\mathcal{N}$ be the conductor of E and let $\mathcal{N}_p$ be as defined in (6). Recall that S is the set of prime ideals $\mathfrak{P}$ dividing 2. We define the *even parts* of $\mathcal{N}$ and $\mathcal{N}_p$ by

$$\mathcal{N}^{\text{even}} = \prod_{\mathfrak{P}\in S} \mathfrak{P}^{\upsilon_{\mathfrak{P}}(\mathcal{N})} \quad \text{and} \quad \mathcal{N}_p^{\text{even}} = \prod_{\mathfrak{P}\in S} \mathfrak{P}^{\upsilon_{\mathfrak{P}}(\mathcal{N}_p)}.$$

We define the *odd parts* of $\mathcal{N}$ and $\mathcal{N}_p$ by

$$\mathcal{N}^{\text{odd}} = \frac{\mathcal{N}}{\mathcal{N}^{\text{even}}} \quad \text{and} \quad \mathcal{N}_p^{\text{odd}} = \frac{\mathcal{N}}{\mathcal{N}_p^{\text{even}}}.$$

Lemma 3.3. *Let (a, b, c) be a nontrivial solution to the Fermat equation (2) with odd prime exponent p satisfying $\mathcal{G}_{a,b,c} = \mathfrak{m}\cdot\mathfrak{r}^2$, where $\mathfrak{m} \in \mathcal{H}$, and $\mathfrak{r}$ is an odd prime ideal such that $\mathfrak{r} \neq \mathfrak{m}$. Write E for the Frey curve in (3). Then at all $\mathfrak{q} \notin S \cup \{\mathfrak{m}\}$, the local minimal model $E_{\mathfrak{q}}$ is semistable and satisfies $p \mid \upsilon_{\mathfrak{q}}(\Delta_{\mathfrak{q}})$. Moreover,*

$$\mathcal{N}^{\text{odd}} = \mathfrak{m}^2 \cdot \mathfrak{r}^{0 \text{ or } 1} \cdot \prod_{\substack{\mathfrak{q} \mid abc \\ \mathfrak{q}\notin S\cup\{\mathfrak{m},\mathfrak{r}\}}} \mathfrak{q} \quad \textit{and} \quad \mathcal{N}_p^{\textit{odd}} = \mathfrak{m}^2. \tag{10}$$

Proof. Clearly, if $\mathfrak{q} \nmid 2abc$ then E has good reduction at $\mathfrak{q}$; hence $\mathfrak{q} \nmid \mathcal{N}, \mathcal{N}_p$. Note also that

$$\min\{\upsilon_{\mathfrak{r}(a^p)}, \upsilon_{\mathfrak{r}(b^p)}, \upsilon_{\mathfrak{r}(c^p)}\} = 2p.$$

By Lemma 3.2, E has good or multiplicative reduction at $\mathfrak{r}$, and in either case, $p \mid \upsilon_{\mathfrak{r}}(\Delta_{\mathfrak{r}})$, proving also the correctness of the exponents of $\mathfrak{r}$ in $\mathcal{N}$ and $\mathcal{N}_p$.

Recall that $\mathfrak{m} \in \mathcal{H}$ satisfies $\mathfrak{m} = 1 \cdot \mathcal{O}_K$, or $\mathfrak{m}$ is an odd prime ideal. In the former case, there is nothing to prove, so suppose that $\mathfrak{m}$ is an odd prime ideal. As E has full 2-torsion over K, the wild part of the conductor of E/K at $\mathfrak{m}$ vanishes (see [Silverman 1994, page 380]). Moreover,

$$\min\{\upsilon_{\mathfrak{m}}(a^p), \upsilon_{\mathfrak{m}}(b^p), \upsilon_{\mathfrak{m}}(c^p)\} = p.$$

By Lemma 3.2, E/K has additive reduction at $\mathfrak{m}$. Thus the exponent of $\mathfrak{m}$ in both $\mathcal{N}$ and $\mathcal{N}_p$ is 2.

Suppose that $\mathfrak{q} \mid abc$ and $\mathfrak{q} \notin S \cup \{\mathfrak{m}, \mathfrak{r}\}$. Since $\mathcal{G}_{a,b,c} = \mathfrak{m} \cdot \mathfrak{r}^2$, the prime $\mathfrak{q}$ divides precisely one of a, b, c. From (8), $\mathfrak{q} \nmid c_4$ so the model (3) is minimal and has multiplicative reduction at $\mathfrak{q}$, and $p \mid \upsilon_{\mathfrak{q}}(\Delta)$. By (6), we see that $\mathfrak{q} \nmid \mathcal{N}_p$. $\square$

Corollary 3.4. *Let $2 \le d \le 23$ be squarefree and let $K = \mathbb{Q}(\sqrt{d})$. Let (a, b, c) be a nontrivial solution to the Fermat equation (2). We may scale (a, b, c) so that it remains integral, and*

$$\mathcal{G}_{a,b,c} = \mathfrak{m} \cdot \mathfrak{r}^2, \quad \mathcal{N}_p^{odd} = \mathfrak{m}^2,$$

where

(a) *if $d \neq 10, 15$ then $\mathfrak{m} = 1 \cdot \mathcal{O}_K$;*
(b) *if $d = 10$ then $\mathfrak{m} = 1 \cdot \mathcal{O}_K$ or $\mathfrak{m} = (3, 1 + \sqrt{10})$;*
(c) *if $d = 15$ then $\mathfrak{m} = 1 \cdot \mathcal{O}_K$ or $\mathfrak{m} = (3, \sqrt{15})$;*

and $\mathfrak{r}$ is an odd prime ideal such that $\mathfrak{r} \neq \mathfrak{m}$.

Proof. For $2 \le d \le 23$, where $d \neq 10, 15$ and is squarefree, we have $\mathrm{Cl}(K)$ is trivial and so $\mathcal{H} = \{1 \cdot \mathcal{O}_K\}$. For $d = 10$, we have $\mathrm{Cl}(K) = \{[1 \cdot \mathcal{O}_K], [(3, 1 + \sqrt{10})]\}$, and we choose $\mathcal{H} = \{1 \cdot \mathcal{O}_K,\ (3, 1 + \sqrt{10})\}$. For $d = 15$, we have $\mathrm{Cl}(K) = \{[1 \cdot \mathcal{O}_K],\ [(3, \sqrt{15})]\}$, and we choose $\mathcal{H} = \{1 \cdot \mathcal{O}_K,\ (3, \sqrt{15})\}$. The corollary follows immediately from Lemmas 3.1 and 3.3. $\square$

4. Scaling by units and the even part of the level

In the previous section, we scaled (a, b, c) so that $\mathcal{G}_{a,b,c} = \mathfrak{m} \cdot \mathfrak{r}^2$, where $\mathfrak{m} \in \mathcal{H}$, and we computed the odd parts of the conductor $\mathcal{N}$ and level $\mathcal{N}_p$. Let $\mathcal{O}_K^*$ be the unit group of K. In this section, we study the effect on $\mathcal{N}$ and $\mathcal{N}_p$ of scaling (a, b, c) by units. Note that scaling (a, b, c) by units does not affect $\mathcal{G}_{a,b,c}$; it is plain from the proofs in the previous section that this leaves the odd parts of $\mathcal{N}$ and $\mathcal{N}_p$ unchanged. Applying an even permutation to (a, b, c) results in an isomorphic Frey curve, whereas applying an odd permutation replaces the Frey curve with its twist by -1, and so has the same effect as scaling (a, b, c) by -1.

Lemma 4.1. *Suppose K is a quadratic field and 2 is inert in $\mathcal{O}_K$. Let $\mathfrak{P} = 2\mathcal{O}_K$, and suppose $\mathfrak{P} \nmid abc$. Then after suitably permuting (a, b, c), we have $\upsilon_{\mathfrak{P}}(\mathcal{N}) = 4$. Moreover, E has potentially good reduction at $\mathfrak{P}$.*

Proof. We can write $K = \mathbb{Q}(\sqrt{d})$, where $d \equiv -3 \pmod 8$. Thus $\mathcal{O}_{\mathfrak{P}} = \mathbb{Z}_2[\omega]$, where $\omega^2 + \omega + 1 = 0$. The residue field of $\mathfrak{P}$ is $\mathbb{F}_2[\tilde{\omega}] \cong \mathbb{F}_4$. Write $A = a^p$, $B = b^p$, and $C = c^p$. Now $\mathfrak{P} \nmid ABC$ and $A + B + C = 0$. Then A, B, C are congruent modulo $\mathfrak{P}$ to 1, ω, ω^2 in some order. By rearranging, we may suppose that $C \equiv \omega^2 \pmod{\mathfrak{P}}$, and we will decide later on which of A, B are congruent to 1 and ω modulo $\mathfrak{P}$. Let E denote the Frey curve in (3). It follows from (8) that $\upsilon_{\mathfrak{P}}(\Delta) = 4$ and $\upsilon_{\mathfrak{P}}(c_4) \geq 5$. In particular, $\upsilon_{\mathfrak{P}}(j) \geq 11$, and so E has potentially good reduction at $\mathfrak{P}$. Furthermore, the Frey curve is minimal at $\mathfrak{P}$ and has additive reduction. We will follow the steps of Tate's algorithm as in [Silverman 1994, page 366]. Let $\tilde{E}$ denote the reduction of E modulo $\mathfrak{P}$. It is easy to check that the point $(\tilde{C}, \tilde{1})$ is singular on $\tilde{E}$. Now we shift the model E, replacing X by $X + C$ and Y by $Y + 1$, which has the effect of sending the point $(\widetilde{C}, \tilde{1})$ on the special fibre to $(\tilde{0}, \tilde{0})$. Write $a_1, \ldots, a_6$ for the a-invariants of the resulting model. Then

$$a_6 = C^3 + (B - A)C^2 - ABC - 1.$$

By Step 3 of Tate's algorithm, we know that if $\mathfrak{P}^2 \nmid a_6$ then $\upsilon_{\mathfrak{P}}(\mathcal{N}) = \upsilon_{\mathfrak{P}}(\Delta) = 4$. Suppose $\mathfrak{P}^2 \mid a_6$. Now swapping A and B replaces a_6 by

$$a_6' = C^3 + (A - B)C^2 - ABC - 1.$$

Observe that $\upsilon_{\mathfrak{P}}(a_6' - a_6) = \upsilon_{\mathfrak{P}}(2(A - B)C^2) = 1$. Hence $\mathfrak{P}^2 \nmid a_6'$. Thus we may always permute A, B, C so that $\upsilon_{\mathfrak{P}}(\mathcal{N}) = 4$. $\square$

Remark. Under the hypotheses of Lemma 4.1, it follows from Ogg's formula [Silverman 1994, Section IV.11] that the possible exponents of $\mathfrak{P}$ in the conductor are 2, 3, 4. Lemma 4.1 shows that we can always permute the solution so that the exponent of $\mathfrak{P}$ in the conductor is 4, avoiding having to compute newforms at the smaller levels. Unfortunately, we have found that it is not possible by permuting the solution to force the exponent to be smaller in general.

Lemma 4.2. *Suppose K is a quadratic field and let $\mathfrak{P} \in S$. Suppose (a, b, c) is a nontrivial solution to the Fermat equation, with $\mathfrak{P} \nmid \mathcal{G}_{a,b,c}$. The Frey curve E has potentially multiplicative reduction at $\mathfrak{P}$ if and only if*

(a) *either $f(\mathfrak{P}/2) = 1$ (i.e., 2 splits or ramifies in K),*

(b) *or $f(\mathfrak{P}/2) = 2$ (i.e., 2 is inert in K) and $\mathfrak{P} \mid abc$.*

Moreover, if the reduction at $\mathfrak{P}$ is multiplicative then $p \nmid \upsilon_{\mathfrak{P}}(\Delta_{\mathfrak{P}})$.

Proof. Suppose (a) or (b) holds. We claim that $\mathfrak{P} \mid abc$. If (b) holds, this is true by hypothesis. If (a) holds, then the residue field at $\mathfrak{P}$ is $\mathbb{F}_2$. It follows from $a^p + b^p + c^p = 0$ that $\mathfrak{P}$ divides at least one of a, b, c, establishing our claim. Moreover, as $\mathfrak{P} \nmid \mathcal{G}_{a,b,c}$, we see that $\mathfrak{P}$ divides precisely one of a, b, c. Let $t = \upsilon_{\mathfrak{P}}(abc) \geq 1$. By (8),

$$\upsilon_{\mathfrak{P}}(c_4) = 4\upsilon_{\mathfrak{P}}(2), \quad \upsilon_{\mathfrak{P}}(\Delta) = 4\upsilon_{\mathfrak{P}}(2) + 2pt.$$

Thus,

$$\upsilon_{\mathfrak{P}}(j) = 8\upsilon_{\mathfrak{P}}(2) - 2pt < 0 \tag{11}$$

as $p \geq 17$. Thus we have potentially multiplicative reduction at $\mathfrak{P}$. The converse follows from Lemma 4.1.

To complete the proof suppose that the reduction is multiplicative, and let c_4', c_6' and $\Delta' = \Delta_{\mathfrak{P}}$ be the corresponding invariants of a local minimal model. Now $\mathfrak{P} \nmid c_4'$, but $j' = (c_4')^3/\Delta'$. From (11), $p \nmid \upsilon_{\mathfrak{P}}(j)$, and hence $p \nmid \upsilon_{\mathfrak{P}}(\Delta_{\mathfrak{P}})$. ☐

Lemma 4.3. *Let $K_{\mathfrak{P}}$ be a local field and E an elliptic curve over $K_{\mathfrak{P}}$ with potentially multiplicative reduction. Let c_4, c_6 be the usual c-invariants of E. Let $L = K_{\mathfrak{P}}(\sqrt{-c_6/c_4})$ and let $\Delta(L/K_{\mathfrak{P}})$ be the discriminant of this local extension. Then the conductor of $E/K_{\mathfrak{P}}$ is*

$$f(E/K_{\mathfrak{P}}) = \begin{cases} 1 & \textit{if } \upsilon_{\mathfrak{P}}(\Delta(L/K_{\mathfrak{P}})) = 0, \\ 2\upsilon_{\mathfrak{P}}(\Delta(L/K_{\mathfrak{P}})) & \textit{otherwise.} \end{cases}$$

Proof. Let E' be the quadratic twist of E by $-c_6/c_4$. Then E' is a Tate curve (see, for example, [Silverman 1994, Section V.5]). The lemma now follows from [Rohrlich 1994, Section 18]. ☐

Lemma 4.4. *Let (a, b, c) be a nontrivial solution to the Fermat equation such that $\mathcal{G}_{a,b,c}$ is odd. Suppose 2 is either split or ramified in K, or that 2 is inert and $2 \mid abc$. Let*

$$\mathfrak{b} = \prod_{\mathfrak{P} \in S} \mathfrak{P}^{2\upsilon_{\mathfrak{P}}(2)+1},$$

and write

$$\Phi : \mathcal{O}_K^* \to (\mathcal{O}_K/\mathfrak{b})^*/((\mathcal{O}_K/\mathfrak{b})^*)^2$$

for the natural map. Choose a set $\lambda_1, \ldots, \lambda_k \in \mathcal{O}_K \backslash \mathfrak{b}$ that represent the elements of the cokernel of Φ. For $1 \leq i \leq k$, and for $\mathfrak{P} \in S$, let $\Delta_{\mathfrak{P}}^{(i)}$ be the discriminant of the local extension $K_{\mathfrak{P}}(\sqrt{\lambda_i})/K_{\mathfrak{P}}$, and let

$$\epsilon_{\mathfrak{P}}^{(i)} = \begin{cases} 1 & \textit{if } \upsilon_{\mathfrak{P}}(\Delta_{\mathfrak{P}}^{(i)}) = 0, \\ 2\upsilon_{\mathfrak{P}}(\Delta_{\mathfrak{P}}^{(i)}) & \textit{otherwise.} \end{cases}$$

Then we may scale (a,b,c) by an element of $\mathcal{O}_K^*$ so that for some i and for every $\mathfrak{P} \in S$, we have $\upsilon_{\mathfrak{P}}(\mathcal{N}) = \epsilon_{\mathfrak{P}}^{(i)}$.

Proof. Write $\mathcal{O} = \mathcal{O}_K$. By Lemma 4.2, we have potentially multiplicative reduction at $\mathfrak{P}$ for all the primes $\mathfrak{P} \in S$. Write c_4, c_6 for the usual c-invariants of the Frey curve E. Since $\mathcal{G}_{a,b,c}$ is odd but $\mathfrak{P} \mid abc$ for all $\mathfrak{P} \in S$, we have from (8) and the relation $a^p + b^p + c^p = 0$ that $\upsilon_{\mathfrak{P}}(c_4) = 4\upsilon_{\mathfrak{P}}(2)$ and $\upsilon_{\mathfrak{P}}(c_6) = 6\upsilon_{\mathfrak{P}}(2)$. Write $\gamma = -c_6/4c_4$. Then $\gamma \in \mathcal{O}_{\mathfrak{P}}^*$ for all $\mathfrak{P} \in S$, and

$$K_{\mathfrak{P}}(\sqrt{\gamma}) = K_{\mathfrak{P}}(\sqrt{-c_6/c_4}).$$

Now the exponent of $\mathfrak{P}$ in the conductor $\mathcal{N}$ of the Frey curve can be expressed by Lemma 4.3 in terms of the discriminant of the extension $K_{\mathfrak{P}}(\sqrt{\gamma})/K_{\mathfrak{P}}$.

We shall make use of the isomorphism

$$(\mathcal{O}/\mathfrak{b})^*/((\mathcal{O}/\mathfrak{b})^*)^2 \cong \prod_{\mathfrak{P} \in S} \mathcal{O}_{\mathfrak{P}}^*/(\mathcal{O}_{\mathfrak{P}}^*)^2,$$

which follows from the Chinese remainder theorem, and Hensel's lemma. Observe that scaling (a,b,c) by a unit $\eta \in \mathcal{O}_K^*$ scales γ by η^p. Now, as p is odd, it follows from the definition of Φ and the above isomorphism that we can scale (a,b,c) by some $\eta \in \mathcal{O}_K^*$ so that there is some $1 \le i \le k$ with $\gamma/\lambda_{\mathfrak{P}}^i$ a square in $\mathcal{O}_{\mathfrak{P}}$ for each $\mathfrak{P} \in S$. Therefore,

$$K_{\mathfrak{P}}(\sqrt{\gamma}) = K_{\mathfrak{P}}(\sqrt{\lambda_{\mathfrak{P}}^i}),$$

and the lemma follows from Lemma 4.3. □

Remark. Let u be the fundamental unit of the real quadratic field K. Observe that if $\lambda \in \mathcal{O}_K \backslash \mathfrak{b}$ represents an element of the cokernel of Φ, then for every integer k, the same element of the cokernel is also represented by $\lambda' = \pm u^k \lambda$. The local extension $K_{\mathfrak{P}}(\sqrt{\lambda'})/K_{\mathfrak{P}}$ depends only on the choice of sign $\pm$ and the parity of k. To keep the even part of the level small, we replace each representative λ by whichever one of λ, $-\lambda$, $u\lambda$, $-u\lambda$ minimizes the norm of the even part of the level $\mathcal{N}_p$.

5. Possibilities for $\mathcal{N}_p$

Corollary 5.1. *Let $2 \le d \le 23$ be squarefree and let $K = \mathbb{Q}(\sqrt{d})$. Let (a,b,c) be a nontrivial solution to the Fermat equation (2) with odd prime exponent p. We may scale (a,b,c) so that it remains integral, $\mathcal{G}_{a,b,c}$ and $\mathcal{N}_p^{\mathrm{odd}}$ are as in Corollary 3.4 and $\mathcal{N}_p^{\mathrm{even}} = \mathcal{N}^{\mathrm{even}}$ is as given in Table 1.*

Proof. The proof is a straightforward application of Lemma 4.1 and Lemma 4.4. When $\mathfrak{P} \mid abc$ (which includes all cases where 2 splits or ramifies), the third column of Table 1 lists our choices $\lambda_1, \dots, \lambda_k$ of representatives for the cokernel of Φ as in Lemma 4.4. In these cases, the even part of the conductor $\mathcal{N}^{\mathrm{even}}$ is given by Lemma 4.4. □

d	S	λs	$\mathcal{N}^{\mathrm{even}} = \mathcal{N}_p^{\mathrm{even}}$
2	$\mathfrak{P} = (\sqrt{2})$	$1, \quad -1-2\sqrt{2}$	$\mathfrak{P}$
3	$\mathfrak{P} = (1+\sqrt{3})$	1 $-1+2\sqrt{3}$	$\mathfrak{P}$ $\mathfrak{P}^4$
5	$\mathfrak{P} = (2)$	$1, \quad -5+2\sqrt{5}$ $\mathfrak{P} \nmid abc$	$\mathfrak{P}$ $\mathfrak{P}^4$
6	$\mathfrak{P} = (-2+\sqrt{6})$	1 $1+\sqrt{6}$	$\mathfrak{P}$ $\mathfrak{P}^8$
7	$\mathfrak{P} = (3+\sqrt{7})$	$1, \quad 21-8\sqrt{7}$ $-1+2\sqrt{7}, \quad -5+2\sqrt{7}$	$\mathfrak{P}$ $\mathfrak{P}^4$
10	$\mathfrak{P} = (2, \sqrt{10})$	$1, \quad 7-2\sqrt{10}$	$\mathfrak{P}$
11	$\mathfrak{P} = (3+\sqrt{11})$	1 $-1+2\sqrt{11}$	$\mathfrak{P}$ $\mathfrak{P}^4$
13	$\mathfrak{P} = (2)$	$1, \quad -5+2\sqrt{13}$ $\mathfrak{P} \nmid abc$	$\mathfrak{P}$ $\mathfrak{P}^4$
14	$\mathfrak{P} = (4+\sqrt{14})$	$1, \quad -3$ $1+\sqrt{14}, \quad -3+\sqrt{14}$	$\mathfrak{P}$ $\mathfrak{P}^8$
15	$\mathfrak{P} = (2, 1+\sqrt{15})$	$1, \quad -15+4\sqrt{15}$ $-1+2\sqrt{15}, \quad 7-2\sqrt{15}$	$\mathfrak{P}$ $\mathfrak{P}^4$
17	$\mathfrak{P}_1 = (\frac{3+\sqrt{17}}{2}),\ \mathfrak{P}_2 = (\frac{3-\sqrt{17}}{2})$	$1,\ 17-4\sqrt{17},\ -9+2\sqrt{17},\ -5+2\sqrt{17}$	$\mathfrak{P}_1 \cdot \mathfrak{P}_2$
19	$\mathfrak{P} = (13+3\sqrt{19})$	1 $-1+2\sqrt{19}$	$\mathfrak{P}$ $\mathfrak{P}^4$
21	$\mathfrak{P} = (2)$	$1, \quad -5+2\sqrt{21}$ $(7-\sqrt{21})/2, \quad (3+3\sqrt{21})/2$ $\mathfrak{P} \nmid abc$	$\mathfrak{P}$ $\mathfrak{P}^4$ $\mathfrak{P}^4$
22	$\mathfrak{P} = (14+3\sqrt{22})$	1 $1+\sqrt{22}$	$\mathfrak{P}$ $\mathfrak{P}^8$
23	$\mathfrak{P} = (5+\sqrt{23})$	$1, \quad 115+24\sqrt{23}$ $-1+2\sqrt{23}, \quad -163-34\sqrt{23}$	$\mathfrak{P}$ $\mathfrak{P}^4$

Table 1. Quantities required for Corollary 5.1 and its proof.

6. Irreducibility

We begin with a proposition that gathers some well-known facts in the literature.

Proposition 6.1. *Let E be an elliptic curve over a number field K and let p be a rational prime.*

(i) *If $\mathfrak{q} \nmid p$ and is a prime of good or multiplicative reduction then*

$$\bar{\rho}_{E,p}|_{I_\mathfrak{q}} \sim \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix}.$$

(ii) *If $\mathfrak{p} \mid p$ and is a prime of good ordinary reduction, or of multiplicative reduction, then*

$$\bar{\rho}_{E,p}|_{I_\mathfrak{p}} \sim \begin{pmatrix} \chi & * \\ 0 & 1 \end{pmatrix},$$

where χ is the mod p cyclotomic character.

(iii) *If $\mathfrak{p} \mid p$ and is a prime of good supersingular reduction with ramification degree e, then either*

$$\bar{\rho}_{E,p}|_{I_\mathfrak{p}} \sim \begin{pmatrix} \psi_2^e & 0 \\ 0 & \psi_2^{pe} \end{pmatrix}, \tag{12}$$

where $\psi_2 : I_\mathfrak{p} \to \mathbb{F}_{p^2}^*$ is a level 2 fundamental character, or

$$\bar{\rho}_{E,p}|_{I_\mathfrak{p}} \sim \begin{pmatrix} \psi_1^f & 0 \\ 0 & \psi_1^{e-f} \end{pmatrix}, \tag{13}$$

where $\psi_1 : I_\mathfrak{p} \to \mathbb{F}_p^*$ is the level 1 fundamental character, and f is some integer satisfying $0 < f < e$.

Proof. See [Serre 1972, §1.11, 1.12] and the proof of [Kraus 1996, Lemma 1]. □

Corollary 6.2. *Let E be an elliptic curve over a quadratic field K. Let p be a rational prime. Suppose $\bar{\rho}_{E,p}$ is reducible and E has supersingular reduction at some $\mathfrak{p}$ dividing p. Then $(p) = \mathfrak{p}^2$, and*

$$\bar{\rho}_{E,p}|_{I_\mathfrak{p}} \sim \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_1 \end{pmatrix}.$$

Proof. Write e for the ramification degree of $\mathfrak{p}$. As K is quadratic, $e = 1$ or 2. We apply part (iii) of Proposition 6.1. Suppose first that $\bar{\rho}_{E,p}|_{I_\mathfrak{p}}$ is given by (12). Now the characters ψ_2 and ψ_2^2 are not $\mathbb{F}_p$-valued, contradicting the reducibility of $\bar{\rho}_{E,p}$.

It follows that $\bar{\rho}_{E,p}|_{I_\mathfrak{p}}$ is given by (13), where f is an integer satisfying the inequality $1 < f < e$. Thus $e = 2$ and $f = 1$, completing the proof. □

Lemma 6.3. *Let E be an elliptic curve over a number field K of conductor $\mathcal{N}$ and let $p \geq 5$ be a rational prime. Suppose $\bar{\rho}_{E,p}$ is reducible and write*

$$\bar{\rho}_{E,p} \sim \begin{pmatrix} \theta & * \\ 0 & \theta' \end{pmatrix}, \tag{14}$$

where $\theta, \theta' : G_K \to \mathbb{F}_p^*$ are characters. Write $\mathcal{N}_\theta$ and $\mathcal{N}_{\theta'}$ for the respective conductors of these characters. Let q be a prime of K with $\mathfrak{q} \nmid p$.

(a) *If E has good or multiplicative reduction at $\mathfrak{q}$ then $\upsilon_\mathfrak{q}(\mathcal{N}_\theta) = \upsilon_\mathfrak{q}(\mathcal{N}'_\theta) = 0$.*

(b) *If E has additive reduction at $\mathfrak{q}$ then $\upsilon_\mathfrak{q}(\mathcal{N})$ is even and*

$$\upsilon_\mathfrak{q}(\mathcal{N}_\theta) = \upsilon_\mathfrak{q}(\mathcal{N}_{\theta'}) = \tfrac{1}{2}\upsilon_\mathfrak{q}(\mathcal{N}).$$

Proof. If E has good or multiplicative reduction at $\mathfrak{q}$ then, by Proposition 6.1(i), the characters θ, θ' are unramified at $\mathfrak{q}$. This immediately implies (a).

Suppose now that E has additive reduction at $\mathfrak{q}$. Recall that $\theta = \chi/\theta'$, where $\chi : G_K \to \mathbb{F}_p^*$ is the mod p cyclotomic character. As χ is unramified away from p, and therefore unramified at $\mathfrak{q}$, we see that $\upsilon_\mathfrak{q}(\mathcal{N}_\theta) = \upsilon_\mathfrak{q}(\mathcal{N}_{\theta'})$.

Suppose that $\upsilon_\mathfrak{q}(\mathcal{N}_\theta) = \upsilon_\mathfrak{q}(\mathcal{N}_{\theta'}) = 0$; we will deduce a contradiction. Then $\theta|_{I_\mathfrak{q}} = \theta'|_{I_\mathfrak{q}} = 1$. It follows that $\bar{\rho}_{E,p}(I_\mathfrak{q})$ has order 1 or p. Suppose first that E has potentially good reduction at $\mathfrak{q}$. Then (see [Kraus 1990, Introduction]), the order of $\bar{\rho}_{E,p}(I_\mathfrak{q})$ divides 24, and moreover is equal to 1 if and only if E has good reduction at $\mathfrak{q}$. As $p \geq 5$, we have a contradiction. We may therefore suppose that E has potentially multiplicative reduction. It then follows from the theory of the Tate curve [Silverman 1994, Proposition V.6.1] that $\#\bar{\rho}_{E,p}(I_\mathfrak{q}) = 1$ or 2. Again as $p \geq 5$, we have that $\bar{\rho}_{E,p}(I_\mathfrak{q}) = 1$ and so E has multiplicative reduction at $\mathfrak{q}$. This contradicts the fact that E has additive reduction at $\mathfrak{q}$.

Thus $\upsilon_\mathfrak{q}(\mathcal{N}_\theta) = \upsilon_\mathfrak{q}(\mathcal{N}_{\theta'}) = 1+t$, where $t \geq 0$ is the wild part of these conductors of the characters at $\mathfrak{q}$. As E has additive reduction at $\mathfrak{q}$, we can write $\upsilon_\mathfrak{q}(\mathcal{N}) = 2+\delta$, where δ is the wild part of conductor of E at $\mathfrak{q}$. To prove (b), it is sufficient to show that $\delta = 2t$. Let $I_\mathfrak{q}^w$ be the wild inertia subgroup at $\mathfrak{q}$. As $I_\mathfrak{q}^w$ is a pro-q group, where q is the rational prime below $\mathfrak{q}$, and as $p \neq q$, we have

$$\bar{\rho}_{E,p}|_{I_\mathfrak{q}^w} \sim \begin{pmatrix} \theta|_{I_\mathfrak{q}^w} & 0 \\ 0 & \theta'|_{I_\mathfrak{q}^w} \end{pmatrix}.$$

Now the relation $\delta = 2t$ follows straightforwardly from the formula [Silverman 1994, page 380] for the wild part of the conductor of E at $\mathfrak{q}$. □

Suppose E is as in Lemma 6.3. Observe that $\operatorname{Ker}\theta$ is a K-rational subgroup of $E[p]$ of order p. Thus $E' = E/\operatorname{Ker}\theta$ is a p-isogenous elliptic curve defined over K. It is straightforward to show that

$$\bar{\rho}_{E',p} \sim \begin{pmatrix} \theta' & * \\ 0 & \theta \end{pmatrix}.$$

Thus by replacing E with a p-isogenous elliptic curve, we may swap θ and θ' in (14) as we please.

Lemma 6.4. *Let $p \geq 17$, and let $2 \leq d \leq 23$ be squarefree. Let $K = \mathbb{Q}(\sqrt{d})$. Let (a, b, c) be a nontrivial solution to the Fermat equation (2) scaled as in Corollary 5.1. Then $\bar{\rho}_{E,p}$ is irreducible.*

Proof. Suppose that $\bar{\rho}_{E,p}$ is reducible. As E has nontrivial 2-torsion, it gives rise to a K-point on $X_0(2p)$. The quadratic points on $X_0(34)$ have been determined by Ozman [≥ 2015]. These are all defined over $\mathbb{Q}(i)$, $\mathbb{Q}(\sqrt{-2})$ and $\mathbb{Q}(\sqrt{-15})$. Thus we suppose $p \geq 19$.

Let θ, θ', $\mathcal{N}_\theta$, $\mathcal{N}_{\theta'}$ be as in Lemma 6.3.

We shall first complete the proof under the assumption that p is coprime to either $\mathcal{N}_\theta$ or $\mathcal{N}_{\theta'}$. After swapping θ and θ', we can assume that p is coprime to $\mathcal{N}_\theta$. It follows from Lemma 6.3 that $\mathcal{N}_\theta$ is the square root of the additive part of the conductor $\mathcal{N}$. From Lemma 3.3, we know that the odd additive part is $\mathfrak{m}^2$ where the possibilities for $\mathfrak{m}$ are as in Corollary 3.4. The even additive part of $\mathcal{N}$ can be deduced from Table 1. For the cases where 2 is inert and $2 \nmid abc$, the even part of the conductor is $\mathcal{N}^{\text{even}} = (2)^4$ (after appropriate scaling of (a, b, c)) by Lemma 4.1.

Thus for each d, we have a small list of possibilities for $\mathcal{N}_\theta$. Let ∞_1 and ∞_2 be the two real places of K. It follows that θ is a character of the ray class group for the modulus $\mathcal{N}_\theta \infty_1 \infty_2$. Using Magma, we computed this ray class group in all cases and found it to be one of the following groups:

$$0, \quad \mathbb{Z}/2\mathbb{Z}, \quad \mathbb{Z}/4\mathbb{Z}, \quad \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}, \quad \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}.$$

The order of θ divides the exponent of the group, and so it is 1, 2 or 4. If θ has order 1, then E has a point of order p over K. The possibilities for p-torsion over quadratic fields have been determined by Kamienny, Kenku and Momose (see [Kamienny 1992, Theorem 3.1]) and their results imply that $p \leq 13$, giving a contradiction. If θ has order 2, then E has a point of order p over a quadratic extension L/K. The field L has degree 4 over $\mathbb{Q}$. The possibilities for p-torsion over quartic fields have been determined by Derickx, Kamienny, Stein and Stoll [Derickx et al. ≥ 2015] and their results imply that $p \leq 17$, again giving a contradiction. Suppose θ has order 4. Let L be the unique quadratic extension of K cut out by θ^2. Now $\phi = \theta|_{G_L}$ is a quadratic character. Twisting E/L by ϕ gives an elliptic curve defined over L with a point of order p. As before, $p \leq 17$. This completes the proof if p is coprime to either $\mathcal{N}_\theta$ or $\mathcal{N}_{\theta'}$.

From now on we assume that neither $\mathcal{N}_\theta$ nor $\mathcal{N}_{\theta'}$ is coprime to p. Observe that E is semistable at all $\mathfrak{p} \mid p$. We shall divide into cases according to whether p is inert, splits or is ramified in K.

(a) Suppose first that p is inert in K. It follows from Corollary 6.2 that E cannot have good supersingular reduction at $\mathfrak{p} = (p)$. Thus E has either good ordinary or multiplicative reduction at $\mathfrak{p}$. By Proposition 6.1(ii), we see that one of θ, θ' is unramified at $\mathfrak{p}$. It follows that one of $\mathcal{N}_\theta, \mathcal{N}_{\theta'}$ is coprime to p, giving a contradiction.

(b) Suppose now that p ramifies in K. This means that $d = p$ and d is either 19 or 23. Let $\mathfrak{p}$ be the unique prime above p in K. If E has good ordinary or multiplicative reduction at $\mathfrak{p}$ then we obtain a contradiction as in (a). Thus suppose E has good supersingular reduction at $\mathfrak{p}$. We will now apply Proposition 6.5 below to show this cannot happen. The field $K = \mathbb{Q}(\sqrt{d})$ has a prime $\mathfrak{P}$ dividing 2 with residue field $\mathbb{F}_2$, and so by Lemma 4.2, this is a prime of potentially multiplicative reduction for E. In the notation of Proposition 6.5, $\sqrt{\mathcal{A}} = (1)$ or $\mathfrak{P}^2$. In all cases, the ray class group of modulus $\sqrt{\mathcal{A}}\infty_1\infty_2$ is $\mathbb{Z}/2\mathbb{Z}$, and the proposition implies that $4 = \mathrm{Norm}(\mathfrak{P})^2 \equiv 1 \pmod{p}$. As $p = 19$ or 23, we have a contradiction and so E cannot be supersingular at $\mathfrak{P}$.

(c) Suppose p splits as $\mathfrak{p}\mathfrak{p}'$. The primes $\mathfrak{p}$, $\mathfrak{p}'$ are unramified, and again we deduce that E has either good ordinary or multiplicative reduction at these. By Proposition 6.1(ii), we have that precisely one of θ, θ' is ramified at $\mathfrak{p}$ and precisely one of them is ramified at $\mathfrak{p}'$. If θ is unramified at both $\mathfrak{p}$, $\mathfrak{p}'$ then we have a contradiction, and likewise if θ' is unramified at both $\mathfrak{p}$, $\mathfrak{p}'$. We can assume $\mathfrak{p} \mid \mathcal{N}_\theta$, $\mathfrak{p} \nmid \mathcal{N}_{\theta'}$ and $\mathfrak{p}' \nmid \mathcal{N}_\theta$, $\mathfrak{p}' \mid \mathcal{N}_{\theta'}$. Thus, by Proposition 6.1(ii), $\theta|_{I_\mathfrak{p}} = \chi|_{I_\mathfrak{p}}$ and $\theta'|_{I_{\mathfrak{p}'}} = \chi|_{I_{\mathfrak{p}'}}$. We shall write down a small integer $n > 0$ such that θ^n is unramified away from $\mathfrak{p}$. If $\mathfrak{q} \nmid p$ is a prime of potentially multiplicative reduction then θ^2 is unramified at $\mathfrak{q}$. Furthermore, for our Frey curve E, the only odd additive prime is $\mathfrak{q} = \mathfrak{m}$, and Lemma 3.2(iv) implies that $\#\bar{\rho}_{E,p}(I_\mathfrak{q}) = 2$, and so θ^2 is unramified at $\mathfrak{q}$. We are left with primes $\mathfrak{q}$, with $\mathfrak{q} \mid 2$, of potentially good reduction. These only arise for $d = 5$, 13, 21, and in these cases $\upsilon_\mathfrak{q}(\Delta_\mathfrak{q}) = 4$. It follows from [Kraus 1990, Theorem 3] that $\bar{\rho}_{E,p}(I_\mathfrak{q})$ is either cyclic of order 3, 6 or isomorphic to $\mathrm{SL}_2(\mathbb{F}_3)$ and so has order 24. The last case cannot occur, as $\mathrm{SL}_2(\mathbb{F}_3)$ is nonabelian, and any nonabelian reducible subgroup of $\mathrm{GL}_2(\mathbb{F}_p)$ has an element of order p. It follows that θ^6 is unramified at $\mathfrak{q}$. Letting $n = 6$ for $d = 5$, 13, 21, and $n = 2$ for other values of d, we conclude that the character θ^n is unramified away from $\mathfrak{p}$, and that $\theta^n|_{I_\mathfrak{p}} = \chi^n|_{I_\mathfrak{p}}$. Let u be a generator of the subgroup of totally positive units in $\mathcal{O}_K^*$. It follows (see [Kraus 1996, page 249]) that $p \mid \mathrm{Norm}(u^n - 1)$. We computed the factorization of $\mathrm{Norm}(u^n - 1)$ for our values of d and found that none are divisible by primes $p \geq 19$, except when $d = p = 19$ or $d = p = 23$. However, in these cases p ramifies in the field $K = \mathbb{Q}(\sqrt{d})$, and so are covered by case (b). $\square$

Proposition 6.5. *Let $d = p \geq 5$ be a prime and $\mathfrak{p}$ be the unique prime in $K = \mathbb{Q}(\sqrt{d})$ above p. Let E/K be an elliptic curve and denote by $\mathcal{A}$ the additive part of its conductor. Suppose that E has good supersingular reduction at $\mathfrak{p}$ and potentially multiplicative reduction at some prime $\mathfrak{q}_0 \neq \mathfrak{p}$. Suppose further that $\bar{\rho}_{E,p}$ is reducible. Therefore, $\mathcal{A}$ is a square and we let n be the exponent of the ray class group for the modulus $\sqrt{\mathcal{A}}\infty_1\infty_2$. Then,* $\mathrm{Norm}(\mathfrak{q}_0)^n \equiv 1 \pmod{p}$.

Proof. Suppose $\bar{\rho}_{E,p}$ is reducible and let θ, θ' be as in the proof of Lemma 6.4. Write $\epsilon = \theta/\theta'$. By Corollary 6.2, the character ϵ is unramified at $\mathfrak{p}$; it is here that we use the assumption that E has good supersingular reduction at $\mathfrak{p}$. Moreover, as $\epsilon = \theta^2/\chi$, where χ is the cyclotomic character, it follows that for $\mathfrak{q}$, an additive prime with $\mathfrak{q} \nmid p$,

$$\upsilon_{\mathfrak{q}}(\mathcal{N}_\epsilon) \leq \upsilon_{\mathfrak{q}}(\mathcal{N}_\theta) = \tfrac{1}{2}\upsilon_{\mathfrak{q}}(\mathcal{A}).$$

Therefore, the exponent of the ray class group of modulus $\mathcal{N}_\epsilon\infty_1\infty_2$ is a divisor of n. Thus $\epsilon^n = 1$. Let $\sigma_{\mathfrak{q}_0}$ be the Frobenius element of G_K at $\mathfrak{q}_0$. Since $\mathfrak{q}_0$ is of potentially multiplicative reduction the possible pairs of eigenvalues of $\bar{\rho}_{E,p}(\sigma_{\mathfrak{q}_0})$ are $(1, \mathrm{Norm}(\mathfrak{q}_0))$ or $(-1, -\mathrm{Norm}(\mathfrak{q}_0))$ and they correspond to the values of $\theta(\sigma_{\mathfrak{q}_0})$ and $\theta'(\sigma_{\mathfrak{q}_0})$ up to reordering. Thus,

$$1 = \epsilon^n(\sigma_{\mathfrak{q}_0}) = \theta(\sigma_{\mathfrak{q}_0})^n/\theta'(\sigma_{\mathfrak{q}_0})^n \equiv \mathrm{Norm}(\mathfrak{q}_0)^{\pm n} \pmod{p}. \qquad \square$$

7. Proof of Theorem 1

For now let $2 \leq d \leq 23$ be squarefree, and let $K = \mathbb{Q}(\sqrt{d})$. We would like to show that the equation $x^n + y^n = z^n$ has only trivial solutions in K for $n \geq 4$, although as we will see in due course, our proof strategy fails for $d = 5$ and $d = 17$. As in the introduction, we reduce to showing that the Fermat equation (2) has no nontrivial solutions (a, b, c) in $\mathcal{O}_K$ with prime exponent $p \geq 17$. Now suppose (a, b, c) is a nontrivial solution with $p \geq 17$, and scale this as in Corollary 5.1. Let $E = E_{a,b,c}$ be the Frey curve given by (3), and let $\bar{\rho}_{E,p}$ be its mod p representation. We know from Lemma 6.4 that $\bar{\rho}_{E,p}$ is irreducible. We now apply Theorem 3 to deduce that there is a cuspidal Hilbert newform $\mathfrak{f}$ over K of weight $(2, 2)$ and level $\mathcal{N}_p$ (one of the levels predicted by Corollary 5.1) such that $\bar{\rho}_{E,p} \sim \bar{\rho}_{\mathfrak{f},\varpi}$ for some prime $\varpi \mid p$ of $\mathbb{Q}_{\mathfrak{f}}$.

Lemma 7.1. *Let $\mathfrak{q} \nmid \mathcal{N}_p$ be a prime of K, and let*

$$\mathcal{A} = \{a \in \mathbb{Z} : |a| \leq 2\sqrt{\mathrm{Norm}(\mathfrak{q})},\ \mathrm{Norm}(\mathfrak{q}) + 1 - a \equiv 0 \pmod{4}\}.$$

If $\bar{\rho}_{E,p} \sim \bar{\rho}_{\mathfrak{f},\varpi}$ then ϖ divides the principal ideal

$$B_{\mathfrak{f},\mathfrak{q}} = \mathrm{Norm}(\mathfrak{q})\big((\mathrm{Norm}(\mathfrak{q}) + 1)^2 - a_{\mathfrak{q}}(\mathfrak{f})^2\big) \prod_{a\in\mathcal{A}} (a - a_{\mathfrak{q}}(\mathfrak{f})) \cdot \mathcal{O}_{\mathbb{Q}_{\mathfrak{f}}}.$$

Proof. If $\mathfrak{q} \mid p$, then $\mathrm{Norm}(\mathfrak{q})$ is a power of p. Since $\varpi \mid p$, we have ϖ divides $B_{\mathfrak{f},\mathfrak{q}}$. Thus we may suppose $\mathfrak{q} \nmid p$. By assumption $\mathfrak{q} \nmid \mathcal{N}_p$. From the definition of $\mathcal{N}_p$

in (6), the prime $\mathfrak{q}$ is of good or multiplicative reduction for E. If $\mathfrak{q}$ is a prime of good reduction for E, then $a_\mathfrak{q}(E) \equiv a_\mathfrak{q}(\mathfrak{f}) \pmod{\varpi}$. By the Hasse–Weil bounds, we know that $|a_\mathfrak{q}(E)| \le 2\sqrt{\mathrm{Norm}(\mathfrak{q})}$. Moreover, as E has full 2-torsion (and $\mathfrak{q} \nmid 2$, as $\mathfrak{q} \nmid \mathcal{N}_p$), we have $4 \mid \#E(\mathbb{F}_\mathfrak{q})$. Thus $a_\mathfrak{q}(E) \in \mathcal{A}$ and so $\varpi \mid B_{\mathfrak{f},\mathfrak{q}}$. Finally, suppose $\mathfrak{q}$ is a prime of multiplicative reduction for $\mathcal{N}_p$. Then, comparing the traces of the images of Frobenius at $\mathfrak{q}$ under $\bar{\rho}_{E,p}$ and $\bar{\rho}_{\mathfrak{f},\varpi}$, we have

$$\pm(\mathrm{Norm}(\mathfrak{q}) + 1) \equiv a_\mathfrak{q}(\mathfrak{f}) \pmod{\varpi}.$$

It follows that ϖ divides $B_{\mathfrak{f},\mathfrak{q}}$ in this case too. □

Using Magma we computed the newforms $\mathfrak{f}$ at the predicted levels, the fields $\mathbb{Q}_\mathfrak{f}$, and eigenvalues $a_\mathfrak{q}(\mathfrak{f})$ at primes $\mathfrak{q}$ of K small norm. We computed for each $\mathfrak{f}$ at level $\mathcal{N}_p$, the ideal

$$B_\mathfrak{f} := \sum_{\mathfrak{q} \in T} B_{\mathfrak{f},\mathfrak{q}}, \tag{15}$$

where T is the set of prime ideals $\mathfrak{q}$ of K, with $\mathfrak{q} \nmid \mathcal{N}_p$ and with norm less than 60 (this turns out to be sufficient for our purpose). Let $C_\mathfrak{f} := \mathrm{Norm}_{\mathbb{Q}_\mathfrak{f}/\mathbb{Q}}(B_\mathfrak{f})$. If $\bar{\rho}_{E,p} \sim \bar{\rho}_{\mathfrak{f},\varpi}$ then by the above lemma, $\varpi \mid B_\mathfrak{f}$ and so $p \mid C_\mathfrak{f}$. Hence, the isomorphism $\bar{\rho}_{E,p} \sim \bar{\rho}_{\mathfrak{f},\varpi}$ is impossible if $p \nmid C_\mathfrak{f}$. Thus, the newforms satisfying $C_\mathfrak{f} = 0$ are the problematic ones. We computed $C_\mathfrak{f}$ for all newforms $\mathfrak{f}$ at the predicted levels, and found only three fields where $C_\mathfrak{f} = 0$ for some $\mathfrak{f}$. All the others produced values of $C_\mathfrak{f}$ that are not divisible by primes $p \ge 17$. Thus to complete the proof, we have to deal with the cases where $C_\mathfrak{f} = 0$; these are as follows:

(i) $K = \mathbb{Q}(\sqrt{3})$, $\mathcal{N}_p = (1+\sqrt{3})^4$. Here $\mathfrak{f}$ is the unique newform at level $\mathcal{N}_p$. It satisfies $\mathbb{Q}_\mathfrak{f} = \mathbb{Q}$ and corresponds to the elliptic curve

$$E' : y^2 = x(x+1)(x+8+4\sqrt{3})$$

of conductor $(1+\sqrt{3})^4$.

(ii) $K = \mathbb{Q}(\sqrt{5})$, $\mathcal{N}_p = (2)^4$. There are three newforms at level $\mathcal{N}_p$, and all three satisfy $\mathbb{Q}_\mathfrak{f} = \mathbb{Q}$. For all three newforms, $C_\mathfrak{f} = 0$.

(iii) $K = \mathbb{Q}(\sqrt{17})$, $\mathcal{N}_p = (2)$. Here $\mathfrak{f}$ is the unique newform at level $\mathcal{N}_p$. It satisfies $\mathbb{Q}_\mathfrak{f} = \mathbb{Q}$ and corresponds to the elliptic curve

$$W : y^2 = x(x-4+\sqrt{17})\left(x + \frac{-13+5\sqrt{17}}{2}\right)$$

of conductor (2).

Indeed, the remaining eigenforms $\mathfrak{f}$ correspond to elliptic curves with full 2-torsion. It is easy to see from the definitions that $B_{\mathfrak{f},\mathfrak{q}} = 0$ for such an eigenform. It follows that it is futile to enlarge the set T in (15).

Since $d \neq 5, 17$ in the statement of Theorem 1, we only have to complete the proof for $d = 3$. To do this, we must discard the isomorphism $\bar{\rho}_{E,p} \sim \bar{\rho}_{E',p}$, where E' is given in (i) above. The elliptic curve E' has j-invariant $j' = 54000$ and so potentially good reduction at $\mathfrak{P} = (1 + \sqrt{3})$; in particular [Kraus 1990], the order of $\bar{\rho}_{E',p}(I_{\mathfrak{P}})$ is 1, 2, 3, 4, 6, 8 or 24. On the other hand, the Frey curve E has potentially multiplicative reduction, and $p \nmid \upsilon_{\mathfrak{P}}(j)$. By the theory of the Tate curve [Silverman 1994, Proposition V.6.1], we have $p \mid \bar{\rho}_{E',p}(I_{\mathfrak{P}})$, giving a contradiction as $p \geq 17$.

8. Proof of Theorem 2

To complete the proof of Theorem 2, it remains to discard the isomorphism $\bar{\rho}_{E,p} \sim \bar{\rho}_{W,p}$, where W is given in (iii) above. We apply Lemma 1.6 of [Halberstadt and Kraus 2002]—this is proved for $K = \mathbb{Q}$, but the proof for K, a general number field, is identical. Let $\mathfrak{P}_1$, $\mathfrak{P}_2$ be as in Table 1 for $d = 17$. The curve E has multiplicative reduction at $\mathfrak{P}_i$, and the valuations of the minimal discriminants are $-8 + 2pt_i$, where t_1, t_2 are positive integers. The curve W has conductor $(2) = \mathfrak{P}_1\mathfrak{P}_2$ and its minimal discriminant Δ_W satisfies $\upsilon_{\mathfrak{P}_1}(\Delta_W) = 4$ and $\upsilon_{\mathfrak{P}_2}(\Delta_W) = 2$. The quantity

$$\frac{\upsilon_{\mathfrak{P}_1}(\Delta_E)\upsilon_{\mathfrak{P}_2}(\Delta_E)}{\upsilon_{\mathfrak{P}_1}(\Delta_W)\upsilon_{\mathfrak{P}_2}(\Delta_W)} = \frac{(-8 + 2pt_1)(-8 + 2pt_2)}{4 \cdot 2} \equiv 8 \pmod{p}$$

is a square modulo p if and only if $p \equiv 1, 7 \pmod{8}$. It follows from [Halberstadt and Kraus 2002, Lemma 1.6] that $\bar{\rho}_{E,p} \sim \bar{\rho}_{W,p}$ cannot hold if $p \equiv 3, 5 \pmod{8}$, concluding the proof.

9. Computational remarks

In the introduction, we indicated that the above strategy can be applied over other totally real fields (assuming the modularity of the Frey curve). However, the computation of newforms will often be impractical, particularly if the levels predicted by level lowering have large norm. These levels depend crucially on a choice of odd prime ideal representatives $\mathcal{H}$ for $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$. In this section, we illustrate these computational issues by looking at $K = \mathbb{Q}(\sqrt{30})$ and $K = \mathbb{Q}(\sqrt{79})$.

Let $K = \mathbb{Q}(\sqrt{30})$. Here $\mathrm{Cl}(K)$ has order 2, and we can take $\mathcal{H} = \{1 \cdot \mathcal{O}_K, \mathfrak{m}\}$, where $\mathfrak{m}$ is the unique prime above 3. By computations similar to those leading to Corollary 5.1, we obtain four possible levels $\mathcal{N}_p$. One of these is $\mathcal{N}_p = \mathfrak{P}^8 \cdot \mathfrak{m}^2$, where $\mathfrak{P}$ is the unique prime above 2. The dimension of the space of cusp forms of level $\mathcal{N}_p$ is 26108, making the computation of newforms infeasible with the current Magma implementation.

Let $K = \mathbb{Q}(\sqrt{79})$. Here $\mathrm{Cl}(K)$ has order 3, and thus $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$ is trivial; this is the smallest real quadratic field for which $\mathrm{Cl}(K)$ and $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$ differ. By definition, $\mathcal{H} = \{1 \cdot \mathcal{O}_K\}$. We can show by variants of the arguments in Section 6 that $\bar{\rho}_{E,p}$ is irreducible for $p \geq 17$. Moreover the predicted levels $\mathcal{N}_p$ are $\mathfrak{P}$ and $\mathfrak{P}^4$, where $\mathfrak{P} \mid 2$. The dimensions of the corresponding spaces of cusp forms are 156 and 1077 respectively. Here it feasible to compute the newforms, and similar arguments to those in Section 7 allow us to deduce the following.

Theorem 4. *The Fermat equation (1) has only trivial solutions over $K = \mathbb{Q}(\sqrt{79})$ for $n \geq 4$.*

In [Freitas and Siksek 2015], the reader will also find recipes for the possible levels $\mathcal{N}_p$. The objectives of [loc. cit.] are theoretical and there is no need to make the levels $\mathcal{N}_p$ particularly small. The purpose of the following remarks is to illustrate the value of Sections 3 and 4 of the current paper, where a finer analysis of the levels and the effect of scaling the solution is carried out. In [loc. cit.], the set $\mathcal{H}$ is taken to be representatives of $\mathrm{Cl}(K)$ rather than representatives for $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$. For the fields K appearing in Theorem 1, all class groups are either trivial or cyclic of order 2. Therefore there is no difference between $\mathrm{Cl}(K)$ and $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$. For these fields, the main improvement of the current paper lies in Section 4 which radically reduces the possibilities for the even part of the level. However, to extend the computations to other fields, the distinction between $\mathrm{Cl}(K)$ and $\mathrm{Cl}(K)/\mathrm{Cl}(K)^2$ becomes crucial. For example, for $K = \mathbb{Q}(\sqrt{79})$, a set of odd representatives for $\mathrm{Cl}(K)$ is $\{1 \cdot \mathcal{O}_K,\ \mathfrak{m}_1,\ \mathfrak{m}_2\}$, where $\mathfrak{m}_1\mathfrak{m}_2 = 3 \cdot \mathcal{O}_K$. Following the recipe in [loc. cit.], the odd part of the level will be $1 \cdot \mathcal{O}_K$, $\mathfrak{m}_1^2$ or $\mathfrak{m}_2^2$. Thus the possibilities for $\mathcal{N}_p$ include $\mathfrak{P}^4 \cdot \mathfrak{m}_1^2$. The dimension of the space of cusp forms for this level is 12090, which makes the computation of newforms impractical. Finally, we point out that the even part of the level given by the recipe in [loc. cit.] can be as large as $\mathfrak{P}^{10}$. Even if the odd part of the level is taken to be trivial, the dimension of the space of cusp forms of level $\mathfrak{P}^{10}$ is 64596, which again is too large. It is clear that the refinements of Sections 3 and 4 are required for $K = \mathbb{Q}(\sqrt{79})$, and will be needed if the computations of the current paper are to be extended to other totally real fields.

Acknowledgements

We are grateful to the referee for his careful reading of the paper and for pointing out many improvements, and to Bjorn Poonen for suggesting a correction to the statement of Theorem 1.

References

[Aigner 1934] A. Aigner, "Uber die Möglichkeit von $x^4 + y^4 = z^4$ in quadratischen Körpern", *Jahresber. Dtsch. Math.-Ver.* **43** (1934), 226–229. Zbl 0008.29502

[Aigner 1957] A. Aigner, "Die Unmöglichkeit von $x^6 + y^6 = z^6$ und $x^9 + y^9 = z^9$ in quadratischen Körpern", *Monatsh. Math.* **61** (1957), 147–150. MR 19,120e Zbl 0079.06502

[Bosma et al. 1997] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system I: the user language", *J. Symbolic Comput.* **24**:3-4 (1997), 235–265. MR 1484478 Zbl 0898.68039

[Cassels and Fröhlich 1967] J. W. S. Cassels and A. Fröhlich (editors), *Algebraic number theory*, Academic Press, London, 1967. MR 35 #6500 Zbl 0153.07403

[David 2012] A. David, "Caractère d'isogénie et critères d'irréductibilité", preprint, 2012. arXiv 1103.3892v2

[Dembélé and Donnelly 2008] L. Dembélé and S. Donnelly, "Computing Hilbert modular forms over fields with nontrivial class group", pp. 371–386 in *Algorithmic number theory*, edited by A. J. van der Poorten and A. Stein, Lecture Notes in Comput. Sci. **5011**, Springer, Berlin, 2008. MR 2010d:11149 Zbl 1232.11128

[Dembélé and Voight 2013] L. Dembélé and J. Voight, "Explicit methods for Hilbert modular forms", pp. 135–198 in *Elliptic curves, Hilbert modular forms and Galois deformations*, edited by H. Darmon et al., Springer, Basel, 2013. MR 3184337 Zbl 1276.11004

[Derickx et al. $\geq$ 2015] M. Derickx, S. Kamienny, W. Stein, and M. Stoll, "Torsion points on elliptic curves over number fields", in preparation.

[Dickson 1920] L. E. Dickson, *History of the theory of numbers, Volume II: Diophantine analysis*, Carnegie Institution of Washington, 1920. Zbl 47.0100.04

[Freitas and Siksek 2015] N. Freitas and S. Siksek, "The asymptotic Fermat's last theorem for five-sixths of real quadratic fields", *Compos. Math.* (online publication May 2015).

[Freitas et al. 2014] N. Freitas, B. Le Hung, and S. Siksek, "Elliptic curves over real quadratic fields are modular", *Invent. Math.* (online publication November 2014).

[Fujiwara 2006] K. Fujiwara, "Level optimisation in the totally real case", preprint, 2006. arXiv math/0602586v1

[Gross and Rohrlich 1978] B. H. Gross and D. E. Rohrlich, "Some results on the Mordell–Weil group of the Jacobian of the Fermat curve", *Invent. Math.* **44**:3 (1978), 201–224. MR 58 #10911 Zbl 0369.14011

[Halberstadt and Kraus 2002] E. Halberstadt and A. Kraus, "Courbes de Fermat: résultats et problèmes", *J. Reine Angew. Math.* **548** (2002), 167–234. MR 2003h:11068 Zbl 1125.11038

[Jarvis 2004] F. Jarvis, "Correspondences on Shimura curves and Mazur's principle at p", *Pacific J. Math.* **213**:2 (2004), 267–280. MR 2004m:11066 Zbl 1073.11030

[Jarvis and Meekin 2004] F. Jarvis and P. Meekin, "The Fermat equation over $\mathbb{Q}(\sqrt{2})$", *J. Number Theory* **109**:1 (2004), 182–196. MR 2005g:11045 Zbl 1078.11019

[Kamienny 1992] S. Kamienny, "Torsion points on elliptic curves and q-coefficients of modular forms", *Invent. Math.* **109**:2 (1992), 221–229. MR 93h:11054 Zbl 0773.14016

[Kraus 1990] A. Kraus, "Sur le défaut de semi-stabilité des courbes elliptiques à réduction additive", *Manuscripta Math.* **69**:4 (1990), 353–385. MR 91j:11045 Zbl 0792.14014

[Kraus 1996] A. Kraus, "Courbes elliptiques semi-stables et corps quadratiques", *J. Number Theory* **60**:2 (1996), 245–253. MR 97e:11065 Zbl 0865.11050

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

[Ozman $\geq$ 2015] E. Ozman, "Quadratic points on non-hyperelliptic modular curves", in preparation.

[Rajaei 2001] A. Rajaei, "On the levels of mod l Hilbert modular forms", *J. Reine Angew. Math.* **537** (2001), 33–65. MR 2002i:11041 Zbl 0982.11023

[Ribet 1990] K. A. Ribet, "On modular representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ arising from modular forms", *Invent. Math.* **100**:2 (1990), 431–476. MR 91g:11066 Zbl 0773.11039

[Rohrlich 1994] D. E. Rohrlich, "Elliptic curves and the Weil–Deligne group", pp. 125–157 in *Elliptic curves and related topics*, edited by H. Kisilevsky and M. R. Murty, CRM Proc. Lecture Notes **4**, Amer. Math. Soc., Providence, RI, 1994. MR 95a:11054 Zbl 0852.14008

[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

[Silverman 1986] J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics **106**, Springer, New York, 1986. MR 87g:11070 Zbl 0585.14026

[Silverman 1994] J. H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Graduate Texts in Mathematics **151**, Springer, New York, 1994. MR 96b:11074 Zbl 0911.14015

[Tzermias 2004] P. Tzermias, "Low-degree points on Hurwitz–Klein curves", *Trans. Amer. Math. Soc.* **356**:3 (2004), 939–951. MR 2004j:11061 Zbl 1116.14017

[Wiles 1995] A. Wiles, "Modular elliptic curves and Fermat's last theorem", *Ann. of Math.* (2) **141**:3 (1995), 443–551. MR 96d:11071 Zbl 0823.11029

Communicated by John Henry Coates
Received 2014-07-16 Revised 2014-09-23 Accepted 2015-03-09

nunobfreitas@gmail.com *Mathematisches Institut, Universität Bayreuth, 95440 Bayreuth, Germany*

samir.siksek@gmail.com *Mathematics Institute, University of Warwick, Coventry, CV4 7AL, United Kingdom*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.897

Bounded negativity of self-intersection numbers of Shimura curves in Shimura surfaces

Martin Möller and Domingo Toledo

Shimura curves on Shimura surfaces have been a candidate for counterexamples to the bounded negativity conjecture. We prove that they do not serve this purpose: there are only finitely many whose self-intersection number lies below a given bound.

Previously (*Duke Math. J.* **162**:10 (2013), 1877–1894), this result was shown for compact Hilbert modular surfaces using the Bogomolov–Miyaoka–Yau inequality. Our approach uses equidistribution and works uniformly for all Shimura surfaces.

Introduction

Let X be a Shimura surface not isogenous to a product, i.e., an algebraic surface which is the quotient of a two-dimensional hermitian symmetric space G/K by an irreducible arithmetic lattice in G. The aim of this note is to show that Shimura curves on such a Shimura surface do not provide a counterexample to the bounded negativity conjecture. More precisely we show:

Theorem 0.1. *For any Shimura surface X not isogenous to a product and for any real number M, there are only finitely many compact Shimura curves C on X with $C^2 < M$.*

The bounded negativity conjecture claims that for any smooth projective algebraic surface X there is a positive constant B so that for any irreducible curve C on X the self-intersection C^2 is at least $-B$. We emphasize that the above theorem does not decide the validity on any Shimura surface, as there could exist non-Shimura curves with arbitrarily negative self-intersection.

Möller is partially supported by the ERC-StG 257137. Toledo is partially supported by Simons Foundation Collaboration Grant 208853.

MSC2010: primary 14G35; secondary 14J25, 22E40, 37A99, 53C55.

Keywords: bounded negativity, Shimura curves, self-intersections, equidistribution of Shimura curves.

There are two possibilities for the uniformization of X. The first case is Shimura surfaces uniformized by $\mathbb{H}^2$. In this case, $G = \mathrm{SL}_2(\mathbb{R})^2$ and the surfaces are called *quaternionic Shimura surfaces* if Γ is cocompact and *Hilbert modular surfaces* if Γ has cusps. The second case are Shimura surfaces uniformized by the complex 2-ball $\mathbb{B}^2$. In this case, $G = \mathrm{SU}(2, 1)$ and the surfaces are called *Picard modular surfaces*. There are compact and noncompact Picard modular surfaces. The assumption on the Shimura surface is necessary, since the theorem is certainly false in the product situation, e.g., for $X = X(d) \times X(d)$ a product of modular curves or a finite quotient of such a surface: the fiber classes give infinitely many curves with self-intersection zero.

While only the case of compact X is relevant to the bounded negativity conjecture, the proofs for noncompact X are the same. When both X and the curves C are allowed to have cusps the proper formulation is needed; see Theorem 3.6.

Theorem 0.1 was proven for compact Shimura surfaces uniformized by $\mathbb{H}^2$ in [Bauer et al. 2013]. The methods there, based on the logarithmic Bogomolov–Miyaoka–Yau inequality, do not extend to the ball quotient case. Here we give a uniform treatment of both cases based on equidistribution results. As in that paper, we obtain as a consequence:

Corollary 0.2. *There are only finitely many Shimura curves on X that are smooth.*

Intersection numbers of Shimura curves are known to appear as coefficients of modular forms, and coefficients of modular forms are known to grow. This, however, does not directly give a method to prove Theorem 0.1, since in these modularity statements [Hirzebruch and Zagier 1976; Kudla 1978] the Shimura curves are packaged to reducible curves T_N with an unbounded number of components as $N \to \infty$, while the statement here is for every individual Shimura curve.

1. Shimura curves on Shimura surfaces not isogenous to a product

A *Shimura surface not isogenous to a product* is a connected algebraic surface that can be written as a quotient $X = \Gamma \backslash G/K$, where $G = G_{\mathbb{Q}}(\mathbb{R})$ is the set of $\mathbb{R}$-valued points in a connected semisimple $\mathbb{Q}$-algebraic group $G_{\mathbb{Q}}$, $K \subset G$ is a maximal compact subgroup and Γ is an irreducible arithmetic lattice in G. Here a lattice is called irreducible if it does not have a finite-index subgroup that splits as a product of two lattices.

Our geometric definition of Shimura varieties differs from the arithmetic literature on this subject, where Shimura varieties are typically not connected. It is the point of view of the bounded negativity conjecture that requires one to deal with irreducible components of the objects in question. Note that we do not require Γ to be a congruence subgroup either.

Definition. Let $H_{\mathbb{Q}}$ be a $\mathbb{Q}$-algebraic group, Δ an arithmetic lattice in $H_{\mathbb{Q}}$, and $\tau : H_{\mathbb{Q}} \to G_{\mathbb{Q}}$ a $\mathbb{Q}$-morphism such that $\tau(\Delta) \subset \Gamma$. Suppose the τ-preimage of a maximal compact subgroup $K \subset G_{\mathbb{R}}$ is a maximal compact subgroup $K_H \subset H = H_{\mathbb{Q}}(\mathbb{R})$. Then the algebraic curve C in X given by $C = \Delta \backslash H / K_H$ is called a *Shimura curve*.

The aim of this section is to compile the list of possible constructions of Shimura surfaces that contain infinitely many Shimura curves and the possible pairs $(G_{\mathbb{Q}}, H_{\mathbb{Q}})$. This will be used in the equidistribution theorem in the next section. More precisely, we need that all Shimura curves can be generated as the orbit of a fixed subgroup. For this purpose we write $G = G_0 \times W$ with W compact and G_0 without compact factors. There is a corresponding decomposition of the compact subgroup $K = K_0 \times W$, and also for the Shimura curve $H = H_0 \times W_H$ and $K_H = K_{H,0} \times W_H$.

It turns out that there are only two possibilities for G_0, and, for each of them, we can construct all Shimura curves as follows.

Proposition 1.1. *For a given Shimura surface $X = \Gamma \backslash G_0 / K_0 = \Gamma \backslash G / K$ not isogenous to a product, there exists a subgroup $H_0 \cong \mathrm{SL}_2(\mathbb{R})$ of G_0 such that all Shimura curves arise as $C = \Gamma \backslash \Gamma g H_0 / K_{H_0}$ for some $g \in G_0$.*

We start with the possibilities for G_0. There are only two hermitian symmetric domains of dimension two. This leads to the following two cases, as in the introduction. In each case we give a description of the possible Shimura surfaces. Here, and elsewhere, the description of the algebraic groups in question will always be given only up to central isogeny.

Case One: $G_0 = \mathrm{SL}_2(\mathbb{R})^2$. There two possibilities. Either G is the set of $\mathbb{R}$-points of the $\mathbb{Q}$-algebraic group $G_{\mathbb{Q}} = \mathrm{Res}_{F/\mathbb{Q}}(\mathrm{SL}_2(A))$ for a quaternion algebra A over a totally real field F which is unramified at exactly two infinite places of F, or G is the product $\mathrm{Res}_{F/\mathbb{Q}}(\mathrm{SL}_2(A_1)) \times \mathrm{Res}_{F/\mathbb{Q}}(\mathrm{SL}_2(A_2))$ for two quaternion algebras A_1, A_2, each unramified at exactly at one infinite place. For the proofs, first remark that these give F-forms of $\mathrm{SL}_2(\mathbb{R})^2$; see, e.g., [Vignéras 1980, IV.1]. That these are the only possibilities follows from the classification of algebraic groups [Tits 1966]. In more detail, the procedure of [Tits 1966, §3.1] reduces the problem to the classification of F-forms of SL_2. The description in [Serre 1994, III.1.4] of the F-forms of SL_2 in bijective correspondence with quaternion algebras over F gives the above description of the algebraic groups. In both cases, the maximal compact subgroup K in G is $\mathrm{SO}_2(\mathbb{R})^2$ times the compact factors of $G_{\mathbb{R}}$.

In the product case, all lattices are reducible, so we can discard this case in view of our irreducibility hypothesis on X. In the remaining case, in order obtain an arithmetic lattice $\Gamma \subset G$ one has to fix an order $\mathfrak{O} \subset A$ and let $\mathfrak{O}^1 \subset \mathfrak{O}$ be the elements of reduced norm 1. Then Γ is the image in G of a group commensurable with $\mathfrak{O}^1$. See, e.g., [Vignéras 1980] for more details.

Case Two: $G_0 = \mathrm{SU}(2,1)$. In this case the underlying $\mathbb{Q}$-algebraic group is $G_{\mathbb{Q}} = \mathrm{Res}_{F_0/\mathbb{Q}}(G_{F_0}))$, and from the classification of algebraic groups (over number fields) [Tits 1966; Platonov and Rapinchuk 1994], we see that, in the notation of [Tits 1966, p. 55] G_{F_0} must be of type ${}^2A^{(d)}_{2,r}$, where $d \mid 3$, $d \geq 1$ and $2rd \leq 3$. In other words, $G_{F_0} = \mathrm{SU}(h)$, where h is a hermitian form constructed as follows. Start with a totally real field F_0 and take a totally complex quadratic extension F/F_0, i.e., F is a CM field. Then take a central simple division algebra D of degree d (hence dimension d^2) over F, with center F and involution σ of the second kind (not the identity on F), and a hermitian form h on $D^{3/d}$ so that h is isotropic at one real place of F_0 and definite at all other real places (equivalently, isotropic at one conjugate pair of complex places of F, definite at all other pairs).

Thus there are two "types" corresponding to the two possibilities $d = 1$ or $d = 3$:

The *first type* corresponds to $d = 1$. Then $D = F$ and h is a hermitian form on F^3 that is definite except for one pair of places of F, interchanged by complex conjugation. Then $\mathrm{SU}(h)$ is indeed a F_0-algebraic group and the set of $\mathbb{R}$-valued points of $\mathrm{Res}_{F_0/\mathbb{Q}}(\mathrm{SU}(h))$ equals G_0 up to compact factors. The compact subgroup K in G is $S(U(2) \times U(1))$ times the compact factors of $G_{\mathbb{R}}$. Arithmetic lattices Γ *of the first type* are obtained by fixing an order $\mathfrak{O} \subset F$ and taking Γ commensurable with $G \cap \mathrm{SL}_3(\mathfrak{O})$. The integer r above satisfying $2rd \leq 3$ is the F_0-rank of G_{F_0}, or the dimension of the maximal isotropic subspace of h in F^3. The lattice is cocompact if and only if $r = 0$, and $r = 1$ forces F_0 to be $\mathbb{Q}$.

The *second type* corresponds to $d = 3$; in this case, D is central simple division algebra of degree 3 (dimension 9) over F with an involution "of the second kind". The lattices Γ are obtained by fixing an order $\mathfrak{O} \subset D$ and taking Γ commensurable with $G \cap \mathrm{SL}(D)$. Observe that in this case the inequality $2rd \leq 3$ forces r to be 0, and therefore Γ is always cocompact. We will see that lattices of the second type do not have any Shimura curves, so we will not need to consider them.

Shimura curves in X for $G_0 = \mathrm{SL}_2(\mathbb{R})^2$. The Shimura curves in X are totally geodesic complex curves in X, so they are projections to X of totally geodesic holomorphic disks $\mathbb{H} \subset \mathbb{H}^2$, which in turn are orbits of embeddings of $\mathrm{SL}_2(\mathbb{R}) \subset \mathrm{SL}_2(\mathbb{R})^2$. It is well known that, up to biholomorphic isometries, there are only two classes of such disks: factors and diagonals. By the irreducibility hypothesis, the inclusion into one factor does not come from a morphism of the underlying $\mathbb{Q}$-algebraic groups. So $H_0 \subset G_0$ has to be the diagonal embedding, proving Proposition 1.1 in this case. In fact, the possible embeddings are discussed in great detail in [van der Geer 1988] for Hilbert modular surfaces and in [Granath 2002] for quaternionic Shimura surfaces.

Shimura curves in X for $G_0 = \mathrm{SU}(2, 1)$. Fix a Shimura surface X obtained by choosing F_0, F, d, D, σ, h, $\mathfrak{O} \subset D$, Γ. The Shimura curves, being totally geodesic complex curves, are projections to X of orbits in the universal cover of subgroups $H \subset G_0$, all isomorphic to $\mathrm{SU}(1, 1)$ and standardly embedded in $\mathrm{SU}(2, 1)$. The image in X of an H-orbit is a Shimura curve if and only if $H \cap \Gamma$ is a lattice in H. This happens if and only if H is defined over F_0, meaning that the underlying algebraic group G_{F_0} contains an F_0-subgroup H_{F_0} so that, if $\iota : F_0 \to \mathbb{R}$ is the embedding of F_0 with group of real points $G_{F_0,\iota}(\mathbb{R})$ isomorphic to G_0, the inclusion $H_{F_0,\iota}(\mathbb{R}) \subset G_{F_0,\iota}(\mathbb{R})$ agrees with $H \subset G_0$. There are two cases:

No Shimura curves in Shimura surfaces of the second type. The group $\mathrm{SU}(h)$, for h a hermitian form on a central simple division algebra D over F of degree 3 as above, has no subgroup H_{F_0} defined over F_0 with $H_{F_0}(\mathbb{R}) = \mathrm{SU}(1, 1)$ standardly embedded in $\mathrm{SU}(h)(\mathbb{R}) = \mathrm{SU}(2, 1)$.

This is well-known to experts, but we do not know a reference (but see [Garibaldi and Gille 2009, Corollary 4.2] for a more general result). Matthew Stover kindly communicated the following proof:

Let F_0, F, D, σ be as above. The D-valued hermitian form h can be taken to be $h(x, y) = \sigma(x)y$, and the group of F_0-points of the F_0-group in question is

$$\mathrm{SU}(D, \sigma)(F_0) = \{x \in D : \sigma(x)x = e, Nrd(x) = 1\} \subset D,$$

which gives us an $\mathrm{SU}(2, 1)$ as follows: choose an embedding $F \to \mathbb{C}$, use it to form $D \otimes_F \mathbb{C}$, which becomes isomorphic to the algebra $M(3, \mathbb{C})$ of 3×3 complex matrices, under an isomorphism (unique up to conjugation by Skolem–Noether) which takes σ to its conjugate-transpose with respect to a hermitian form h'. Whenever all choices can be made so that h' has signature $(2, 1)$, the group of real points of $\mathrm{SU}(D, \sigma)$ becomes the standard $\mathrm{SU}(2, 1)$. The signature of the hermitian form h' depends only on D, σ and the embedding $F \to \mathbb{C}$.

Note that the F-algebra D is embedded in the algebra $M(3, \mathbb{C})$ by $x \mapsto x \otimes 1$. The F-vector subspace of $M(3, \mathbb{C})$ generated by the subset $\mathrm{SU}(D, \sigma)(F_0)$ is easily seen to be a σ-stable subalgebra of $M(3, \mathbb{C})$ contained in the division algebra D, hence it is itself a division algebra, and easily seen to equal D. Suppose H_{F_0} is an F_0-subgroup of $\mathrm{SU}(D, \sigma)$, so that the corresponding inclusion of real points is a standard embedding of $\mathrm{SU}(1, 1)$ in $\mathrm{SU}(2, 1)$, all inside $M(3, \mathbb{C})$, and let V be the F-vector subspace of $M(3, \mathbb{C})$ generated by the F_0-points of H_{F_0}. This is a noncommutative division subalgebra of D, and it must be a proper subalgebra because $V \otimes_F \mathbb{C}$ is a proper subspace of $D \otimes_F \mathbb{C} = M(3, \mathbb{C})$. Since D has degree 3, it has no proper noncommutative F-subalgebras, so such subgroups cannot exist.

Classification of Shimura curves in Shimura surfaces of the first type. In this case, there are always infinitely many Shimura curves. We continue the same notation: choose an embedding of F in $\mathbb{C}$ so that the hermitian form h is isotropic, then extend h from F^3 to $\mathbb{C}^3$. Interpret the unit ball $G_0/K_0 \cong \mathbb{B}^2 \subset \mathbb{P}^2$ as the collection of h-negative lines in $\mathbb{C}^3$. The Shimura curves in X arise as the quotient of totally geodesic disks $\mathbb{B}^1 \subset \mathbb{B}^2$, and such disks are in bijective correspondence with the h-positive lines. Namely, an h-positive line l determines the hermitian space $(\ell^\perp, h|_{l^\perp})$ of signature $(1, 1)$ and the corresponding space of negative lines $\mathbb{B}^1_l \subset \mathbb{B}^2$. All geodesic disks arise this way. The groups G_ℓ, the stabilizer of ℓ (isomorphic to $U(1, 1)$) and the subgroup H_l fixing l pointwise (isomorphic to $\mathrm{SU}(1, 1)$) act on $(\ell^\perp, h|_{l^\perp})$ and $\mathbb{B}^1_\ell$, both actions being transitive on $\mathbb{B}^1_l$. The disk $\mathbb{B}^1_\ell$ projects to a Shimura curve in X if and only if $H_\ell \cap \Gamma$ a lattice in H_ℓ; in turn:

Lemma 1.2. *The group $H_\ell \cap \Gamma$ is a lattice in H_ℓ if and only if ℓ is an F-rational line, that is, $\ell \cap F^3 \neq \{0\}$.*

Proof. Let $v \in \mathbb{C}^3$ be a basis vector for ℓ, and suppose $\Gamma_\ell = H_\ell \cap \Gamma$ is a lattice in H_ℓ. Since Γ_ℓ fixes ℓ pointwise, v is fixed by all $\gamma \in H_\ell \cap \Gamma$. Since Γ_ℓ leaves $\ell^\perp$ invariant, the remaining eigenvectors of any $\gamma \in \Gamma_\ell$ lie in $\ell^\perp$. Since the action of H_ℓ on $l^\perp$ is isomorphic to the standard action of $\mathrm{SU}(1, 1)$ on $\mathbb{C}^2$ and Γ_ℓ is a lattice in H_ℓ, Γ_l contains hyperbolic elements. Fix such an element γ. Then $\gamma(v) = v$ and the remaining eigenvalues of γ are of absolute value $\neq 1$. Therefore 1 is a simple eigenvalue of γ, and thus the space of solutions of $\gamma(v) = v$ is an F-rational line, as asserted.

For the converse, suppose that ℓ is a rational line, and let $v \in \mathfrak{O}^3$ be a primitive vector which is a basis for ℓ. Let $M_0 = \mathfrak{O} v$ and $M_1 = v^\perp \cap \mathfrak{O}^3$, and let $M = M_0 \oplus M_1$. Then M is an $\mathfrak{O}$-submodule of finite index in $\mathfrak{O}^3$. Consequently, Γ is commensurable with $\Gamma' = \{\gamma \in \mathrm{SU}(h, \mathfrak{O}) : \gamma(M) = M\}$ and $\Gamma \cap H_l$ is commensurable with $\Gamma'_v = \{\gamma \in \Gamma' : \gamma(v) = v\}$, which is a lattice in the group $H_\ell = H_v = \{g \in G : g(v) = v\}$, a group defined over F_0, and isomorphic (over F_0) to $\mathrm{SU}(h|_{M_1 \otimes F})$. This group in turn is isomorphic over $\mathbb{R}$ to $\mathrm{SU}(1, 1)$. Thus $\Gamma \cap H_\ell$ is a lattice in H_ℓ and we obtain a Shimura curve associated to the $\mathbb{Q}$-group $\mathrm{Res}_{F_0/\mathbb{Q}}(\mathrm{SU}(h|_{M_1 \otimes F}))$. □

End of proof of Proposition 1.1. Choose an orthogonal basis v_1, v_2, v_3 for $\mathfrak{O}^3$, where $h(v_i) = a_i \bar{a}_i > 0$ for $i = 1, 2$, $h(v_3) = -a_3 \bar{a}_3 < 0$ and $v_1 \in \ell$. Let e_1, e_2, e_3 be the standard basis for $\mathbb{C}^3$, let $H = H_{e_1} \subset G$ be the subgroup, isomorphic to $\mathrm{SU}(1, 1)$, that fixes e_1, and let $g \in G$ be the linear transformation that takes e_i to v_i/a_i. Then $gHg^{-1} = H_\ell$; therefore H_ℓ is as asserted in Proposition 1.1 □

Remark. From Lemma 1.2 we see that the collection of Shimura curves in X is parametrized by the Γ-equivalence classes of primitive positive vectors in $\mathfrak{O}^3$, that is, primitive vectors $v \in \mathfrak{O}^3$ with $h(v) > 0$. The collection of these equivalence classes is commensurable with $\mathrm{SU}(h, F)\backslash\mathbb{P}(F^3)^+$, where $\mathbb{P}(F^3)^+$ denotes the space of h-positive lines in F^3. The class of $h(v)$ gives a well-defined function

$h : \mathbb{P}(F^3) \to F_0^*/N_{F/F_0}(F^*)$, the *norm residue group*. It can be checked that the class of $h(v)$ is a commensurability invariant and that it takes on infinitely many values; hence we get an infinite number of commensurability classes of subgroups of $\mathrm{SU}(1, 1)$. Observe that the matrix of the conjugating element g of Lemma 1.2 has entries in the finite field extension $F(a_1, a_2, a_3)$ of F.

The compact factors of G, necessary for the $\mathbb{Q}$-structure in the definition of a Shimura surface, play no role in the sequel. We thus simplify notation and write G for G_0 and H for H_0 from now on.

Elliptic elements and cusps. The bounded negativity conjecture (BNC) originally is a question for smooth compact (projective) surfaces. If Γ is cocompact and torsion-free, Shimura surfaces as defined above fall into the scope of this conjecture and the results in the introduction need no explanation.

Any arithmetic lattice contains a neat normal subgroup of finite index. Such subgroups are in particular torsion-free. As quotients by a finite group, the Shimura surfaces come with a ($\mathbb{Q}$-valued) intersection theory. The BNC can be extended to such surfaces, and Theorem 0.1 needs no further explanation.

If Γ is cofinite but not cocompact, our proof of Theorem 0.1 gives a statement about the self-intersection number of the cohomology class of the Shimura curve projected to the complement of the cusp resolution cycles, as we will now explain.

We may suppose that Γ is a neat subgroup. Let X^{BB} be the minimal (Baily–Borel) compactification of $X = \Gamma\backslash G/K$. Since X is not isogenous to a product, $X^{\mathrm{BB}} \setminus X$ has codimension two, and hence $H_c^2(X, \mathbb{Q}) \cong H^2(X^{\mathrm{BB}}, \mathbb{Q})$. Let $\pi : Y \to X^{\mathrm{BB}}$ be a (minimal) smooth resolution of the singularities at the cusps and $j : X \to Y$ the inclusion. We claim that

$$H^2(Y, \mathbb{Q}) = \pi^* H^2(X^{\mathrm{BB}}, \mathbb{Q}) \oplus B, \tag{1}$$

where B is the subspace spanned by cusp resolution curves. Moreover, the direct sum is orthogonal and the intersection form on B is negative-definite. This implies that the sum decomposition is compatible with Poincaré duality, and this will make the arguments in Section 3 work in the noncompact case, too; see Theorem 3.6.

Our claims are stated for the Hilbert modular case in [van der Geer 1988, Sections II.3, VI.1]. In the case of a ball quotient, a neighborhood W of the cusps in Y is a disjoint union of disc bundles over tori, each sitting inside a line bundle of negative degree. It suffices to show that

$$H_2(Y, \mathbb{Q}) = H_2(W, \mathbb{Q}) \oplus \mathrm{Im}(j_* : H_2(X, \mathbb{Q}) \to H_2(Y, \mathbb{Q}))$$

and then apply duality. By Mayer–Vietoris, it suffices to show that $H_1(W \cap X, \mathbb{Q}) \to H_1(W, \mathbb{Q}) \oplus H_1(X, \mathbb{Q})$ is injective. This holds true, since the inclusion of a circle bundle into the corresponding disc bundle induces an injection the level of $H_1(\,\cdot\,, \mathbb{Q})$.

We remark that the BNC (and intersection numbers in general) are very sensitive to blowups. We leave it to the reader to investigate if Theorem 0.1 also holds on Y.

Volume normalization. The hermitian symmetric space G/K comes with a Kähler $(1,1)$-form ω that we normalize, say, so that the minimum value of the curvature of the associated Riemannian metric is -1. We continue assuming that Γ is a neat subgroup, so that X is a manifold with universal cover $\widetilde{X} = G/K$. Then $\omega \wedge \omega$ provides volume forms on $\widetilde{X}$ and X. We let $\mathrm{vol}(X)$ be the volume of the Shimura surface. Rescaling by the volume, we obtain a probability measure ν_X on X induced from the volume form.

Shimura curves are totally geodesic subvarieties in X. Consequently, the restriction of ω is a Kähler form ω_C on C. We let $\mathrm{vol}(C) = \int_C \omega_C$ be the corresponding volume and ν_C the probability measure defined by ω_C.

We need to extend this to the quotients by smaller compact subgroups. Let $K' \subset G$ be a compact subgroup and $K'_H = K' \cap H$. Let ν_G be the Haar measure on G normalized so that the pushforward to G/K gives the above volume form on $\widetilde{X}$ and such that the fibers have volume 1. From ν_G, we obtain measures $\nu_{G/K'}$ on G/K' and finite measures $\nu_{\Gamma\backslash G/K'}$ on $X_{K'} = \Gamma\backslash G/K'$ with $\mathrm{vol}(X) = \mathrm{vol}(X_{K'})$.

Similarly we fix a normalization of a Haar measure ν_H on H by requiring that the fibers of $H \to H/K_H$ have volume 1 and that the pushforward to H/K_H is the volume form coming from the metric with curvature -1, as above.

In this way, given a Shimura curve $C = \Gamma\backslash\Gamma g H/K_H$, the pushforward of ν_H defines a finite measure $\nu_{C,K'}$ on the locally symmetric subspaces $C_{K'} = \Gamma\backslash\Gamma g H/K'_H$ inside $X_{K'}$ with $\mathrm{vol}(C_{K'}) = \mathrm{vol}(C)$.

2. Equidistribution

There are many sources in the literature that deduce equidistribution for Shimura curves from a Ratner-type theorem (notably [Clozel and Ullmo 2005; Ullmo 2007]). We need a slightly stronger equidistribution result, on $\Gamma\backslash G$ or on $\Gamma\backslash G/K'$ for some (not necessarily maximal) compact subgroup K' of G rather than on the algebraic surface X. This follows along known lines from Ratner's result, or rather the version in [Eskin et al. 1996]. We give a proof avoiding technicalities on Shimura data and focusing on the surface case.

The references above contain as special case the following equidistribution:

Proposition 2.1. *Suppose that X is a Shimura surface. If $(C_n)_{n\in\mathbb{N}}$ is a sequence of pairwise different Shimura curves, then $\nu_{C_n} \to \nu_X$ weakly as $n \to \infty$.*

This is a special case of the following stronger result:

Proposition 2.2. *Suppose that $X = \Gamma\backslash G/K$ is a Shimura surface. Let $K' \subset K$ be a closed subgroup, and let $g_n \in G$ be a sequence of points so that the orbits $g_n H \subset G$ project to pairwise-distinct Shimura curves C_n in X. Then on $X' = \Gamma\backslash G/K'$ the sequence of probability measures $\nu_{C_n,K'}$ converges weakly to $\nu_{\Gamma\backslash G/K'}$ as $n \to \infty$.*

Corollary 2.3. *Suppose that $X = \Gamma\backslash G/K$ is a Shimura surface. If $(C_n)_{n\in\mathbb{N}}$ is a sequence of pairwise different Shimura curves, then* $\mathrm{vol}(C_n) \to \infty$ *as $n \to \infty$.*

Proof of Corollary 2.3. With the above volume normalization, it suffices to prove the claim for the lifts of the Shimura curves C_n to $X'' = \Gamma\backslash G$. Let C_n'' denote these lifts. We apply the preceding proposition for $K' = \{e\}$. Equidistribution implies in particular that Shimura curves are dense; i.e., for any finite collection of open sets U_i, $i \in I$, there exists N_0 such that for $n > N_0$ the intersection $C_n'' \cap U_i$ is nonempty for all i. Since X'' is foliated by H-orbits and ν_G is locally the product of ν_H and a transversal measure, it suffices to take for U_i sufficiently many open sets locally trivializing the foliation, namely $U_i = V_i \times W_i$ with V_i an H-orbit, such that $\nu_H(V_i) = O(1)$ but the transversal measure of W_i is $O(1/n^2)$. Then we can fit $O(n)$ such sets into X, and each time C_n'' intersects some U_i it picks up a volume of $O(1)$. □

Proof of Proposition 2.2. We first observe that, if the proposition holds for $K' = \{e\}$, then it holds for any other $K' \subset K$. Namely, under the projection $\pi : X'' = \Gamma\backslash G \to X' = \Gamma\backslash G/K'$, we have, by the volume normalization above, that the pushforward measures satisfy $\pi_*(\nu_{X''}) = \nu_{X'}$ and $\pi_*(\nu_{C_n,e}) = \nu_{C_n,K'}$. Thus we will assume $K' = \{e\}$. For this choice of K' we have that $X' = \Gamma\backslash G$. Thus we'll simply write X' for $\Gamma\backslash G$ and ν_n' for $\nu_{C_n,e}$.

The proof consists of two parts: (1) prove that ν_n' has convergent subsequences ν_{n_j}'; (2) prove that the limit of any convergent subsequence must be $\nu_{X'}$.

If Γ is cocompact, that is, X' is compact, then the space of probability measures on X' is compact in the weak-* topology, so ν_n' has a convergent subsequence. If X is not compact, then a subsequence converges to a measure on the one-point compactification $X' \cup \{\infty\}$, but these measures may "escape to infinity", e.g., converge to the delta function at ∞. An example of this "escape of mass" is given in the introduction to [Eskin et al. 1997]. The main result there is that there is no escape of mass when the image of $Z(H)$ in X' is compact (where $Z(H)$ is the centralizer of H in G). More precisely, compactness of the image of $Z(H)$ in X' implies (see [Eskin et al. 1997, Theorem 1.1]) that for every $\varepsilon > 0$ there exists a compact subset $W \subset \Gamma\backslash G$ such that every H-orbit gives measure at least $1-\varepsilon$ to W. Hence the sequence ν_n' indeed converges in the space of probability measures on X'.

In our situation $Z(H)$ itself is compact: it is finite in Case One and $U(1)$ in Case Two, and thus we always have convergence, thereby proving (1). (Compactness of $Z(H)$ generally holds for Shimura varieties if one discards the obvious exception of product situations; see [Ullmo 2007].)

To prove (2) we may assume ν_n' converges weakly to a probability measure ν'; we must prove $\nu' = \nu_{X'}$. This follows a pattern which is by now standard: (i) use, as in [Eskin et al. 1996], Ratner's theorem on unipotent flows to prove that ν is *algebraic*, i.e., supported on an L-orbit of some connected algebraic group $H \subseteq L \subseteq G$ that intersects Γ in a lattice; (ii) prove $L = G$. We formulate (i) as the following lemma:

Lemma 2.4. *Suppose ν_n' converges weakly to ν'. Then there exists a closed connected subgroup L, $H \subset L \subset G$, such that ν' is an L-invariant measure supported on $\Gamma\backslash\Gamma cL$ for some $c \in G$ and such that $c^{-1}\Gamma c \cap L$ is a lattice in L. Moreover, there exist a sequence $x_n \in \Gamma g_n H$ converging to c and an n_0 such that cLc^{-1} contains the subgroup generated by $x_n H x_n^{-1}$ for $n \geq n_0$.*

We formulated this lemma following closely the wording of [Eskin and Oh 2006, Proposition 2.1] (see also [Eskin et al. 1996, Theorem 1.7]) because it can be proved from [Mozes and Shah 1995, Theorem 1.1] in same way. Namely, start from the fact that ν_n' is supported on the H-orbit $\Gamma\backslash\Gamma g_n H \subset \Gamma\backslash G$, which is isomorphic to $(g_n^{-1}\Gamma g_n \cap H)\backslash H$ and is H-invariant. Since $g_n^{-1}\Gamma g_n$ is a lattice in H, which, in our case, is locally isomorphic to $\mathrm{SL}(2, \mathbb{R})$, we can choose a unipotent one-parameter subgroup $u(t)$ in H and apply the Moore ergodicity theorem, as in the proof of [Eskin and Oh 2006, Proposition 2.1], to show that ν_n' is an ergodic $u(t)$-invariant measure, thus checking that the first hypothesis of [Mozes and Shah 1995, Theorem 1.1] is satisfied. We continue, in this way, following the proof of [Eskin and Oh 2006, Proposition 2.1] until the proof of Lemma 2.4 is complete.

Finally the groups $x_n H x_n^{-1}$ cannot all be equal to H, since this would give $\gamma_n \in \Gamma$ so that $g_n H g_n^{-1} = \gamma_n H \gamma_n^{-1}$, contradicting the hypothesis that the curves C_n are pairwise different. We conclude that $H \subsetneq L$ and thus $L = G$ by the following lemma. $\square$

Lemma 2.5. *Let (G, H) be as in Case One or Case Two. If L is a connected real Lie group with $H \subsetneq L \subset G$ and $\Gamma \cap L$ is a lattice in L, then $L = G$.*

Proof. This is easily verified on the level of Lie algebras. Since $\mathrm{Lie}(L)$ contains an element not in $\mathrm{Lie}(H)$, bracketing with suitable elements of $\mathrm{Lie}(H)$ allows one to produce a generating set of $\mathrm{Lie}(G)$. $\square$

3. The current of integration of a Shimura curve

Any Shimura curve C, in fact any codimension-one subvariety of the Shimura surface X, defines a closed $(1, 1)$-current on X. On the other hand, the Shimura surface comes with a natural $(1, 1)$-form, the Kähler form ω. The aim of this section is to translate the equidistribution result (a convergence of measures) into a convergence statement for the classes of these currents, suitably normalized. We start with the compact case and explain at the end of this section the necessary modification in the noncompact case. Recall that a $(1, 1)$-current on a complex surface X is a continuous linear functional on $A_c^{1,1}(X)$, the space of compactly supported (1, 1)-forms on X. This space $(A_c^{1,1}(X))^\vee$ contains both the complex curves $C \subset X$ and the smooth forms $\eta \in A^{1,1}(X)$ by the formulas

$$C \to \left(\alpha \to \int_C \alpha\right), \quad \eta \to \left(\alpha \to \int_X \eta \wedge \alpha\right) \quad \text{for all } \alpha \in A_c^{1,1}(X).$$

The cohomology of X can be computed either from the complex of forms or from the complex of currents. Recall also that, if X is Kähler and ω denotes the Kähler form, then $\mathrm{vol}(X) = \int_X \omega \wedge \omega$, the Kähler form of C is $\omega_C = \omega_X|_C$ and $\mathrm{vol}(C) = \int_C \omega_C$.

Proposition 3.1. *Let $X = \Gamma\backslash G/K$ be a smooth Shimura surface and let $g_n \in G$ be any sequence of points such that the Shimura curves $C_n = \Gamma\backslash\Gamma g_n H/K$ are pairwise distinct. Then*

$$C_n/\,\mathrm{vol}(C_n) \to \omega \quad \textit{in } A_c^{1,1}(X)^\vee, \quad \textit{hence in } H^{1,1}(X).$$

This and the finite-dimensionality of the Picard group allows us to deduce our main result.

Corollary 3.2. *Let $X = \Gamma\backslash G/K$ be a compact, smooth Shimura surface and let $g_n \in G$ be any sequence of points such that the Shimura curves $C_n = \Gamma\backslash g_m H/K$ are pairwise distinct. Then*

$$C_n^2 \sim \mathrm{vol}(X)\,\mathrm{vol}(\Gamma\backslash\Gamma g_n H)^2 \quad \textit{for } n \to \infty.$$

In particular, for any M, there are only finitely many Shimura curves C on X with $C^2 < M$.

Proof. For the first statement, fix a basis $\gamma_0 = \omega, \gamma_1, \ldots, \gamma_s$ of $H^{1,1}(X)$. Taking γ_i for $i \geq 1$ orthogonal to γ_0, we may suppose that the dual basis is $\lambda^{-1}\omega = \gamma_0^\vee, \gamma_1^\vee, \ldots, \gamma_s^\vee$ for some $\lambda \in \mathbb{C}$; in fact, $\lambda = \int_X \omega \wedge \omega = \mathrm{vol}(X)$. If C is a curve in X, thus representing a (1, 1)-class, the Poincaré dual is represented by

$$\mathrm{PD}(C) = \sum_{i=0}^{s} \left(\int_C \gamma_i\right) \gamma_i^\vee.$$

Consequently, letting $A_n = \mathrm{vol}(C_n)$, by Proposition 3.1,

$$\begin{aligned}\frac{1}{A_n^2} C_n \cdot C_n = \frac{1}{A_n^2} \int_{C_n} \mathrm{PD}(C_n) &= \sum_{i=0}^{s} \left(\frac{1}{A_n}\int_{C_n} \gamma_i\right)\left(\frac{1}{A_n}\int_{C_n} \gamma_i^\vee\right) \\ &\longrightarrow \sum_{i=0}^{s} \left(\int_X \omega \wedge \gamma_i\right)\left(\int_X \omega \wedge \gamma_i^\vee\right) = \lambda = \mathrm{vol}(X). \quad (2)\end{aligned}$$

The second statement follows from the first and from Corollary 2.3. □

Integrating on the projectivized tangent bundle. We now prepare for the proof of Proposition 3.1. For this purpose we work on the universal cover $\widetilde{X} = G/K$ of X. First of all, for any (two-dimensional) Kähler manifold X there is a natural map

$$\mathbb{P}T\widetilde{X} \to \Lambda_{1,1}T\widetilde{X} = (\Lambda^{1,1}T^*\widetilde{X})^\vee,$$

defined pointwise at any $x \in \widetilde{X}$ by $[v] \mapsto v \wedge \bar{v}/|v|^2$ for $v \in T_x\widetilde{X} \setminus \{0\}$. Dually, an element $\alpha \in (\Lambda^{1,1}T^*\widetilde{X})$ defines a real-valued function

$$\varphi_\alpha : \mathbb{P}T\widetilde{X} \to \mathbb{R}, \quad \varphi_\alpha([v]) = \alpha\left(\frac{v \wedge \bar{v}}{|v|^2}\right).$$

Using this map we can write the intersection with α as the integral of a real-valued function against the volume form of $\mathbb{P}TX$. In Case Two, $\mathbb{P}T\widetilde{X} = G/K'$ is a homogeneous space with an invariant volume, where $K' = \mathrm{U}(1) \times \mathrm{U}(1)$. In Case One, we will need to pass to a G-invariant real subbundle of $\mathbb{P}T\widetilde{X}$, also of the form G/K' for $K' = \mathrm{U}(1)$.

We start with Case Two. Recall that we scaled the Kähler form ω so that $\mathrm{vol}(X) = \int_X \omega \wedge \omega$.

Lemma 3.3. *Let X be a two-dimensional Kähler manifold, choose a two-form η on $\mathbb{P}TX$ that restricts to the area form η_x of each fiber $\mathbb{P}T_xX$, $x \in X$, scaled to give total area 1 to each fiber. Then, for all $(1,1)$-forms α on X and for each $x \in X$, we have*

$$(\omega \wedge \alpha)_x = \left(\int_{\mathbb{P}T_xX} \varphi_\alpha \eta_x\right)(\omega \wedge \omega)_x.$$

Therefore we have

$$\int_X \omega \wedge \alpha = \int_{\mathbb{P}TX} \varphi_\alpha\, \eta \wedge \omega \wedge \omega,$$

where we have written simply ω for the pullback to $\mathbb{P}TX$ of the form ω on X.

Proof. In suitable local coordinates at x, the Kähler form at x is

$$\omega_x = \frac{\sqrt{-1}}{2}(dz_1 \wedge d\bar{z}_1 + dz_2 \wedge d\bar{z}_2).$$

Writing $\alpha = \frac{\sqrt{-1}}{2}\sum \alpha_{i\bar{j}} dz_i \wedge d\bar{z}_j$, we have (suppressing the factors of $\frac{\sqrt{-1}}{2}$)

$$(\omega \wedge \alpha)_x = (\alpha_{1\bar{1}} + \alpha_{2\bar{2}})(dz_1 \wedge d\bar{z}_1 \wedge dz_2 \wedge d\bar{z}_2) = \frac{\alpha_{1\bar{1}} + \alpha_{2\bar{2}}}{2}(\omega \wedge \omega)_x.$$

On the other hand, if we let e_1, e_2 denote the basis for T_xX dual to dz_1, dz_2, and write $v = v_1e_1 + v_2e_2 \in T_xX$, the first factor of the right-hand side is

$$\int_{\mathbb{P}^1} \alpha\left(\frac{(v_1e_1+v_2e_2)\wedge\overline{(v_1e_1+v_2e_2)}}{|v_1|^2+|v_2|^2}\right)\eta_x$$
$$= \alpha_{1\bar{1}}\int_{\mathbb{P}^1}\frac{|v_1|^2}{|v_1|^2+|v_2|^2}\,\eta_x+\alpha_{2\bar{2}}\int_{\mathbb{P}^1}\frac{|v_2|^2}{|v_1|^2+|v_2|^2}\,\eta_x+\int_{\mathbb{P}^1}\frac{2\,\mathrm{Im}(\alpha_{1\bar{2}}v_1\bar{v}_2)}{|v_1|^2+|v_2|^2}\,\eta_x. \quad (3)$$

The involution $(v_1 : v_2) \to (v_2 : v_1)$ is an isometry of $\mathbb{P}^1$. The last integrand is sent to its negative by this involution, so the last integral vanishes. The first two integrals are interchanged by this involution, therefore they are equal. Since the sum of the two integrands is visibly identically 1, each of the first two integrals has value $\frac{1}{2}$. Thus the first two terms give $\frac{1}{2}\alpha_{1\bar{1}}$ and $\frac{1}{2}\alpha_{2\bar{2}}$ respectively, hence the first statement of the lemma follows. The second follows from the first and Fubini's theorem. □

Remark. The first statement in the lemma is equivalent to the well-known fact in linear algebra that the trace of a hermitian matrix equals the average value over the unit sphere of the associated hermitian form.

Corollary 3.4. *If X is a Shimura surface covered by the ball, then for all $(1,1)$-forms α on X we have*

$$\int_X \omega\wedge\alpha = \int_{\mathbb{P}TX}\varphi_\alpha\, d\nu_{\Gamma\backslash G/K'},$$

where $\nu_{\Gamma\backslash G/K'}$ is the volume form on $\mathbb{P}TX$ introduced above.

Proof. If $\widetilde{X} = \mathbb{B}^2 = G/K$, then $\eta\wedge\omega\wedge\omega$ in Lemma 3.3 is a G-invariant volume form on $\mathbb{P}T\widetilde{X}$. Moreover, ω and η have been scaled to give the correct normalization. □

Now we address the corresponding statement in Case One. If the Shimura surface X is covered by $\mathbb{H}^2$, then $\mathbb{P}T\widetilde{X}$ is no longer a homogeneous space for G, but it has some natural homogeneous subbundles. Equivalently, the action of K on $\mathbb{P}T_x\widetilde{X} \cong \mathbb{P}^1$ is not transitive, but has some distinguished orbits: two zero-dimensional orbits, corresponding to the tangents to the two factors of $\mathbb{H}^2$, and an orbit of real dimension 1 corresponding to the graphs of isometries between the two factors. Explicitly, if we choose coordinates z_1, z_2 as above, this time adapted to the product structure of $\widetilde{X}$, and with dual basis e_1, e_2 each tangent to one of the factors, and writing $v = v_1e_1 + v_2e_2$ as above, the action of $K \cong \mathrm{U}(1)\times\mathrm{U}(1)$ on $\mathbb{P}T_x\widetilde{X} \cong \mathbb{P}^1$ leaves invariant the points with homogeneous coordinates $(1:0)$ and $(0:1)$ and the real submanifold $\{(v_1 : v_2) : |v_1| = |v_2|\} = \{(1 : e^{i\theta})\} \cong S^1$.

Let us call this submanifold $\mathbb{S}T_x\widetilde{X}$ and let $\mathbb{S}T\widetilde{X} \cong G/K'$ denote the corresponding bundle over $\widetilde{X} \cong G/K$ with fiber $K/K' \cong \mathbb{S}T_x\widetilde{X} \cong S^1$. Then a calculation just as in the proof of Lemma 3.3 gives us:

Lemma 3.5. *Let X be a Shimura surface covered by $\mathbb{H}^2$, choose a one-form η on $\mathbb{S}TX$ that restricts to the angle form $\eta_x = d\theta$ of each fiber $\mathbb{S}T_xX$, scaled to give total area 1 to each fiber. Then, for any $(1,1)$ form α on X and for each $x \in X$, we have*

$$(\omega \wedge \alpha)_x = \left(\int_{\mathbb{S}T_xX} \varphi_\alpha \eta_x\right)(\omega \wedge \omega)_x.$$

Therefore we have

$$\int_X \omega \wedge \alpha = \int_{\mathbb{S}TX} \varphi_\alpha\, \eta \wedge \omega \wedge \omega = \int_{\mathbb{S}TX} \varphi_\alpha\, d\nu_{\Gamma\backslash G/K'},$$

where $\nu_{\Gamma\backslash G/K'}$ is the volume form on $\mathbb{S}TX$ introduced above.

Proof of Proposition 3.1. To show convergence in $H^{1,1}(X)$ it suffices to show that

$$\frac{1}{\mathrm{vol}(C_n)} \int_{C_n} \alpha \to \int_X \omega \wedge \alpha$$

for any $\alpha \in H^{1,1}(X)$. In Case Two, by Corollary 3.4 it suffices to show that

$$\frac{1}{\mathrm{vol}(C_n)} \int_{C_n} \alpha \to \int_{\mathbb{P}TX} \varphi_\alpha\, d\nu_{\Gamma\backslash G/K'}.$$

A local verification, just using the definition of φ_α and the fact that $\nu_{C_n,K'}$ was defined to give measure 1 to the fibers K/K', implies that $\int_{C_n} \alpha = \int_{\mathbb{P}TC_n} \varphi_\alpha\, d\nu_{C_n,K'}$. Since $\nu_{C_n,K'}$ is supported on $\mathbb{P}TC_n \subset \mathbb{P}TX$, it is thus sufficient to show that

$$\int_{\mathbb{P}TX} \varphi_\alpha\, d\nu_{C_n,K'} \to \int_{\mathbb{P}TX} \varphi_\alpha\, d\nu_{\Gamma\backslash G/K'}.$$

We have reformulated our claim in terms of a convergence of measures, integrating against a globally defined function φ_α. Proposition 2.2 completes the proof. In Case One, the proof is the same, replacing $\mathbb{P}TX$ by $\mathbb{S}TX$ and the reference to Corollary 3.4 by Lemma 3.5. □

The noncompact case. Recall that we denoted by Y a minimal resolution of the singularities of the Baily–Borel compactification X^{BB}. By [Mumford 1977, Theorem 3.1, Proposition 1.1], the Kähler class ω extends to a closed current on Y. Moreover, $\omega \in \pi^* H^2(X^{\mathrm{BB}}, \mathbb{Q})$ by [Mumford 1977, Proposition 3.4(b)]. The statement of Proposition 3.1 now reads

$$p_{B^\perp}(C_n)/\,\mathrm{vol}(C_n) \to \omega \quad \text{in } \pi^* H^2(X^{\mathrm{BB}}, \mathbb{Q}),$$

where $p_B^\perp$ is the orthogonal projection onto the complement of B. The same proof as above works. In order to show the analog

$$(p_{B^\perp} C_n)^2 \sim \mathrm{vol}(\Gamma\backslash\Gamma g_n H)^2 \quad \text{for } n \to \infty$$

of Corollary 3.2, we apply Poincaré duality to $\pi^* H^2(X^{\mathrm{BB}}, \mathbb{Q})$. Since this is a perfect pairing, the proof of Corollary 3.2 applies without changes:

Theorem 3.6. *For X as above and for any real number M, there are only finitely many Shimura curves C on X with $(p_{B^\perp} C)^2 < M$.*

In particular, for the collection of compact Shimura curves in X, we obtain Theorem 0.1.

Acknowledgements

This note is the result of the Mathematisches Forschungsinstitut Oberwolfach mini-workshop "the bounded negativity conjecture". We thank the organizers of this workshop and the MFO administration for organizing the simultaneous mini-workshop "Kähler groups", which brought the two authors together. We thank Alex Eskin for discussing details on equidistribution results, and Madhav Nori and Matthew Stover for discussions on algebraic group details. We thank the referee for a thorough report and helpful suggestions. Toledo thanks Misha Gromov for a conversation in 1990 on the idea that Shimura curves approximate the Kähler class.

The paper [Koziarz and Maubon 2014], with some overlap with ours, also resulted from the MFO workshop.

References

[Bauer et al. 2013] T. Bauer, B. Harbourne, A. L. Knutsen, A. Küronya, S. Müller-Stach, X. Roulleau, and T. Szemberg, "Negative curves on algebraic surfaces", *Duke Math. J.* **162**:10 (2013), 1877–1894. MR 3079262 Zbl 1272.14009

[Clozel and Ullmo 2005] L. Clozel and E. Ullmo, "Équidistribution de sous-variétés spéciales", *Ann. of Math.* (2) **161**:3 (2005), 1571–1588. MR 2006j:11083 Zbl 1099.11031

[Eskin and Oh 2006] A. Eskin and H. Oh, "Ergodic theoretic proof of equidistribution of Hecke points", *Ergodic Theory Dynam. Systems* **26**:1 (2006), 163–167. MR 2006j:11068 Zbl 1092.11023

[Eskin et al. 1996] A. Eskin, S. Mozes, and N. Shah, "Unipotent flows and counting lattice points on homogeneous varieties", *Ann. of Math.* (2) **143**:2 (1996), 253–299. MR 97d:22012 Zbl 0852.11054

[Eskin et al. 1997] A. Eskin, S. Mozes, and N. Shah, "Non-divergence of translates of certain algebraic measures", *Geom. Funct. Anal.* **7**:1 (1997), 48–80. MR 98d:22006 Zbl 0872.22009

[Garibaldi and Gille 2009] S. Garibaldi and P. Gille, "Algebraic groups with few subgroups", *J. Lond. Math. Soc.* (2) **80**:2 (2009), 405–430. MR 2010i:20050 Zbl 1183.20053

[van der Geer 1988] G. van der Geer, *Hilbert modular surfaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) **16**, Springer, Berlin, 1988. MR 89c:11073 Zbl 0634.14022

[Granath 2002] H. Granath, *On quaternionic Shimura surfaces*, Ph.D. thesis, Chalmers Tekniska Hogskola, 2002, Available at http://search.proquest.com/docview/305513982. MR 2715512

[Hirzebruch and Zagier 1976] F. Hirzebruch and D. Zagier, "Intersection numbers of curves on Hilbert modular surfaces and modular forms of Nebentypus", *Invent. Math.* **36** (1976), 57–113. MR 56 #11909 Zbl 0332.14009

[Koziarz and Maubon 2014] V. Koziarz and J. Maubon, "On the equistribution of totally geodesic submanifolds in compact locally symmetric spaces and application to boundedness results for negative curves and exceptional divisors", preprint, 2014. arXiv 1407.6561v2

[Kudla 1978] S. S. Kudla, "Intersection numbers for quotients of the complex 2-ball and Hilbert modular forms", *Invent. Math.* **47**:2 (1978), 189–208. MR 80a:10044 Zbl 0399.10030

[Mozes and Shah 1995] S. Mozes and N. Shah, "On the space of ergodic invariant measures of unipotent flows", *Ergodic Theory Dynam. Systems* **15**:1 (1995), 149–159. MR 95k:58096 Zbl 0818.58028

[Mumford 1977] D. Mumford, "Hirzebruch's proportionality theorem in the noncompact case", *Invent. Math.* **42** (1977), 239–272. MR 81a:32026 Zbl 0365.14012

[Platonov and Rapinchuk 1994] V. Platonov and A. Rapinchuk, *Algebraic groups and number theory*, Pure and Applied Mathematics **139**, Academic Press, Boston, 1994. MR 95b:11039 Zbl 0841.20046

[Serre 1994] J.-P. Serre, *Cohomologie galoisienne*, 5th ed., Lecture Notes in Mathematics **5**, Springer, Berlin, 1994. MR 96b:12010 Zbl 0812.12002

[Tits 1966] J. Tits, "Classification of algebraic semisimple groups", pp. 33–62 in *Algebraic Groups and Discontinuous Subgroups* (Boulder, CO 1965), Proc. Sympos. Pure Math. **9**, Amer. Math. Soc., Providence, R.I., 1966. MR 37 #309 Zbl 0238.20052

[Ullmo 2007] E. Ullmo, "Equidistribution de sous-variétés spéciales, II", *J. Reine Angew. Math.* **606** (2007), 193–216. MR 2008g:11103 Zbl 1137.11043

[Vignéras 1980] M.-F. Vignéras, *Arithmétique des algèbres de quaternions*, Lecture Notes in Mathematics **800**, Springer, Berlin, 1980. MR 82i:12016 Zbl 0422.12008

Communicated by Hélène Esnault
Received 2014-08-16 Revised 2015-03-02 Accepted 2015-04-07

moeller@math.uni-frankfurt.de *Institut für Mathematik, Goethe-Universität Frankfurt, Robert-Mayer-Strasse 6–8, D-60325 Frankfurt am Main, Germany*

toledo@math.utah.edu *Department of Mathematics, University of Utah, 155 S 1400 E, Salt Lake City, UT 84112, United States*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.913

Singularities of locally acyclic cluster algebras

Angélica Benito, Greg Muller, Jenna Rajchgot and Karen E. Smith

We show that locally acyclic cluster algebras have (at worst) canonical singularities. In fact, we prove that locally acyclic cluster algebras of positive characteristic are strongly F-regular. In addition, we show that upper cluster algebras are always Frobenius split by a canonically defined splitting, and that they have a free canonical module of rank one. We also give examples to show that not all upper cluster algebras are F-regular if the local acyclicity is dropped.

1. Introduction

Fomin and Zelevinsky introduced cluster algebras at the close of the twentieth century as a way to study total positivity in a wide range of contexts. Since then, cluster algebra structures have been discovered in many unexpected corners of mathematics (and physics), including Teichmüller theory [Gekhtman et al. 2005; Fock and Goncharov 2007], discrete integrable systems [Fomin and Zelevinsky 2003], knot theory [Muller 2012; Musiker et al. 2013], and mirror symmetry [Shen and Goncharov 2015; Kontsevich and Soibelman 2013], just to name a few.

Locally acyclic cluster algebras, introduced in [Muller 2013], are a large class of cluster algebras which are simultaneously flexible enough to include many interesting examples — including many fundamental examples from representation theory and most examples from Teichmüller theory — yet restrictive enough to avoid the pathological behavior sometimes found in general cluster algebras. For example, locally acyclic cluster algebras are finitely generated and normal, while a general cluster algebra may fail to be either. The main theorem of this paper is that *locally acyclic cluster algebras have (at worst) canonical singularities.* In fact, we show that locally acyclic cluster algebras of prime characteristic are *strongly F-regular,* a strong form of Frobenius split which implies many nice restrictions on the singularities; for example, F-regular varieties are normal, Cohen–Macaulay,

Smith acknowledges the financial support of NSF grant DMS-1001764 and the Clay Foundation. Benito is partially support by MTM2012-35849.

MSC2010: primary 13F60; secondary 13A35, 14B05.

Keywords: cluster algebras, locally acyclic cluster algebras, singularities, F-regularity, Frobenius splitting.

pseudorational, and have Kawamata log terminal singularities (if the canonical class divisor is $\mathbb{Q}$-Cartier) or canonical singularities (if the canonical class is Cartier). These characteristic-p results imply the corresponding statements in characteristic zero as well. For a survey, see, e.g., [Smith 1997b] or [Smith and Zhang 2015].

Associated to a cluster algebra $\mathcal{A}$ is its *upper cluster algebra* $\mathcal{U}$. This related algebra has the same fraction field and satisfies $\mathcal{A} \subseteq \mathcal{U}$ (see [Berenstein et al. 2005]). We show that all upper cluster algebras in positive characteristic have a "cluster" Frobenius splitting, which can be expressed explicitly in terms of any cluster. We also prove the closely related result that upper cluster algebras have a free canonical module, which is generated by any log volume form in a cluster of cluster variables. The latter of these results is found in the Appendix.

The inclusion $\mathcal{A} \subseteq \mathcal{U}$ need not be equality, though it is in the case when $\mathcal{A}$ is locally acyclic [Muller 2014]. When equality fails, a general philosophy is that $\mathcal{U}$ is better behaved than $\mathcal{A}$. In this direction, we show that if an upper cluster algebra $\mathcal{U}$ fails to be F-regular, then $\mathcal{A}$ also fails to be F-regular, and we provide an example of this situation. Taking the ground field to be of characteristic zero, this gives an example of a finitely generated upper cluster algebra $\mathcal{U}$ which has *log canonical but not log terminal singularities.* We also provide an example where $\mathcal{A} \neq \mathcal{U}$ and $\mathcal{A}$ is pathological (e.g., $\mathcal{A}$ is non-Noetherian), but $\mathcal{U}$ is nevertheless strongly F-regular.

All of our results and arguments are also valid for cluster algebras given by an arbitrary skew-symmetrizable matrix. However, we have written the exposition in the slightly less general setting of cluster algebras given by quivers for the sake of accessibility. Experts will have no trouble adapting the arguments to the more general setting.

2. Cluster algebras

A cluster algebra is a commutative domain with some extra combinatorial structure. It comes equipped with a (usually infinite) set of generators called *cluster variables*, which can be recursively generated from a *seed*: a quiver decorated with a free generating set for a field.

We will consider cluster algebras over an arbitrary field k, although in the literature they are usually defined over $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{Z}$. The choice of scalars is mostly irrelevant to the definitions, and most proofs of standard results go through without change. As such, we will cite the original results without comment, and only address the differences as needed.

Seeds and mutations. Let k be a field, and let $\mathcal{F}$ be a purely transcendental finite extension of k. A *seed* for $\mathcal{F}$ over k consists of the following data:

- A *quiver* Q without loops or directed 2-cycles.

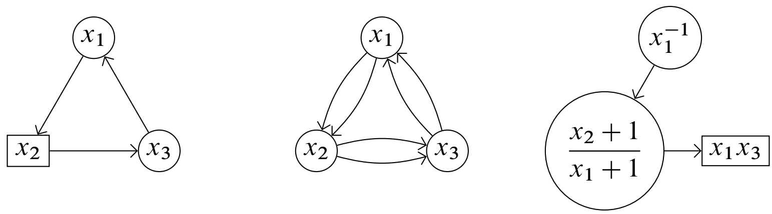


Figure 1. Three examples of seeds in $\mathcal{F} = \mathsf{k}(x_1, x_2, x_3)$.

- A bijection from the vertices of Q to a set of algebraically independent generators $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ for $\mathcal{F}$ over k. The image x_i of a vertex i is called the *cluster variable* at that vertex, and the set $\boldsymbol{x}$ is called a *cluster*.
- A subset of the vertices of Q designated as *frozen*; the rest are called *mutable*. We impose the nonstandard convention that every vertex which touches no arrow is frozen.[1]

Seeds will usually be denoted as a pair $(\mathsf{Q}, \boldsymbol{x})$, with the other data suppressed. The number of vertices of Q (denoted n hereafter) is the *rank* of the seed, and the number of mutable vertices (denoted m hereafter) is the *mutable rank*.

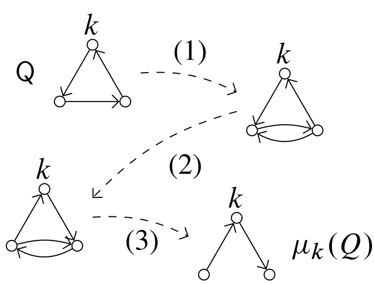
Seeds may be drawn as a quiver with circles ○ for mutable vertices, and rectangles □ for frozen vertices, each with the corresponding cluster variable inscribed (e.g., Figure 1).

A seed $(\mathsf{Q}, \boldsymbol{x})$ may be *mutated* at any mutable vertex k to produce a new seed $(\mu_k(\mathsf{Q}), \mu_k(\boldsymbol{x}))$ for $\mathcal{F}$. The quiver $\mu_k(\mathsf{Q})$ is constructed in three steps:

(1) For each pair of arrows $i \to k \to j$ through the vertex being mutated, add an arrow $i \to j$.

(2) Reverse the orientation of every arrow incident to k.

(3) Cancel any directed 2-cycles in pairs.



[1]This convention allows us to define cluster algebras in characteristic two, and otherwise produces the same definition as the usual convention in every other characteristic. The point is that this convention prevents the numerator in the mutation formula (2-1) from being 2, which in characteristic two would mean that a mutation at that vertex would *not produce* another valid cluster variable.

The cluster variables in $\mu_k(\boldsymbol{x})$ are the same as those in $\boldsymbol{x}$, except for the cluster variable at vertex k, which becomes

$$x_k' := \Big(\prod_{i\to k} x_i + \prod_{k\to i} x_i\Big) x_k^{-1}, \tag{2-1}$$

where the products are over all arrows into or out of k, respectively. Frozen vertices stay frozen.

Mutating at the same vertex twice in a row returns the original seed. That is, mutation is an involution on the set of seeds of $\mathcal{F}/\mathsf{k}$. Two seeds are *mutation-equivalent* if they are related by a sequence of mutations.

Cluster algebras. Fix a seed $(\mathsf{Q}, \boldsymbol{x})$ for $\mathcal{F}$ over k. The union of all the clusters which appear in mutation-equivalent seeds defines the complete set of *cluster variables* in the ambient field $\mathcal{F}$, naturally grouped into overlapping *clusters* consisting of those appearing together in a seed. The *cluster algebra* $\mathcal{A}(\mathsf{Q}, \boldsymbol{x})$ determined by $(\mathsf{Q}, \boldsymbol{x})$ is the sub-k-algebra of $\mathcal{F}$ generated by all of the cluster variables and the inverses of the frozen variables. The cluster algebra only depends on the mutation-equivalence class of the initial seed, and so the initial seed $(\mathsf{Q}, \boldsymbol{x})$ will often be omitted from the notation.

A fundamental property of cluster algebras is the *Laurent phenomenon* [Fomin and Zelevinsky 2002], which states that each cluster variable can be expressed as a Laurent polynomial in *any* cluster. Put differently, the localization of $\mathcal{A}$ at any cluster $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ is the ring of Laurent polynomials in $\boldsymbol{x}$ over k:

$$\mathcal{A} \hookrightarrow \mathcal{A}[x_1^{-1}, x_2^{-1}, \dots, x_n^{-1}] = \mathsf{k}[x_1^{\pm 1}, x_2^{\pm 1}, \dots, x_n^{\pm 1}] \subset \mathcal{F}.$$

Every cluster in $\mathcal{A}$ defines such an inclusion. This naturally leads to the following definition: the *upper cluster algebra* $\mathcal{U}$ of $\mathcal{A}$ is the intersection of each of these Laurent rings, taken inside the ambient field $\mathcal{F}$:

$$\mathcal{U} := \bigcap_{\text{clusters } \boldsymbol{x} \subset \mathcal{A}} \mathsf{k}[x_1^{\pm 1}, x_2^{\pm 1}, \dots, x_n^{\pm 1}] \subset \mathcal{F}.$$

By the Laurent phenomenon, there is an inclusion $\mathcal{A} \subseteq \mathcal{U}$. This inclusion is not always equality (see [Berenstein et al. 2005, Proposition 1.26]), but it is an equality in all of the simplest examples, and in many of the most important examples.

Lemma 2.1. *Upper cluster algebras are normal.*

Proof. Laurent rings over fields are regular, and hence normal. Since an intersection of normal rings inside their common fraction field is normal, upper cluster algebras are normal. $\square$

Cluster localization. Under certain circumstances, localizing a cluster algebra at one or more cluster variables is again a cluster algebra. This important idea was first discussed in [Muller 2013] and further developed in [Muller 2014], to which we refer for more details.

Given a seed $(\mathsf{Q}, \boldsymbol{x})$ over k and a designated subset $\{k_1, k_2, \dots, k_a\}$ of its mutable vertices, we can make a new seed $(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ by making those vertices frozen. Because mutations for $\mathsf{Q}^\dagger$ are all mutations for Q, there is a natural containment

$$\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger) \subseteq \mathcal{A}(\mathsf{Q}, \boldsymbol{x})[x_{k_1}^{-1}, x_{k_2}^{-1}, \dots, x_{k_a}^{-1}]. \tag{2-2}$$

If this is an equality, $\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ is called a *cluster localization* of $\mathcal{A}(\mathsf{Q}, \boldsymbol{x})$.

Although it can be difficult to determine whether a particular localization is a cluster localization, there is one situation where it is easy. Indeed, we have inclusions

$$\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger) \subseteq \mathcal{A}(\mathsf{Q}, \boldsymbol{x})[x_{k_1}^{-1}, x_{k_2}^{-1}, \dots, x_{k_a}^{-1}] \subseteq \mathcal{U}(\mathsf{Q}, \boldsymbol{x})[x_{k_1}^{-1}, x_{k_2}^{-1}, \dots, x_{k_a}^{-1}] \subseteq \mathcal{U}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger), \tag{2-3}$$

where the first and third inclusions follow from the fact that the mutations for $\mathsf{Q}^\dagger$ are a subset of the mutations for Q (and the middle inclusion follows from the Laurent phenomenon for $\mathcal{A}(\mathsf{Q})$). Thus the inclusion in (2-2) is always equality whenever $\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger) = \mathcal{U}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$. One extreme case is where we freeze all vertices: since obviously $\mathcal{A} = \mathcal{U}$ when no mutations can happen, it follows that localizing at any full cluster $\boldsymbol{x}$ is a cluster localization. More generally, $\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger) = \mathcal{U}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ will necessarily hold if "enough" mutable vertices become frozen.

For example, if we freeze enough variables to break any directed cycles in Q, we arrive at an *acyclic* quiver $\mathsf{Q}^\dagger$. By definition, a quiver is acyclic if it has no directed cycles through mutable vertices; a *cluster algebra is acyclic* if it admits some acyclic seed. Because acyclic cluster algebras are known to equal their upper cluster algebras (by Theorem 2.6 below), the chain of inclusions (2-3) above implies that $\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ is a cluster localization whenever $\mathsf{Q}^\dagger$ is acyclic.

Cluster covers. The idea of cluster localization is powerful when a cluster algebra can be covered by cluster localizations.

Definition 2.2. For a cluster algebra $\mathcal{A}$, a set $\{\mathcal{A}_i\}_{i\in I}$ of cluster localizations of $\mathcal{A}$ is called a *cluster cover* if the corresponding open subschemes cover $\operatorname{Spec}(\mathcal{A})$, that is, if

$$\operatorname{Spec}(\mathcal{A}) = \bigcup_{i\in I} \operatorname{Spec}(\mathcal{A}_i).$$

If a cluster algebra admits a cluster cover, any geometric property, such as normality, smoothness, or even different classes of singularities, can be checked locally on the cluster localizations. Another property which may be checked on a cover is whether $\mathcal{A} = \mathcal{U}$:

Lemma 2.3 [Muller 2014, Lemma 3.3.2]. *If $\{\mathcal{A}_i\}_{i\in I}$ is a cluster cover of $\mathcal{A}$, and $\mathcal{A}_i = \mathcal{U}_i$ for each $i \in I$, then $\mathcal{A} = \mathcal{U}$.*

A powerful observation proposed in [Muller 2013] is that many notable classes of cluster algebras admit covers by *acyclic cluster algebras*.

Definition 2.4. A cluster algebra is *locally acyclic* if it admits a cluster cover by acyclic cluster algebras.

The class of locally acyclic cluster algebras is *much wider* than the class of acyclic cluster algebras. The latter class is well-understood and very nicely behaved, but far too restrictive to be itself a major class. On the other hand, *locally* acyclic cluster algebras include, for example, cluster algebras of Grassmannians, cluster algebras of *marked surfaces* with at least two marked points on the boundary [Muller 2013, Theorem 10.6], as well as cluster algebras of double Bruhat cells and more generally, *positroid cells* [Muller and Speyer 2014]. Because the geometric properties of locally acyclic cluster algebras follow nicely from the acyclic case, there is now substantial interest in identifying locally acyclic cluster algebras.

Proposition 2.5 [Muller 2013]. *A locally acyclic cluster algebra over $\Bbbk$ is finitely generated over $\Bbbk$ and equal to its upper cluster algebra. A locally acyclic cluster algebra is normal and a local complete intersection* (*hence Gorenstein*).

This follows with little fuss from the acyclic case, due to Berenstein, Fomin, and Zelevinsky.[2]

Theorem 2.6 [Berenstein et al. 2005, Corollary 1.17; Muller 2014, Corollary 4.2.2]. *Let $(\mathsf{Q}, \boldsymbol{x})$ be an acyclic seed. Then the cluster algebra $\mathcal{A}(\mathsf{Q})$ is a finitely generated complete intersection, equal to its upper cluster algebra $\mathcal{U}(\mathsf{Q})$.*

Remark 2.7. It is important to note that not every cluster algebra admits a cover by proper cluster localizations. For example, the Markov cluster algebra generated from the middle seed in Figure 1 cannot by covered by proper cluster localizations. Indeed, one easily checks that $\mathcal{A}$ can be $\mathbb{N}$-graded, with every cluster variable homogeneous of degree one. So, any nontrivial cluster localization $\operatorname{Spec}(\mathcal{A}_i)$ of $\operatorname{Spec}(\mathcal{A})$ necessary misses the unique homogeneous maximal ideal generated by the cluster variables.

3. Frobenius splittings

Frobenius splittings. Every domain[3] R over a field of positive characteristic p has a canonical ring map, the *Frobenius endomorphism*

$$F : R \to R, \quad \text{defined by } x \mapsto x^p.$$

[2] The proof in [Berenstein et al. 2005] assumes an additional condition, that the cluster algebra is totally coprime. However, it was shown in [Muller 2014] that this condition is unnecessary.

[3] The assumption that R is a domain is completely unnecessary, but it simplifies our discussion and is sufficient for our purposes.

The Frobenius map is an R-module map if we equip the target copy of R with the R-module structure it gets via restriction of scalars. In practice, it is convenient to denote the target copy of R by some other notation. We denote the target copy by $R^{1/p}$ and its elements by $r^{1/p}$, which is consistent with viewing the target copy of R as (the canonically isomorphic ring) $R^{1/p}$ inside the algebraic closure of the fraction field of R. In this case, the elements of r act on elements $x^{1/p} \in R^{1/p}$ by $r \cdot x^{1/p} = (r^p x)^{1/p}$, the usual multiplication $r x^{1/p}$ in the fraction field. In this notation, the Frobenius map becomes the inclusion

$$\begin{aligned} R &\xhookrightarrow{F} R^{1/p}, \\ r &\mapsto (r^p)^{1/p} = r, \end{aligned} \tag{3-1}$$

of R into the overring $R^{1/p}$ of p-th roots.

We say that R is *F-finite* if $R^{1/p}$ is a finitely generated R-module. This is a fairly weak condition, satisfied, for example, by every finitely generated algebra over a perfect field k.

A famous theorem of Kunz [1969, Theorem 2.1] states that an F-finite domain R is regular if and only if $R^{1/p}$ is locally free over R. More generally, one should expect that the closer $R^{1/p}$ is to being locally free over R, the milder the singularities of R. Frobenius split rings and strongly F-regular rings are examples of rings in which some degree of "freeness" is retained of $R^{1/p}$ over R.

Definition 3.1. A domain R is *Frobenius split* if the map (3-1) splits in the category of R-modules. A choice of splitting $\phi : R^{1/p} \to R$ is called a *Frobenius splitting*.

Example 3.2. Every field k of characteristic p is Frobenius split, since $\mathsf{k}^{1/p}$ is a vector space over the subfield k. For a perfect field k, the Frobenius endomorphism is a field isomorphism, and its inverse is the unique Frobenius splitting of k.

Example 3.3. Polynomial rings are Frobenius split. Define the *standard splitting* of the polynomial ring $\mathsf{k}[x_1, x_2, \dots, x_n]$ to be given by

$$\phi : (\mathsf{k}[x_1, x_2, \dots, x_n])^{1/p} \to \mathsf{k}[x_1, x_2, \dots, x_n],$$

$$\phi((\lambda x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n})^{1/p}) = \begin{cases} \phi(\lambda^{1/p}) x_1^{a_1/p} x_2^{a_2/p} \cdots x_n^{a_n/p} & \text{if } a_1, a_2, \dots, a_n \in p\mathbb{Z}, \\ 0 & \text{otherwise,} \end{cases}$$

where the map $\phi : \mathsf{k}^{1/p} \to \mathsf{k}$ on scalars λ is taken to be any fixed splitting of Frobenius.

Remark 3.4. The standard splitting of a polynomial ring is a Frobenius splitting, and will be the source of Frobenius splittings of cluster algebras. It depends on a choice of generators $\boldsymbol{x}$ and if k is not perfect, it depends on a choice of Frobenius splitting for k. We suppress the dependence on the choice of a Frobenius splitting of k by assuming our ground field comes with a fixed Frobenius splitting. In any case, when k is perfect, there is a unique splitting.

The standard splitting of a polynomial ring induces a splitting, also called the *standard splitting*, of the k-Laurent ring $L = \mathsf{k}[x_1^{\pm 1}, \dots, x_n^{\pm 1}]$, using the exact same formula as above. An isomorphism between two k-Laurent rings will commute with the standard splitting, so it does not depend on a choice of presentation.

The standard splitting of a k-Laurent ring has the following key property.

Lemma 3.5. *If L is a finitely generated Laurent ring over a perfect*[4] *field* k *of characteristic p, then the standard splitting ϕ freely generates* $\operatorname{Hom}_L(L^{1/p}, L)$ *as an $L^{1/p}$-module.*

Explicitly, every L-module map $L^{1/p} \to L$ (including every Frobenius splitting) can be written as the composition

$$L^{1/p} \xrightarrow{m_s} L^{1/p} \xrightarrow{\phi} L$$

of the standard splitting ϕ and "multiplication by $s^{1/p}$" map m_s for some *unique* $s \in L$. We denote this composition by $\phi \circ s^{1/p}$. We include a short proof, although it may be well-known to experts.

Proof. Let $L = \mathsf{k}[x_1^{\pm 1}, x_2^{\pm 1}, \dots, x_n^{\pm 1}]$. As an L-module, $L^{1/p}$ has a basis consisting of monomials $\boldsymbol{x}^{\boldsymbol{a}} = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$ for which $0 \le a_i < p$. For any $\psi \in \operatorname{Hom}_L(L^{1/p}, L)$, define

$$s := \sum_{\boldsymbol{a} \,|\, 0 \le a_i < p} \psi(\boldsymbol{x}^{\boldsymbol{a}})^p \boldsymbol{x}^{-\boldsymbol{a}}.$$

Then, for any $\boldsymbol{b}$ with $0 \le b_i < p$,

$$\phi((s\boldsymbol{x}^{\boldsymbol{b}})^{1/p}) = \phi\Big(\Big(\sum \psi(\boldsymbol{x}^{\boldsymbol{a}})^p \boldsymbol{x}^{\boldsymbol{b}-\boldsymbol{a}}\Big)^{1/p}\Big) = \sum \psi(\boldsymbol{x}^{\boldsymbol{a}}) \phi(\boldsymbol{x}^{\boldsymbol{b}-\boldsymbol{a}})^{1/p} = \psi(\boldsymbol{x}^{\boldsymbol{b}}).$$

Since $\phi \circ s^{1/p}$ and ψ agree on a basis for $L^{1/p}$, they coincide. ☐

Remark 3.6. In fact, for any local or graded Gorenstein F-finite ring S of characteristic p, the module $\operatorname{Hom}_S(S^{1/p}, S)$ is a free rank-one $S^{1/p}$-module, since in this case, $\operatorname{Hom}_S(S^{1/p}, S)$ is a canonical module for $S^{1/p}$.[5] The point of Lemma 3.5 is that for a Laurent ring L, a Frobenius splitting gives a natural generator for $\operatorname{Hom}_L(L^{1/p}, L)$, and hence for the canonical module of L.

One special case is for a field. If $\mathcal{F}$ is a field, then $\operatorname{Hom}_{\mathcal{F}}(\mathcal{F}^{1/p}, \mathcal{F})$ is a one-dimensional vector space over $\mathcal{F}^{1/p}$, so we can take any nonzero mapping to be a basis. In particular, if we fix a splitting $\phi : \mathcal{F}^{1/p} \to \mathcal{F}$, then every $\psi : \mathcal{F}^{1/p} \to \mathcal{F}$ is the composition $\psi = \phi \circ s^{1/p}$ for some *unique* $s \in \mathcal{F}$.

[4]Perfect is not necessary here, but it suffices for our purposes and simplifies the discussion.

[5]This is a special case of the general "upper shriek" formula for a finite extension $R \to S$ that $\omega_S \cong \operatorname{Hom}_R(S, \omega_R)$ [Hartshorne 1966; Bruns and Herzog 1993, Theorem 3.3.7b]. Note that is S if Gorenstein, then so is the isomorphic ring $S^{1/p}$; hence $\omega_S \cong S$.

Frobenius splittings of upper cluster algebras. As we now prove, upper cluster algebras are *always* Frobenius split. Indeed, there is a natural *cluster splitting* which is compatible with the cluster structure:

Theorem 3.7. *Suppose that $\mathcal{U}$ is an upper cluster algebra over a field k of positive characteristic. For any cluster $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, the standard splitting of $\mathsf{k}[x_1^{\pm 1}, x_2^{\pm 1}, \ldots, x_n^{\pm 1}]$ restricts to a splitting of $\mathcal{U}$. This splitting of $\mathcal{U}$ does not depend on the choice of cluster.*

The point of the proof is the following simple but *crucial observation*: a subalgebra R of a Frobenius split algebra S is Frobenius split if $\phi(R^{1/p}) \subseteq R$, where ϕ is some Frobenius splitting for S.

Proof. Let $\boldsymbol{x}' = \{x_1', x_2, \ldots, x_n\}$ be the mutation of $\boldsymbol{x}$ at 1, and let $P_1 = x_1 x_1'$ be the numerator of the mutation (see (2-1)). The standard splitting $\phi_{\boldsymbol{x}}$ of the Laurent ring $L_{\boldsymbol{x}}$ extends to a splitting of the fraction field $\mathcal{F}$ by localization; we check that this splitting restricts to the standard splitting $\phi_{\boldsymbol{x}'}$ on the Laurent ring $L_{\boldsymbol{x}'}$.

$$\phi_{\boldsymbol{x}}((\boldsymbol{x}'^{\alpha})^{1/p}) = \phi_{\boldsymbol{x}}((x_1'^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n})^{1/p}) = \phi_{\boldsymbol{x}}((P_1^{\alpha_1} x_1^{-\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n})^{1/p}).$$

Since P_1 does not contain x_1, the expression inside $\phi_{\boldsymbol{x}}$ is $x_1^{-\alpha_1}$ times a rational function of $x_2, \ldots, x_n$. It follows that this is zero, unless $\alpha_1 = p\beta_1$ for some $\beta_1 \in \mathbb{Z}$. In this case, $\phi_{\boldsymbol{x}}((\boldsymbol{x}'^{\alpha})^{1/p}) = \phi_{\boldsymbol{x}}((x_1'^{\beta_1 p} x_2^{\alpha_2} \cdots x_n^{\alpha_n})^{1/p}) = x_1'^{\beta_1} \phi_{\boldsymbol{x}}((x_2^{\alpha_2} \cdots x_n^{\alpha_n})^{1/p})$. Since this last expression is a Laurent monomial in $\boldsymbol{x}$, we find that

$$\phi_{\boldsymbol{x}}((\boldsymbol{x}'^{\alpha})^{1/p}) := \begin{cases} \boldsymbol{x}'^{\beta} & \text{if } \alpha = p\beta, \\ 0 & \text{otherwise,} \end{cases}$$

and so $\phi_{\boldsymbol{x}} = \phi_{\boldsymbol{x}'}$ on $\mathcal{F}$. Iterating this argument, we see that every cluster $\boldsymbol{x}$ gives the same splitting on $\mathcal{F}$. Since this splitting preserves each Laurent ring $L_{\boldsymbol{x}}$, it preserves their intersection $\mathcal{U}$. □

The cluster splitting of $\mathcal{U}$ inherits the key property from Lemma 3.5.

Theorem 3.8. *Let $\mathcal{U}$ be an upper cluster algebra over a perfect field k. The cluster splitting ϕ of $\mathcal{U}$ freely generates $\mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p}, \mathcal{U})$ as a $\mathcal{U}^{1/p}$-module.*

Proof. Consider a $\mathcal{U}$-module map $\psi : \mathcal{U}^{1/p} \to \mathcal{U}$. This map induces, by localization, an $\mathcal{F}$-linear map $\psi : \mathcal{F}^{1/p} \to \mathcal{F}$, which we also denote (somewhat abusively) by ψ. Since $\mathrm{Hom}_{\mathcal{F}}(\mathcal{F}^{1/p}, \mathcal{F})$ is a one-dimensional vector space over $\mathcal{F}^{1/p}$ generated by the (localization of the) standard splitting ϕ, we can write ψ as $\phi \circ s^{1/p}$ for some *unique* $s \in \mathcal{F}$. We aim to show that $s \in \mathcal{U}$. This will complete the proof, as every $\psi \in \mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p}, \mathcal{U})$ will then be the composition of the standard splitting with premultiplication by a unique $s^{1/p}$ in $\mathcal{U}^{1/p}$.

To show that $s \in \mathcal{U}$, it suffices to show that $s \in L_{\boldsymbol{x}}$, where $L_{\boldsymbol{x}}$ is the Laurent ring on *any* cluster $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. Note that $L_{\boldsymbol{x}}$ is the localization of $\mathcal{U}$

at the cluster variables $\{x_1, x_2, \dots, x_n\}$. Thus the map $\psi : \mathcal{U}^{1/p} \to \mathcal{U}$ induces an $L_{\boldsymbol{x}}$-module map $\psi_{\boldsymbol{x}} : (L_{\boldsymbol{x}})^{1/p} \to L_{\boldsymbol{x}}$, which we again call ψ. By Lemma 3.5, there is a unique $s_{\boldsymbol{x}} \in L_{\boldsymbol{x}}$ such that $\psi_{\boldsymbol{x}}(r^{1/p}) = \phi((s_{\boldsymbol{x}} r)^{1/p})$ for all $r \in L_{\boldsymbol{x}}$. But now, localizing further to the fraction field $\mathcal{F}$, this map is of course the same as the map $\phi \circ s^{1/p}$ from the first paragraph; that is, $\phi \circ s^{1/p} = \phi \circ s_{\boldsymbol{x}}^{1/p}$. So by the uniqueness of s, we see that $s = s_{\boldsymbol{x}} \in L_{\boldsymbol{x}}$. Since this works for any cluster $\boldsymbol{x}$, it follows that $s \in \mathcal{U}$. □

Remark 3.9. The existence of the cluster splitting of $\mathcal{U}$ is closely related to the fact that the *canonical module* $\omega_{\mathcal{U}/\mathsf{k}}$ is free (see Remark 3.6). This is addressed in the Appendix, which also describes the relation to Frobenius splittings.

4. F-regularity of locally acyclic cluster algebras

Strong F-regularity is a strengthened form of Frobenius splitting, first introduced by Hochster and Huneke [1988]. Strongly F-regular rings have many nice properties: they are Cohen–Macaulay, normal, and have pseudorational singularities, to name a few. Our main theorem in this section is that locally acyclic cluster algebras are strongly F-regular.

Strong F-regularity. Fix a domain R of characteristic $p > 0$. We continue to assume that R is F-finite, meaning that $R^{1/p}$ is finitely generated over R. This is always satisfied for algebras finitely generated over a perfect field.

Strong F-regularity will be a splitting condition on iterates of the Frobenius map. For any natural number e, let $F^e : R \to R$ denote the e-th iterate of Frobenius, so that $F^e(r) = r^{p^e}$ for all $r \in R$. As in the opening paragraphs of Section 3, it is convenient to replace the target copy of R by the canonically isomorphic ring R^{1/p^e} and view the Frobenius map as the inclusion

$$R \hookrightarrow R^{1/p^e}$$

inside the algebraic closure of the fraction field of R.

If $R \hookrightarrow R^{1/p}$ splits, it is easy to see that every iterate $R \hookrightarrow R^{1/p^e}$ splits as well. Indeed, if $\phi : R^{1/p} \to R$ is a Frobenius splitting, then there is a naturally induced R-module splitting $\phi^e : R^{1/p^e} \to R$ induced by composition

$$R^{1/p^e} \xrightarrow{\phi^{1/p^{e-1}}} R^{1/p^{e-1}} \longrightarrow \cdots \longrightarrow R^{1/p} \xrightarrow{\phi} R.$$

In particular, upper cluster algebras also have *cluster splittings* ϕ^e for the inclusions $\mathcal{U} \hookrightarrow \mathcal{U}^{1/p^e}$, and one easily checks (using the same proof) that ϕ^e is a generator for $\operatorname{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$ as a $\mathcal{U}^{1/p^e}$-module as in Theorem 3.8.

Definition 4.1. An F-finite domain R is *strongly F-regular* if for every nonzero element $x \in R$, there exists $e \in \mathbb{N}$ and $\psi \in \operatorname{Hom}_R(R^{1/p^e}, R)$ such that $\psi(x^{1/p^e}) = 1$.

Though not apparent from its definition, strong F-regularity is a geometric property which restricts how bad singularities can be. The next two well-known theorems are examples of this. See also [Smith and Zhang 2015] for a recent survey of F-regularity.

Theorem 4.2 [Hochster and Huneke 1989, Theorem 3.1c]. *An F-finite regular ring is strongly F-regular.*

Theorem 4.3. *A Noetherian strongly F-regular ring is*:

(1) *Frobenius split*;

(2) *Cohen–Macaulay and normal* [Hochster and Huneke 1989, Theorem 3.1d];

(3) *pseudorational* (see [Smith 1997a]);

(4) *Kawamata log terminal whenever it is $\mathbb{Q}$-Gorenstein* [Hara and Watanabe 2002] (*or more generally Kawamata log terminal in the sense of Schwede and Smith's result [2010]*: *there exists a boundary divisor Δ such that the pair (X, Δ) is Kawamata log terminal*).

Like most good geometric properties, strong F-regularity is a local condition; this is essential for our application to locally acyclic cluster algebras.

Lemma 4.4 [Hochster and Huneke 1989, Theorem 3.1a]. *A domain R is strongly F-regular if and only if $R_{\mathfrak{p}}$ is strongly F-regular, for each prime ideal $\mathfrak{p}$.*

In practice, to determine whether or not R is strongly F-regular, it often suffices to check the condition in the definition for a single element x.

Proposition 4.5 [Hochster and Huneke 1989, Theorem 3.3]. *Let R be a Noetherian F-finite domain which is Frobenius split. If there is some nonzero $c \in R$ such that*

(1) *$R_c = R[c^{-1}]$ is strongly F-regular*, *and*

(2) *there exists $e \in \mathbb{N}$ and $\psi \in \operatorname{Hom}_R(R^{1/p^e}, R)$ such that $\psi(c^{1/p^e}) = 1$*,

then R is strongly F-regular.

Proof. This is a well-known result lacking a precise reference easy for nonexperts to parse (see [Hochster and Huneke 1989, Theorem 3.1a] or [Smith 2000, Theorem 3.10]), so we include a proof. Take any nonzero $x \in R$. By (1), there exists $\psi \in \operatorname{Hom}_{R_c}({R_c}^{1/p^f}, R_c)$ such that $\psi(x^{1/p^f}) = 1$. Since $\operatorname{Hom}_{R_c}(R_c^{1/p^n}, R_c) = \operatorname{Hom}_R(R^{1/p^n}, R) \otimes_R R_c$, we know $\psi = (1/c^q)\widetilde{\psi}$ for some $\widetilde{\psi} \in \operatorname{Hom}_R(R^{1/p^n}, R)$ and some natural number q, which without loss of generality can be assumed a power of p. So $\widetilde{\psi}(x^{1/p^n}) = c^q$. Now, because R is Frobenius split, a splitting $\phi \in \operatorname{Hom}_R(R^{1/q}, R)$ will send $(c^q)^{1/q}$ to c. Composing this with the map given in (2) will produce a map sending $x^{1/qp^{e+n}}$ to 1. So R is strongly F-regular. $\square$

Such an element c is a *test element* for R. These types of test elements were first defined in [Hochster and Huneke 1989]; for a recent survey of test elements in this context, see [Smith and Zhang 2015] (more basic) or [Schwede and Tucker 2012] (more advanced).

***F*-regularity of locally acyclic cluster algebras.** We now establish the main result of this section, the F-regularity of locally acyclic cluster algebras.

Theorem 4.6. *A locally acyclic cluster algebra $\mathcal{A}$ over an F-finite field k of prime characteristic is strongly F-regular.*

Proof. The assumption on the field ensures that $\mathcal{A}$ is F-finite. Strong F-regularity is a local condition (see Lemma 4.4), and so it can be checked on an open affine cover. Since locally acyclic cluster algebras admit an open affine cover by acyclic cluster algebras, it suffices to prove the theorem for acyclic cluster algebras.

Fix an *acyclic* seed $(\mathsf{Q}, \boldsymbol{x})$ for $\mathcal{A}$. We induce on the number of mutable vertices to prove that $\mathcal{A}$ is strongly F-regular.

First, suppose there is only one mutable variable; call it x_1. Then

$$\mathcal{A} = \mathsf{k}[x_1, x_1', x_2^{\pm 1}, \dots, x_n^{\pm 1}] / \langle x_1 x_1' - p_1^+ - p_1^- \rangle,$$

where p_1^+ and p_1^+ are monomials in $x_2, \dots, x_n$ with disjoint supports. This is a localization of the hypersurface algebra

$$S = \mathsf{k}[x_1, x_1', x_2, \dots, x_n] / \langle x_1 x_1' - p_1^+ - p_1^- \rangle.$$

Since at least one of p_1^+ and p_1^- is not 1,[6] the corresponding Jacobian ideal contains a monomial in $x_2, \dots, x_n$, and so the Jacobian ideal becomes trivial in the localization to $\mathcal{A}$. Hence, $\mathcal{A}$ is regular, so it strongly F-regular by Theorem 4.2.

Assume now by induction that any acyclic quiver with $m-1$ mutable vertices defines a strongly F-regular cluster algebra.

Let $(\mathsf{Q}, \boldsymbol{x})$ be an acyclic seed with m mutable vertices. Since Q is acyclic, we can find a vertex which is mutable *and* admits no arrows to any other mutable vertex — a *sink*. Label that vertex x_1, and the remaining mutable vertices $x_2, \dots, x_r$. Let $(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ be the same seed but with x_1 also frozen. Since $(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ is also acyclic, $\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger) = \mathcal{U}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ and so

$$\mathcal{A}(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger) = \mathcal{A}[x_1^{-1}]$$

is a cluster localization. The seed $(\mathsf{Q}^\dagger, \boldsymbol{x}^\dagger)$ is acyclic with $m-1$ mutable vertices, and so by the inductive hypothesis, $\mathcal{A}[x_1^{-1}]$ is strongly F-regular.

Since $\mathcal{A}$ is acyclic, the cluster algebra $\mathcal{A}$ coincides with the upper cluster algebra $\mathcal{U}$ (Theorem 2.6), and so the cluster splitting from Theorem 3.7 is a Frobenius splitting

[6] Due to the assumption that mutable vertices must have at least one incident arrow.

for $\mathcal{A}$. Hence, by Proposition 4.5, it suffices to check only the test element $c := x_1$ in $\mathcal{A}$. We construct a ψ sending x_1^{1/p^e} to 1 directly, using the cluster splitting ϕ.

Let p_1^+ and p_1^- be the monomials appearing in the mutation formula at x_1, so that $x_1x_1' = p_1^+ + p_1^-$. Choose e large enough that each exponent appearing in p_1^+ or p_1^- is less than p^e. Since there are no arrows from x_1 to other mutable vertices, p_1^- is a monomial only in the frozen variables; in particular, it is invertible. Consider the map

$$\psi = \phi^e \circ \left(\frac{x_1'}{p_1^-}\right)^{1/p^e},$$

where ϕ^e is the cluster splitting of $\mathcal{A} \hookrightarrow \mathcal{A}^{1/p^e}$. Then

$$\psi(x_1^{1/p^e}) = \phi^e\left(\left(\frac{x_1x_1'}{p_1^-}\right)^{1/p^e}\right) = \phi^e\left(\left(\frac{p_1^+}{p_1^-}+1\right)^{1/p^e}\right).$$

Since p^e is greater than any exponent in the Laurent monomial p_1^+/p_1^-, we know that ϕ^e kills that term, and so $\psi(x_1^{1/p^e}) = 1$. By Proposition 4.5, this shows that $\mathcal{A}$ is strongly F-regular. This completes the inductive step and the proof. $\square$

Characteristic zero consequences. So far, our results are for cluster algebras over a field of positive characteristic. By a standard miracle, these results imply similar consequences over fields of characteristic zero.

We first need to check that locally acyclic cluster algebras over $\mathbb{Z}$ behave as expected when tensored with a field k. Let $\mathcal{A}_{\mathbb{Z}}$ denote a cluster algebra over $\mathbb{Z}$. Choosing any seed in $\mathcal{A}_{\mathbb{Z}}$ and replacing the cluster with a cluster over k determines a cluster algebra $\mathcal{A}_{\mathsf{k}}$ over k, which is well-defined up to canonical isomorphism.

Lemma 4.7. *If $\mathcal{A}_{\mathbb{Z}}$ is locally acyclic, then $\mathcal{A}_{\mathsf{k}}$ is locally acyclic and $\mathsf{k} \otimes_{\mathbb{Z}} \mathcal{A}_{\mathbb{Z}} \simeq \mathcal{A}_{\mathsf{k}}$.*

Proof. If $\mathcal{A}_{\mathbb{Z}}$ is acyclic, then any acyclic seed of $\mathcal{A}_{\mathbb{Z}}$ corresponds to an acyclic seed of $\mathcal{A}_{\mathsf{k}}$ with the same quiver. The presentations of $\mathcal{A}_{\mathbb{Z}}$ and $\mathcal{A}_{\mathsf{k}}$ from Theorem 2.6 coincide except for the ring of scalars, and so $\mathsf{k} \otimes_{\mathbb{Z}} \mathcal{A}_{\mathbb{Z}} \simeq \mathcal{A}_{\mathsf{k}}$.

If $\mathcal{A}_{\mathbb{Z}}$ is locally acyclic, let $\{(\mathcal{A}_i)_{\mathbb{Z}}\}_{i\in I}$ be a cover by acyclic cluster algebras. By the previous paragraph, $\mathsf{k} \otimes_{\mathbb{Z}} (\mathcal{A}_i)_{\mathbb{Z}} \simeq (\mathcal{A}_i)_{\mathsf{k}}$ is an acyclic cluster localization of $\mathcal{A}_{\mathsf{k}}$. Since extension of scalars sends covers to covers, $\{(\mathcal{A}_i)_{\mathsf{k}}\}_i$ is a cover of $\mathcal{A}_{\mathsf{k}}$. Since the map $\mathsf{k} \otimes_{\mathbb{Z}} \mathcal{A}_{\mathbb{Z}} \to \mathcal{A}_{\mathsf{k}}$ is locally an isomorphism, it is an isomorphism. $\square$

With this in hand, we may prove one of our main theorems.

Theorem 4.8. *A locally acyclic cluster algebra over a field k of characteristic zero has* (*at worst*) *canonical singularities.*

Proof. Let $\mathcal{A}$ be a locally acyclic cluster algebra over k, and let $\mathcal{A}_{\mathbb{Z}}$ be the corresponding locally acyclic cluster algebra over $\mathbb{Z}$. By the preceding lemma, for any prime $p \in \mathbb{Z}$,

$$\mathbb{F}_p \otimes_{\mathbb{Z}} \mathcal{A}_{\mathbb{Z}} \simeq \mathcal{A}_{\mathbb{F}_p}$$

is locally acyclic. By Theorem 4.6, $\mathcal{A}_{\mathbb{F}_p}$ is strongly F-regular. On the other hand, $\mathcal{A}$ is Gorenstein by Proposition 2.5. Thus, by [Smith 1997a, Theorem 4.3], $\mathcal{A}_{\mathsf{k}}$ has (at worst) *rational singularities*, but Gorenstein rational singularities are canonical (see the discussion in [Elkik 1981]). □

5. The nonlocally acyclic setting

What can be said about cluster algebras and upper cluster algebras which are not locally acyclic? We provide examples which demonstrate that strong F-regularity is still possible, but not necessary. We also support the general philosophy that $\mathcal{U}$ should be better-behaved than $\mathcal{A}$ by proving that F-regularity of $\mathcal{A}$ implies F-regularity of $\mathcal{U}$.

We end this section by showing that related algebras, called the *lower bound algebras*, are always Frobenius split. We do not know whether or not lower bound algebras are always strongly F-regular (they are, at least, always Noetherian complete intersections). Nor do we know whether an arbitrary cluster algebra is Frobenius split in general.

***F*-regularity of $\mathcal{A}$ implies *F*-regularity of $\mathcal{U}$.** In this section, we consider a completely arbitrary cluster algebra $\mathcal{A}$ (possibly infinitely generated) over a perfect field.

Lemma 5.1. *Fix any integer $e \geq 1$ and let $\varphi \in \operatorname{Hom}_{\mathcal{A}}(\mathcal{A}^{1/p^e}, \mathcal{A})$. Then φ extends uniquely to a map in $\operatorname{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$.*

Proof. The map φ extends, by localization, to the Laurent ring generated by the cluster variables in any given cluster. Hence, it preserves $\mathcal{U}$, the common intersection of all of these Laurent rings. □

Proposition 5.2. *If $\mathcal{A}$ is strongly F-regular then $\mathcal{U}$ is strongly F-regular.*

Proof. Suppose that c is a nonzero element of $\mathcal{U}$. Since $\mathcal{A}$ and $\mathcal{U}$ have the same fraction field, there is an $a \in \mathcal{A}$ for which ac is a nonzero element of $\mathcal{A}$. Because $\mathcal{A}$ is strongly F-regular, there is an integer $e \geq 1$ and a map $\varphi \in \operatorname{Hom}_{\mathcal{A}}(\mathcal{A}^{1/p^e}, \mathcal{A})$ for which $\varphi((ac)^{1/p^e}) = 1$. By Lemma 5.1, we may extend φ uniquely to a map $\widetilde{\varphi} : \mathcal{U}^{1/p^e} \to \mathcal{U}$.

Let $m_a : \mathcal{U}^{1/p^e} \to \mathcal{U}^{1/p^e}$ be the multiplication map given by $m_a(x^{1/p^e}) = a^{1/p^e} x^{1/p^e}$. Then the composition $\widetilde{\varphi} \circ m_a$ is an element of $\operatorname{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$ which maps c^{1/p^e} to 1. □

In what follows, we focus on the F-regularity of upper cluster algebras.

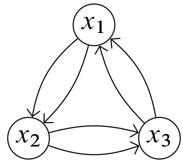


Figure 2. The seed for the Markov cluster algebra in $\mathcal{F} = \mathsf{k}(x_1, x_2, x_3)$.

The Markov upper cluster algebra. Consider the seed $(\mathsf{Q}, \boldsymbol{x})$ defined in Figure 2. Observe that it has three mutable vertices and no frozen vertices.

Introduced in [Berenstein et al. 2005], the *Markov cluster algebra* $\mathcal{A} = \mathcal{A}(\mathsf{Q}, \boldsymbol{x})$ is a standard source of counterexamples and pathologies. For example, $\mathcal{A} = \mathcal{A}(\boldsymbol{x}, \mathsf{B})$ is not a locally acyclic cluster algebra, and indeed $\mathcal{A} \subsetneq \mathcal{U}$ [Berenstein et al. 2005, Theorem 1.26]. Moreover, the Markov cluster algebra $\mathcal{A}$ is not Noetherian [Muller 2013].

Nevertheless, the Markov upper cluster algebra $\mathcal{U} = \mathcal{U}(\mathsf{Q}, \boldsymbol{x})$ is quite well-behaved. It was shown in [Matherne and Muller 2014] that it can be presented as the hypersurface algebra

$$\mathcal{U} \cong \mathsf{k}[x_1, x_2, x_3, M] / \langle x_1 x_2 x_3 M - x_1^2 - x_2^2 - x_3^2 \rangle.$$

Equivalently, the upper cluster algebra $\mathcal{U}$ is generated inside the field $\mathcal{F}$ by x_1, x_2, x_3 and the element

$$M = \frac{x_1^2 + x_2^2 + x_3^2}{x_1 x_2 x_3}.$$

Proposition 5.3. *If* $\operatorname{char}(\mathsf{k}) \neq 2, 3$, *then the Markov upper cluster algebra* $\mathcal{U}$ *is strongly* F*-regular.*

Proof. Since $\mathcal{U}$ is Frobenius split by Theorem 3.7, and the localization of $\mathcal{U}$ at $x_1 x_2 x_3$ is a Laurent ring, $x_1 x_2 x_3$ is a test element for $\mathcal{U}$. Consider now the morphism $\varphi : \mathcal{U}^{1/p^e} \to \mathcal{U}^{1/p^e}$ defined by $\varphi(-) = \phi^e((1/6 \cdot M^3)^{1/p^e} -)$, where ϕ^e is the iterated cluster splitting of $\mathcal{U}$ defined on page 921 and M is described as before. This morphism φ maps $(x_1 x_2 x_3)^{1/p^e}$ to 1:

$$\begin{aligned}\varphi((x_1x_2x_3)^{1/p^e}) &= \phi\left(\left(\frac{1}{6}\left(\frac{x_1^2+x_2^2+x_3^2}{x_1x_2x_3}\right)^3 x_1x_2x_3\right)^{1/p^e}\right)\\ &= \phi\left(\left(\frac{1}{6}\,\frac{\sum_{i+j\le 3, i,j\neq 1} c_{i,j}x_1^{2i}x_2^{2j}x_3^{6-2i-2j} + 6x_1^2x_2^2x_3^2}{(x_1x_2x_3)^2}\right)^{1/p^e}\right)\\ &= 1,\end{aligned}$$

where the $c_{i,j}$ are some combinatorial coefficients. This shows that $\mathcal{U}$ is strongly F-regular. $\square$

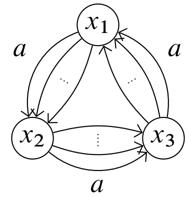


Figure 3. A seed in $\mathcal{F} = \mathsf{k}(x_1, x_2, x_3)$, where $a \geq 2$.

A non-F-regular upper cluster algebra. Generalizing the previous setting, consider the seed $(\mathsf{Q}, \boldsymbol{x})$ defined in Figure 3 for some integer $a \geq 2$.

Let $\mathcal{U} = \mathcal{U}(\mathsf{Q}, \boldsymbol{x})$ denote the associated upper cluster algebra. As shown in [Matherne and Muller 2014], this *generalized Markov upper cluster algebra* can be presented as

$$\mathcal{U} \cong \mathsf{k}[x_1, x_2, x_3, M] / \langle x_1 x_2 x_3 M - x_1^a - x_2^a - x_3^a \rangle.$$

Proposition 5.4. *If $a \geq 3$, then $\mathcal{U}$ is not strongly F-regular.*

Proof. Notice that $\mathcal{U}$ is graded, with

$$\deg(x_1) = \deg(x_2) = \deg(x_3) = 1, \quad \deg(M) = a - 3.$$

When $a \geq 3$, every homogeneous element in $\mathcal{U}$ has degree at least 0. As a consequence, the span of the positive degree elements forms a nonzero ideal I.

The cluster splitting ϕ sends positive degree elements to positive degree elements or zero, so $\phi^e(I^{1/p^e}) \subseteq I$ for any e. By Theorem 3.8, any $\psi \in \mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$ can be written as

$$\psi = \phi^e(s^{1/p^e} -)$$

for some $s \in \mathcal{U}$. Since $(sI)^{1/p^e} \subseteq I^{1/p^e}$, we see that $\psi(I^{1/p^e}) \subseteq I^{1/p^e}$ for any $\psi \in \mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$. In particular, for any $c \in I$, there is no $\psi \in \mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$ such that $\psi(c) = 1 \notin I$, and so $\mathcal{U}$ is not strongly F-regular. $\square$

By Proposition 5.2, this extends to the cluster algebra as well.

Corollary 5.5. *If $a \geq 3$, then $\mathcal{A}(\mathsf{Q}, \boldsymbol{x})$ is not strongly F-regular.*

Remark 5.6. The positive degree elements in $\mathcal{U}$ form an ideal (x_1, x_2, x_3) stable under all maps in $\mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$, that is, compatible with respect to every element of $\mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p^e}, \mathcal{U})$. This is the *test ideal* of $\mathcal{U}$. See [Schwede and Tucker 2012].

Lower bound algebras. Fix a seed $(\mathsf{Q}, \boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. As before, let

$$p_i^+ := \prod_{j \to i} x_j \quad \text{and} \quad p_i^- := \prod_{j \leftarrow i} x_j,$$

and let $x_i' := (p_i^+ + p_i^-)x_i^{-1}$. The algebra $\mathcal{L}(\mathsf{Q}, \boldsymbol{x})$ defined by

$$\mathcal{L}(\mathsf{Q}, \boldsymbol{x}) := \mathsf{k}[x_1, x_2, \ldots, x_n, x_1', x_2', \ldots, x_n'] \subseteq \mathcal{F} = \mathsf{k}(x_1, \ldots, x_n)$$

is called the *lower bound algebra* associated to the seed $(\mathsf{Q}, \boldsymbol{x})$. Notice that $\mathcal{L}(\mathsf{Q}, \boldsymbol{x}) \subseteq \mathcal{A}(\mathsf{Q}, \boldsymbol{x})$. This inclusion is an equality if and only if Q is an acyclic quiver (see [Berenstein et al. 2005, Theorem 1.20]).

Lemma 5.7. *The kernel L of the surjective ring homomorphism*

$$\mathsf{k}[x_1, \ldots, x_n, y_1, \ldots, y_n] \to \mathcal{L}(\mathsf{Q}, \boldsymbol{x}), \quad x_i \mapsto x_i, \; y_i \mapsto x_i',$$

is a prime component of the ideal $I := \langle x_1y_1 - (p_1^+ + p_1^-), \ldots, x_ny_n - (p_n^+ + p_n^-)\rangle$.

Proof. Since $\mathcal{L}(\mathsf{Q}, \boldsymbol{x})$ is a domain, L is a prime ideal. To see that L is a component of I, let $S = \mathsf{k}[x_1, \ldots, x_n, y_1, \ldots, y_n]$, and observe that

$$IS[(x_1 \cdots x_n)^{-1}] = \langle y_1 - x_1', \ldots, y_n - x_n' \rangle.$$

Since $L = \langle y_1 - x_1', \ldots, y_n - x_n' \rangle \cap S$, it follows that $(IS[(x_1 \cdots x_n)^{-1}]) \cap S = L$, and thus L is a prime component of I. □

Proposition 5.8. *The lower bound algebra $\mathcal{L}(\mathsf{Q}, \boldsymbol{x})$ is Frobenius split.*

Proof. Fix any prime $p > 0$, let $S := \mathsf{k}[x_1, \ldots, x_n, y_1, \ldots, y_n]$, and let $\mathcal{B}$ denote the S-module basis of $S^{1/p}$ consisting of all monomials

$$x_1^{a_1/p} \cdots x_n^{a_n/p} y_1^{a_{n+1}/p} \cdots y_n^{a_{2n}/p}, \quad 0 \le a_i < p.$$

Define $\psi : S^{1/p} \to S$ to be the S-linear map which takes value 1 on the basis element $x_1^{(p-1)/p} \cdots x_n^{(p-1)/p} y_1^{(p-1)/p} \cdots y_n^{(p-1)/p}$ and 0 on all other elements of $\mathcal{B}$.

We will construct a Frobenius splitting of S which descends to a Frobenius splitting of $\mathcal{L}$. To this end, let

$$f = \prod_{1 \le i \le n} (x_iy_i - p_i^+ - p_i^-),$$

and observe that premultiplication of ψ by $f^{(p-1)/p}$ is a Frobenius splitting of S. Indeed, since the p^+ and p^- never involved any y, all monomials appearing in $f^{(p-1)}$ involve each y_i to a power less than p; all of these are killed by ψ except for the monomial term $x_1^{(p-1)} \cdots x_n^{(p-1)} y_1^{(p-1)} \cdots y_n^{(p-1)}$, whose p-th root gets sent to 1 by ψ. So, $\psi(f^{(p-1)/p} \cdot 1) = 1$, and $\psi(f^{(p-1)/p} -)$ is thus a Frobenius splitting of S. Furthermore, if $J = \langle f \rangle$, then

$$\psi(f^{(p-1)/p} J^{1/p}) \subseteq J.$$

That is, J is a compatibly split ideal. Because sums and prime components of compatibly split ideals are compatibly split (see, for example, [Brion and Kumar

2005, Chapter 1.2]), the ideal $L \subseteq S$ that defines the lower bound algebra $\mathcal{L}(\mathsf{Q}, \boldsymbol{x})$ is compatibly split. The Frobenius splitting $\psi(f^{(p-1)/p}-) : S^{1/p} \to S$ therefore descends to a Frobenius splitting of the lower bound algebra. □

Appendix: The canonical module of an upper cluster algebra

This appendix considers the canonical module of an upper cluster algebra $\mathcal{U}$ over a field k.[7] Since upper cluster algebras need not be Noetherian [Speyer 2013], we must be careful which definition we use.

Canonical modules. Let S be a normal domain over k such that the fraction field $\mathcal{F}(S)$ has transcendence degree n over k. Define the *canonical module* of S over k to be the S-module

$$\omega_{S/\mathsf{k}} := (\Lambda^n_S \Omega_{S/\mathsf{k}})^{**}.$$

If S is regular (such as a field), then the double dual in the definition is unnecessary, and $\omega_{S/\mathsf{k}} = \Lambda^n_S \Omega_{S/\mathsf{k}}$. This construction commutes with localization; in particular, there is a natural embedding

$$\omega_{S/\mathsf{k}} \subseteq \omega_{\mathcal{F}(S)/\mathsf{k}} = \Lambda^n_{\mathcal{F}(S)} \Omega_{\mathcal{F}(S)/\mathsf{k}}$$

into the canonical module of the fraction field.

The log volume form. Let $\mathcal{U}$ be an upper cluster algebra over k. The algebra $\mathcal{U}$ is normal and the transcendence degree of its fraction field over k is the rank n. For a cluster $\boldsymbol{x}$ with functions $\{x_1, x_2, \ldots, x_n\}$ indexed by $\{1, 2, \ldots, n\}$, let $L_{\boldsymbol{x}}$ denote the k-Laurent ring in that cluster, and define the *log volume form*

$$\mu_{\boldsymbol{x}} := \frac{dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n}{x_1 x_2 \cdots x_n} \in \omega_{L_{\boldsymbol{x}}/\mathsf{k}}.$$

Note that a permutation of the indices may change the sign of this element.

Proposition A.1. *The canonical module $\omega_{L_{\boldsymbol{x}}/\mathsf{k}}$ is free of rank one over $L_{\boldsymbol{x}}$ and generated by the log volume form $\mu_{\boldsymbol{x}}$.*

The log volume form is an invariant of the cluster algebra, up to sign.

Proposition A.2. *For two different clusters $\boldsymbol{x}, \boldsymbol{y}$ of $\mathcal{U}$, we have $\mu_{\boldsymbol{x}} = \pm \mu_{\boldsymbol{y}}$.*

Proof. It is sufficient to check the proposition for a single mutation. Let $\boldsymbol{x}' = \{x_1, x_2, \ldots, x_i', \ldots, x_n\}$, where $x_i' = (p_i^+ + p_i^-)/x_i$. Then

$$\frac{dx_i'}{x_i'} = \frac{d(p_i^+ + p_i^-)}{x_i x_i'} - \frac{(p_i^+ + p_i^-)\,dx_i}{x_i^2 x_i'} = \frac{d(p_i^+ + p_i^-)}{p_i^+ + p_i^-} - \frac{dx_i}{x_i}.$$

[7] The results remain true over $\mathbb{Z}$.

Since p_i^+ and p_i^- are monomials in $\{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$,

$$\frac{dx_1 \wedge \cdots \wedge dx_i' \wedge \cdots \wedge dx_n}{x_1 \cdots x_i' \cdots x_n} = -\frac{dx_1 \wedge \cdots \wedge dx_i \wedge \cdots \wedge dx_n}{x_1 \cdots x_i \cdots x_n}.$$

Hence, $\mu_{\boldsymbol{x}} = -\mu_{\boldsymbol{x}'}$. Iterating mutations or permuting the indices will change this form by at most a sign. □

Canonical modules of upper cluster algebras. Since either log volume form freely generates the canonical module after localizing to a cluster Laurent ring, it follows that they *freely generate* the canonical module of $\mathcal{U}$.

Theorem A.3. *For an upper cluster algebra $\mathcal{U}$ over a field*[8] k, *the canonical module $\omega_{\mathcal{U}/\mathsf{k}}$ is free of rank one over $\mathcal{U}$ and generated by a log volume form in any cluster.*

Proof. Fix a log volume form μ in some cluster. For any cluster $\boldsymbol{x}$, the localization $L_{\boldsymbol{x}} \otimes \Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}}$ equals $L_{\boldsymbol{x}}\mu$ by Proposition A.1. Let

$$\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}} \to f(\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}})$$

be the quotient by the maximal torsion submodule. Then $f(\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}})$ is contained inside the localization $L_{\boldsymbol{x}}\mu$, which is contained inside the generic canonical module $\Lambda^n_{\mathcal{F}(\mathcal{U})} \Omega_{\mathcal{F}(\mathcal{U})/\mathsf{k}}$. Intersecting over all clusters, we obtain a map

$$\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}} \to f(\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}}) \subseteq \bigcap_{\text{clusters } \boldsymbol{x}} (L_{\boldsymbol{x}}\mu) = \Bigg(\bigcap_{\text{clusters } \boldsymbol{x}} L_{\boldsymbol{x}}\Bigg)\mu = \mathcal{U}\mu.$$

Define $\mu^* \in (\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}})^* = \operatorname{Hom}_{\mathcal{U}}(\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}}, \mathcal{U})$ to be the composition of the above map with the $\mathcal{U}$-module map $\mathcal{U}\mu \to \mathcal{U}$ which sends μ to 1.

Consider another $\mathcal{U}$-module map $\psi : \Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}} \to \mathcal{U}$. Since $\mathcal{U}$ is torsion-free, ψ factors through $f(\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}})$. Localizing ψ at a cluster $\boldsymbol{x}$ gives an $L_{\boldsymbol{x}}$-module map $\psi_{\boldsymbol{x}} : L_{\boldsymbol{x}}\mu \to L_{\boldsymbol{x}}$. Let $s_{\boldsymbol{x}} := \psi_{\boldsymbol{x}}(\mu)$, and note that $\psi_{\boldsymbol{x}}(a\mu) = as_{\boldsymbol{x}}$ for all $a \in L_{\boldsymbol{x}}$.

Localizing at a different cluster $\boldsymbol{y}$ gives a map $\psi_{\boldsymbol{y}} : L_{\boldsymbol{y}}\mu \to L_{\boldsymbol{y}}$, which is of the form $\psi_{\boldsymbol{y}}(a\mu) = as_{\boldsymbol{y}}$ for some $s_{\boldsymbol{y}} \in L_{\boldsymbol{y}}$. Since there is some nonzero $b \in U$ such that $b\mu \in f(\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}})$ (the product of the variables in any cluster suffices),

$$bs_{\boldsymbol{x}} = b\psi_{\boldsymbol{x}}(\mu) = \psi(b\mu) = b\psi_{\boldsymbol{y}}(\mu) = bs_{\boldsymbol{y}},$$

and it follows that $s_{\boldsymbol{x}} = s_{\boldsymbol{y}}$ in $L_{\boldsymbol{x}} \cap L_{\boldsymbol{y}}$. Repeating for all clusters, $s_{\boldsymbol{x}} \in \mathcal{U}$. Therefore, $\psi(a\mu) = as_{\boldsymbol{x}} = \mu^*(s_{\boldsymbol{x}} a\mu)$ for all $a \in \mathcal{U}$; this proves that μ^* freely generates the dual module

$$(\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}})^* = \mathcal{U}\mu^*.$$

Dualizing both sides demonstrates that $\omega_{\mathcal{U}/\mathsf{k}} = \mathcal{U}\mu$. □

[8]The theorem remains true when k is an arbitrary normal domain.

There are examples where the log volume forms are not in $\Omega^n_{\mathcal{U}/\mathsf{k}}$; hence, the double dual in the definition of $\omega_{\mathcal{U}/\mathsf{k}}$ is necessary.

Corollary A.4. *A Noetherian upper cluster algebra over a field is Gorenstein.*

Canonical modules and Frobenius splittings. We sketch the relation between canonical modules and Frobenius splittings here; further details may be found in [Brion and Kumar 2005, Section 1.3].

Let k be a field of positive characteristic $p \neq 2$, and let X be a smooth, locally finite-type scheme over k. The Frobenius map becomes a flat, finite morphism

$$f : X \to X.$$

Then the push-forward functor $f_* : \mathrm{Coh}(X) \to \mathrm{Coh}(X)$ has a right-adjoint

$$f^! : \mathrm{Coh}(X) \to \mathrm{Coh}(X),$$

together with an adjunction map

$$\mathrm{tr} : f_* f^! \to \mathrm{Id}$$

called the *trace*.

The coherent sheaf $f^!(\mathcal{O}_X)$ and its trace map can be connected with Frobenius splittings as follows. On any open affine subscheme $\mathrm{Spec}(R) \subseteq X$:

- $f^!(\mathcal{O}_X)[\mathrm{Spec}(R)]$ is isomorphic to $\mathrm{Hom}_R(R^{1/p}, R)$ as an $R^{1/p}$-module; equivalently, to $\mathrm{Hom}_{R^p}(R, R^p)$ as an R-module.
- $f_* f^!(\mathcal{O}_X)[\mathrm{Spec}(R)]$ is isomorphic to $\mathrm{Hom}_R(R^{1/p}, R)$ as an R-module.
- The trace map is given by the R-module map
$$\mathrm{Hom}_R(R^{1/p}, R) \to R,$$
which sends a map $f : R^{1/p} \to R$ to $f(1) \in R$.

Hence, the sheaf of Frobenius splittings is isomorphic to $\mathrm{tr}^{-1}(1) \subset f_* f^!(\mathcal{O}_X)$.

Duality theory for the morphism f gives an alternate description of $f^!(\mathcal{O}_X)$ in terms of the canonical sheaf $\omega_{X/\mathsf{k}}$.

Theorem A.5 [Brion and Kumar 2005, §1.3]. *There are natural isomorphisms*[9]

$$f^!(\mathcal{O}_X) \xrightarrow{\sim} \mathcal{H}om_X(f^*\omega_{X/\mathsf{k}}, \omega_{X/\mathsf{k}}) \xrightarrow{\sim} (\omega_{X/\mathsf{k}})^{1-p},$$

$$f_* f^!(\mathcal{O}_X) \xrightarrow{\sim} \mathcal{H}om_X(\omega_{X/\mathsf{k}}, f_*\omega_{X/\mathsf{k}}) \xrightarrow{\sim} f_*((\omega_{X/\mathsf{k}})^{1-p}),$$

and a map $\tau : f_*\omega_{X/\mathsf{k}} \to \omega_{X/\mathsf{k}}$, such that the trace map is given by

$$f_* f^!(\mathcal{O}_X) \xrightarrow{\sim} \mathcal{H}om_X(\omega_{X/\mathsf{k}}, f_*\omega_{X/\mathsf{k}}) \xrightarrow{\tau\circ -} \mathcal{H}om_X(\omega_{X/\mathsf{k}}, \omega_{X/\mathsf{k}}) \xrightarrow{\sim} \mathcal{O}_X.$$

[9]The negative exponent on $(\omega_{X/\mathsf{k}})^{1-p}$ denotes a positive power of the dual sheaf $\omega^*_{X/\mathsf{k}}$. This is a markedly different use of exponential notation than $R^{1/p}$.

Back to cluster algebras. Let $\mathcal{U}$ be an upper cluster algebra. For any seed $(\mathsf{Q}, \boldsymbol{x})$, the Laurent phenomenon says that freezing every mutable vertex gives the localization $\mathsf{k}[x_1^{\pm 1}, \dots, x_n^{\pm 1}]$ in the cluster $\boldsymbol{x} = \{x_1, \dots, x_n\}$. Geometrically, this means there is an open subscheme

$$\mathrm{Spec}(\mathsf{k}[x_1^{\pm 1}, \dots, x_n^{\pm 1}]) \subseteq \mathrm{Spec}(\mathcal{U}).$$

Let us call subschemes of this form *cluster tori*. Let $X \subseteq \mathrm{Spec}(\mathcal{U})$ be the *union of the cluster tori*, as $(\mathsf{Q}, \boldsymbol{x})$ runs over all seeds.

While the scheme $\mathrm{Spec}(\mathcal{U})$ is generally neither smooth nor locally finite type over k, the open subscheme X is both. Hence, by Theorem A.5, we have isomorphisms

$$f^!(\mathcal{O}_X) \xrightarrow{\sim} \mathcal{H}om_X(f^*\omega_{X/\mathsf{k}}, \omega_{X/\mathsf{k}}) \xrightarrow{\sim} (\omega_{X/\mathsf{k}})^{1-p}.$$

Proposition A.6. *Let X be the union of the cluster tori in Spec$(\mathcal{U})$.*

(1) *$\mathcal{O}_X(X)$ is isomorphic to $\mathcal{U}$ as a k-algebra.*

(2) *$f^!(\mathcal{O}_X)(X)$ is isomorphic to* $\mathrm{Hom}_{\mathcal{U}^p}(\mathcal{U}, \mathcal{U}^p)$ *as a $\mathcal{U}$-module.*

(3) *$\omega_{X/\mathsf{k}}(X)$ is isomorphic to $\omega_{\mathcal{U}/\mathsf{k}}$ as a $\mathcal{U}$-module.*

On any cluster torus, these isomorphisms restrict to the obvious isomorphisms.

Proof. The first isomorphism is a standard fact about upper cluster algebras; see, e.g., [Matherne and Muller 2014, Proposition 3.4]. The other two isomorphisms follow from Theorems 3.8 and A.3, which show that $\mathrm{Hom}_{\mathcal{U}^p}(\mathcal{U}, \mathcal{U}^p)$ and $\omega_{\mathcal{U}/\mathsf{k}}$ are each free of rank one over $\mathcal{U}$ with a distinguished generator (the cluster splitting and either log volume form, respectively). On each cluster torus, the sheaves $f^!(\mathcal{O}_X)$ and $\omega_{X/\mathsf{k}}$ are free and generated by the restriction of the generator. Hence, a global section of $f^!(\mathcal{O}_X)$ or $\omega_{X/\mathsf{k}}$ can be written as the distinguished generator times a rational function which is Laurent in each cluster, that is, an element of $\mathcal{U}$. $\square$

As a consequence, we have an isomorphism of $\mathcal{U}$-modules

$$\mathrm{Hom}_{\mathcal{U}^p}(\mathcal{U}, \mathcal{U}^p) \xrightarrow{\sim} (\omega_{\mathcal{U}/\mathsf{k}})^{1-p},$$

where $\omega_{\mathcal{U}/\mathsf{k}}$ to a negative power means $\omega^*_{\mathcal{U}/\mathsf{k}} = (\Lambda^n_{\mathcal{U}} \Omega_{\mathcal{U}/\mathsf{k}})^*$ to a positive power.

The connection between cluster splittings and log volume forms starts to become clear. Theorem A.3 establishes that $\omega_{\mathcal{U}/\mathsf{k}}$ is free of rank one as a $\mathcal{U}$-module. Hence, $\mathrm{Hom}_{\mathcal{U}}(\mathcal{U}, \mathcal{U}^p)$ is free of rank one as a $\mathcal{U}$-module, or equivalently, $\mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p}, \mathcal{U})$ is free of rank one as a $\mathcal{U}^{1/p}$-module.

To choose a distinguished generator, we observe that $\omega_{\mathcal{U}/\mathsf{k}}$ has two natural generators (the log volume forms) which differ by a sign. Since p is odd, the $(1-p)$-th power of the two log volume forms coincide, so $(\omega_{\mathcal{U}/\mathsf{k}})^{1-p}$ has a canonical generator. This determines a canonical generator in $\mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p}, \mathcal{U})$ over $\mathcal{U}^{1/p}$; all that remains is to observe that it coincides with the cluster splitting.

Proposition A.7. *If $\mu \in \omega_{\mathcal{U}/\mathsf{k}}$ is either log volume form, then the image of μ^{1-p} under the map*

$$(\omega_{\mathcal{U}/\mathsf{k}})^{1-p} \xrightarrow{\sim} \mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p}, \mathcal{U})$$

is the cluster spitting $\phi : \mathcal{U}^{1/p} \to \mathcal{U}$.

The reader is cautioned that, as written, this is not a module map; rather, it intertwines the $\mathcal{U}$-action on $(\omega_{\mathcal{U}/\mathsf{k}})^{1-p}$ and the $\mathcal{U}^{1/p}$-action on $\mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p}, \mathcal{U})$.

Proof. If $L_{\boldsymbol{x}}$ is the k-Laurent ring in some cluster $\boldsymbol{x}$, the localization map $\mathcal{U} \subset L_{\boldsymbol{x}}$ induces localization maps

$$(\omega_{\mathcal{U}/\mathsf{k}})^{1-p} \hookrightarrow (\omega_{L_{\boldsymbol{x}}/\mathsf{k}})^{1-p} \quad \text{and} \quad \mathrm{Hom}_{\mathcal{U}}(\mathcal{U}^{1/p}, \mathcal{U}) \hookrightarrow \mathrm{Hom}_{L_{\boldsymbol{x}}}(L_{\boldsymbol{x}}^{1/p}, L_{\boldsymbol{x}}).$$

It suffices to check that μ^{1-p} is sent to the standard splitting of $L_{\boldsymbol{x}}$; this is essentially [Brion and Kumar 2005, Lemma 1.3.6]. □

Acknowledgements

We are grateful to MSRI for financial support and for organizing the workshop in August 2012 designed to foster interactions between commutative algebra and cluster algebras, where this collaboration was born.

We are also indebted to Allen Knutson and David Speyer, who graciously shared their unpublished work on Frobenius splittings of cluster algebras.

References

[Berenstein et al. 2005] A. Berenstein, S. Fomin, and A. Zelevinsky, "Cluster algebras III: upper bounds and double Bruhat cells", *Duke Math. J.* **126**:1 (2005), 1–52. MR 2005i:16065 Zbl 1135.16013

[Brion and Kumar 2005] M. Brion and S. Kumar, *Frobenius splitting methods in geometry and representation theory*, Progress in Mathematics **231**, Birkhäuser, Boston, 2005. MR 2005k:14104 Zbl 1072.14066

[Bruns and Herzog 1993] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, Cambridge Studies in Advanced Mathematics **39**, Cambridge University Press, 1993. MR 95h:13020 Zbl 0788.13005

[Elkik 1981] R. Elkik, "Rationalité des singularités canoniques", *Invent. Math.* **64**:1 (1981), 1–6. MR 83a:14003 Zbl 0498.14002

[Fock and Goncharov 2007] V. V. Fock and A. B. Goncharov, "Dual Teichmüller and lamination spaces", pp. 647–684 in *Handbook of Teichmüller theory. Vol. I*, edited by A. Papadopoulos, IRMA Lect. Math. Theor. Phys. **11**, Eur. Math. Soc., Zürich, 2007. MR 2008k:32033 Zbl 1162.32009

[Fomin and Zelevinsky 2002] S. Fomin and A. Zelevinsky, "Cluster algebras I: foundations", *J. Amer. Math. Soc.* **15**:2 (2002), 497–529. MR 2003f:16050 Zbl 1021.16017

[Fomin and Zelevinsky 2003] S. Fomin and A. Zelevinsky, "Y-systems and generalized associahedra", *Ann. of Math.* (2) **158**:3 (2003), 977–1018. MR 2004m:17010 Zbl 1057.52003

[Gekhtman et al. 2005] M. Gekhtman, M. Shapiro, and A. Vainshtein, "Cluster algebras and Weil–Petersson forms", *Duke Math. J.* **127**:2 (2005), 291–311. MR 2006d:53103 Zbl 1079.53124

[Hara and Watanabe 2002] N. Hara and K.-I. Watanabe, "F-regular and F-pure rings vs. log terminal and log canonical singularities", *J. Algebraic Geom.* **11**:2 (2002), 363–392. MR 2002k:13009 Zbl 1013.13004

[Hartshorne 1966] R. Hartshorne, *Residues and duality*, Lecture Notes in Mathematics **20**, Springer, Berlin-New York, 1966. MR 36 #5145 Zbl 0212.26101

[Hochster and Huneke 1988] M. Hochster and C. Huneke, "Tightly closed ideals", *Bull. Amer. Math. Soc.* (*N.S.*) **18**:1 (1988), 45–48. MR 89b:13003 Zbl 0674.13003

[Hochster and Huneke 1989] M. Hochster and C. Huneke, "Tight closure and strong F-regularity", pp. 119–133 in *Colloque en l'honneur de Pierre Samuel* (Orsay, 1987), Mém. Soc. Math. France (N.S.) **38**, Société Mathématique de France, Paris, 1989. MR 91i:13025 Zbl 0699.13003

[Kontsevich and Soibelman 2013] M. Kontsevich and Y. Soibelman, "Wall-crossing structures in Donaldson–Thomas invariants, integrable systems and mirror symmetry", preprint, 2013. arXiv 1303.3253

[Kunz 1969] E. Kunz, "Characterizations of regular local rings for characteristic p", *Amer. J. Math.* **91** (1969), 772–784. MR 40 #5609 Zbl 0188.33702

[Matherne and Muller 2014] J. Matherne and G. Muller, "Computing upper cluster algebras", *Int. Math. Res. Not.* **2014** (online publication March 2014).

[Muller 2012] G. Muller, "Skein algebras and cluster algebras of marked surfaces", preprint, 2012. arXiv 1204.0020

[Muller 2013] G. Muller, "Locally acyclic cluster algebras", *Adv. Math.* **233** (2013), 207–247. MR 2995670 Zbl 1279.13032

[Muller 2014] G. Muller, "$\mathcal{A} = \mathcal{U}$ for locally acyclic cluster algebras", *SIGMA Symmetry Integrability Geom. Methods Appl.* **10** (2014), Paper 094, 8. MR 3261850 Zbl 06359376

[Muller and Speyer 2014] G. Muller and D. E. Speyer, "Cluster algebras of Grassmannians are locally acyclic", preprint, 2014. arXiv 1401.5137

[Musiker et al. 2013] G. Musiker, R. Schiffler, and L. Williams, "Bases for cluster algebras from surfaces", *Compos. Math.* **149**:2 (2013), 217–263. MR 3020308 Zbl 1263.13024

[Schwede and Smith 2010] K. Schwede and K. E. Smith, "Globally F-regular and log Fano varieties", *Adv. Math.* **224**:3 (2010), 863–894. MR 2011e:14076 Zbl 1193.13004

[Schwede and Tucker 2012] K. Schwede and K. Tucker, "A survey of test ideals", pp. 39–99 in *Progress in commutative algebra 2*, edited by C. Francisco et al., Walter de Gruyter, Berlin, 2012. MR 2932591 Zbl 1254.13007

[Shen and Goncharov 2015] L. Shen and A. B. Goncharov, "Geometry of canonical bases and mirror symmetry", *Invent. Math.* (online publication February 2015).

[Smith 1997a] K. E. Smith, "F-rational rings have rational singularities", *Amer. J. Math.* **119**:1 (1997), 159–180. MR 97k:13004 Zbl 0910.13004

[Smith 1997b] K. E. Smith, "Vanishing, singularities and effective bounds via prime characteristic local algebra", pp. 289–325 in *Algebraic geometry Santa Cruz 1995*, edited by J. Kollár et al., Proc. Sympos. Pure Math. **62**, Amer. Math. Soc., Providence, RI, 1997. MR 99a:14026 Zbl 0913.13004

[Smith 2000] K. E. Smith, "Globally F-regular varieties: applications to vanishing theorems for quotients of Fano varieties", *Michigan Math. J.* **48** (2000), 553–572. MR 2001k:13007 Zbl 0994.14012

[Smith and Zhang 2015] K. E. Smith and W. Zhang, "Frobenius splitting in commutative algebra", in *Commutative algebra and noncommutative algebraic geometry*, vol. 1, Mathematical Sciences Research Institute Publications **67**, Cambridge University Press, New York, 2015.

[Speyer 2013] D. Speyer, "An infinitely generated upper cluster algebra", preprint, 2013. arXiv 1305.6867

Communicated by Bjorn Poonen
Received 2014-09-04 Revised 2015-01-12 Accepted 2015-03-18

abenitos@umich.edu *Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, United States*

morilac@umich.edu *Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, United States*

rajchgot@umich.edu *Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, United States*

kesmith@umich.edu *Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, United States*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.937

On an analytic version of Lazard's isomorphism

Georg Tamme

We prove a comparison theorem between locally analytic group cohomology and Lie algebra cohomology for locally analytic representations of a Lie group over a nonarchimedean field of characteristic 0. The proof is similar to that of van Est's isomorphism and uses only a minimum of functional analysis.

Introduction

In a seminal paper, Lazard [1965] established two basic theorems concerning the cohomology of a compact $\mathbb{Q}_p$-analytic Lie group G with Lie algebra $\mathfrak{g}$. Firstly, if V is a finite-dimensional $\mathbb{Q}_p$-vector space with continuous G-action, the natural map from locally analytic group cohomology $H^*_{\mathrm{an}}(G, V)$, defined in terms of locally analytic cochains, to continuous group cohomology $H^*_{\mathrm{cont}}(G, V)$ is an isomorphism. Secondly, there is a natural isomorphism between the direct limit $\operatorname{colim}_{G'\subset G} H^*_{\mathrm{cont}}(G', V)$, where G' runs through the system of open subgroups of G, and the Lie algebra cohomology $H^*(\mathfrak{g}, V)$. Hence, combining both, there is a natural isomorphism

$$\operatorname*{colim}_{G'\subset G \text{ open}} H^*_{\mathrm{an}}(G', V) \cong H^*(\mathfrak{g}, V). \tag{1}$$

The author is supported by the CRC 1085 *Higher Invariants* (Universität Regensburg).
MSC2010: primary 22E50; secondary 17B56, 22E35.
Keywords: Lazard isomorphism, locally analytic group cohomology, Lia algebra cohomology.

These results play an important role in arithmetic geometry, in particular in the theory of Galois representations, or in the study of p-adic regulators [Huber and Kings 2011].

At least for certain Lie groups, integral and K-analytic versions have been obtained by Huber, Kings, and Naumann [Huber et al. 2011] when K is a finite extension of $\mathbb{Q}_p$. The proofs are based on Lazard's original argument via continuous group cohomology, and are not easily accessible. A somewhat simplified proof has been given by Lechner [2012] using formal group cohomology.

On the other hand, the situation for a real Lie group G is much more transparent. The analogous result is van Est's isomorphism $H^*_d(G, V) \cong H^*(\mathfrak{g}, K; V)$, which relates differentiable group cohomology with relative Lie algebra cohomology for a maximal compact subgroup $K \subseteq G$. Its proof is based on the following observations: The quotient G/K is contractible, hence the de Rham complex $\Omega^*(G/K, V)$ with coefficients in a G-representation V is a resolution of V. Moreover, for any k, the space $\Omega^k(G/K, V)$ is G-acyclic. Hence, $H^*_d(G, V)$ is computed by the G-invariants of the complex $\Omega^*(G/K, V)$, which is precisely the Chevalley–Eilenberg complex computing relative Lie algebra cohomology $H^*(\mathfrak{g}, K; V)$.

It is a natural question whether a similar argument works in the nonarchimedean situation. In this note, we show that this is indeed the case. This gives a direct proof of the isomorphism (1) and generalizes it with respect to the ground field and the coefficients:

Theorem. *Let K be a nonarchimedean field of characteristic 0. Let G be a locally K-analytic Lie group and $G \to \mathrm{Aut}(V)$ a locally analytic representation on a barreled locally convex K-vector space. Denote by $\mathfrak{g}$ the K-Lie algebra of G. Then there are natural isomorphisms*

$$\operatorname*{colim}_{G' \subset G \text{ open}} H^*_{\mathrm{an}}(G', V) \cong H^*(\mathfrak{g}, V),$$

where the colimit is taken over the system of open subgroups of G.

The rough argument is as follows: The de Rham complex $\Omega^*(G, V)$ is a resolution of the locally constant V-valued functions on G. As in the real case, each $\Omega^k(G, V)$ is G-acyclic, hence the cohomology of the locally constant V-valued functions on G is isomorphic to the Lie algebra cohomology $H^*(\mathfrak{g}, V)$ (see Sections 3 and 4 for precise results). The theorem then follows by taking the direct limit over the open subgroups of G.

The proof also shows that, for compact G, one can recover the locally analytic group cohomology from the Lie algebra cohomology as the invariants under the natural G-action: $H^*_{\mathrm{an}}(G, V) \cong H^*(\mathfrak{g}, V)^G$ (see Corollary 21).

Moreover, we describe the comparison map between locally analytic group cohomology and Lie algebra cohomology explicitly on the level of complexes: It is given by differentiating locally analytic cocycles at 1 (see Section 5). As pointed out by the referee, one can use the methods of [Huber and Kings 2011] to show that, on cohomology groups, this comparison map agrees with the one studied by Lazard in the case that K is $\mathbb{Q}_p$ and V is finite-dimensional.

In order to apply usual arguments from homological algebra, we show, following [Flach 2008], that the locally analytic cochain cohomology groups can be interpreted as derived functors of the global section functor on a topos BG (Sections 1 and 2). The nice feature of this is that it gives a quick proof of the main results which requires only a minimum of functional analysis.

An alternative approach to the cohomology of locally analytic representations of Lie groups over finite extensions of $\mathbb{Q}_p$ is due to Kohlhaase [2011]. It is based on relative homological algebra. He obtains similar results under an additional assumption on the group, which, as he proves, is fulfilled in many cases. The cohomology groups he defines are finer than ours in the sense that they themselves carry a locally convex topology. In contrast to the groups we use, they do not always coincide with the cohomology groups defined in terms of locally analytic cochains.

I would like to thank the referee for useful remarks, in particular concerning the comparison of our isomorphism with Lazard's original one.

Notations and conventions. For the whole paper, we let K be a nonarchimedean field of characteristic 0, i.e., K is equipped with a nontrivial nonarchimedean absolute value $|\cdot|$ such that K is complete for the topology defined by $|\cdot|$. By a manifold we will always mean a paracompact, finite-dimensional locally K-analytic manifold. Note that, by [Schneider 2011, Corollary 18.8], any locally K-analytic Lie group is paracompact. For manifolds X, Y, we denote by $C^{\mathrm{an}}(X, Y)$ the set of locally K-analytic maps from X to Y. We will refer to them simply as analytic maps.

1. Locally analytic group cohomology

In this section, we describe the topos-theoretic approach to locally analytic group cohomology. We refer to [Flach 2008] for the case of continuous cohomology.

Denote by La the category of manifolds. We let Sh(La) be the category of sheaves on La with respect to the topology generated by open coverings. For this topology, every representable presheaf is a sheaf, hence we have the Yoneda embedding $y : \mathrm{La} \to \mathrm{Sh(La)}$.

Let G be a Lie group. Then $y(G)$ is a group object in Sh(La). The category of sheaves with a $y(G)$-action is a topos [SGA 4_1 1972, Exposé IV, 2.4], called the classifying topos of $y(G)$. It will be denoted by BG.[1] We denote its global section functor by $\Gamma : BG \to \mathrm{Set}$, $\Gamma(\mathcal{F}) = \mathrm{Hom}_{BG}(*, \mathcal{F}) = \mathcal{F}(*)^G$. Similarly, if X is an

[1]More precisely, we assume the existence of universes and only consider manifolds which are elements of a given universe $\mathcal{U}$. Then Sh(La) and BG are $\mathcal{V}$-topoi for a universe $\mathcal{V}$ with $\mathcal{U} \in \mathcal{V}$.

object of BG, we denote by $\Gamma(X,-) = \mathrm{Hom}_{BG}(X,-)$ the functor of sections over X. As usual, we define cohomology groups via the derived functors of the global section functor:

Definition 2. Let $\mathcal{A}$ be an abelian group object of BG. Then we set

$$H^i(G,\mathcal{A}) := (R^i\Gamma)(\mathcal{A}).$$

Example 3. Let V be a finite-dimensional K-vector space with a linear G-action such that the map $G \times V \to V$ defining the action is analytic. This induces an action $y(G) \times y(V) \to y(V)$, and hence $y(V)$ can naturally be considered as an element of BG. We have $\Gamma(y(V)) = V^G$. In the next section, we will show that the higher cohomology groups $H^i(G, y(V))$ coincide with the cohomology groups defined in terms of locally analytic cochains with coefficients in V.

As another example, let M be a continuous G-module, i.e., a topological abelian group equipped with a linear G-action such that $G \times M \to M$ is continuous. Then we have the sheaf of continuous M-valued functions $C(-,M)$ on La. It also carries a natural action by $y(G)$. It follows from Proposition 8 below that the groups $H^i(G, C(-,M))$ can be identified with the continuous cochain cohomology groups of M.

We want to describe the cohomology groups defined in Definition 2 in terms of a concrete complex. We begin with some general considerations.

Let $\mathcal{T}$ be a topos, and let $\mathcal{G}$ be a group object in $\mathcal{T}$. For objects $\mathcal{A},\mathcal{B}$ of $B\mathcal{G}$ the internal hom $\mathcal{H}om_{B\mathcal{G}}(\mathcal{A},\mathcal{B})$ is given as follows: The underlying object of $\mathcal{T}$ is $\mathcal{H}om_{\mathcal{T}}(\mathcal{A},\mathcal{B})$ and the action of $\mathcal{G}$ is given by the formula

$$(g\phi)(a) = g(\phi(g^{-1}a)).$$

Denote by $i : * \to \mathcal{G}$ the morphism from the trivial group in $\mathcal{T}$ to $\mathcal{G}$. It induces a geometric morphism of topoi (see [SGA 4_1 1972, Exposé IV, 4.5])

$$i : \mathcal{T} \cong B* \to B\mathcal{G}.$$

The left adjoint i^* simply forgets the $\mathcal{G}$-action. The right adjoint is given by $i_*(\mathcal{F}) = \mathcal{H}om_{B\mathcal{G}}({}_l\mathcal{G},\mathcal{F})$, where ${}_l\mathcal{G}$ is $\mathcal{G}$ with its natural left action, viewed as an object ob $B\mathcal{G}$, and $\mathcal{F}$ is viewed as object of $B\mathcal{G}$ with trivial $\mathcal{G}$-action. The functor i^* also has a left adjoint $i_!$ given by $\mathcal{F} \mapsto {}_l\mathcal{G} \times \mathcal{F}$ with $\mathcal{G}$-action via the first factor.

For an object $\mathcal{A} \in B\mathcal{G}$, we denote by $\mathcal{A}^\natural$ the object of $B\mathcal{G}$ with the same underlying object in $\mathcal{T}$ and trivial $\mathcal{G}$-action.

Lemma 4. *For $\mathcal{A},\mathcal{B} \in B\mathcal{G}$ we have*

$$\mathcal{H}om_{B\mathcal{G}}({}_l\mathcal{G} \times \mathcal{A},\mathcal{B}) \cong i_*\,\mathcal{H}om_{\mathcal{T}}(i^*\mathcal{A}, i^*\mathcal{B}).$$

Proof. Let $\mathscr{X}$ be an object of $B\mathscr{G}$. Then we have natural isomorphisms

$$\begin{aligned}\mathrm{Hom}_{B\mathscr{G}}(\mathscr{X}, i_* \mathscr{H}om_{\mathscr{T}}(i^*\mathscr{A}, i^*\mathscr{B})) &\cong \mathrm{Hom}_{\mathscr{T}}(i^*\mathscr{X}, \mathscr{H}om_{\mathscr{T}}(i^*\mathscr{A}, i^*\mathscr{B}))\\ &\cong \mathrm{Hom}_{\mathscr{T}}(i^*(\mathscr{X}\times\mathscr{A}), i^*\mathscr{B})\\ &\cong \mathrm{Hom}_{B\mathscr{G}}(i_! i^*(\mathscr{X}\times\mathscr{A}), \mathscr{B})\\ &\cong \mathrm{Hom}_{B\mathscr{G}}({}_l\mathscr{G}\times(\mathscr{X}\times\mathscr{A})^\natural, \mathscr{B})\\ &\cong \mathrm{Hom}_{B\mathscr{G}}({}_l\mathscr{G}\times\mathscr{X}\times\mathscr{A}, \mathscr{B})\\ &\cong \mathrm{Hom}_{B\mathscr{G}}(\mathscr{X}, \mathscr{H}om_{B\mathscr{G}}({}_l\mathscr{G}\times\mathscr{A}, \mathscr{B})),\end{aligned}$$

where we used the isomorphism ${}_l\mathscr{G}\times(\mathscr{X}\times\mathscr{A})^\natural \cong {}_l\mathscr{G}\times\mathscr{X}\times\mathscr{A}$ given by $(\mathrm{pr}_1, \mathrm{action})$. This implies the lemma. □

We now consider the case $\mathscr{T} = \mathrm{Sh}(\mathrm{La}), \mathscr{G} = y(G)$. For a sheaf $\mathscr{F}$ on La, the sheaf underlying $i_*\mathscr{F}$ is, by the above, given by $X \mapsto \mathscr{H}om_{\mathrm{Sh}(\mathrm{La})}(y(G), \mathscr{F})(X) \cong \mathscr{F}(G\times X)$ (Yoneda lemma).

Remark 5. By our general assumption, every manifold X in La is paracompact. By [Schneider 2011, Proposition 8.7], it is even *strictly paracompact*, i.e., every open covering of X can be refined by a covering by pairwise-disjoint open subsets. This implies in particular that the functor of sections over X is exact on the category of abelian sheaves on La.

Lemma 6. *The functor i_* from abelian sheaves on* La *to abelian group objects in BG is exact.*

Proof. Since i_* is a right adjoint, it is left exact. Consider an epimorphism $\mathscr{A}\to\mathscr{B}$ of abelian sheaves on La. By the above remark, the functor of sections over $G\times X$ is exact, and hence $\mathscr{A}(G\times X)\to\mathscr{B}(G\times X)$ is an epimorphism of abelian groups. From this we deduce that $i_*\mathscr{A}\to i_*\mathscr{B}$ is an epimorphism. □

Corollary 7. *For any abelian sheaf $\mathscr{A}$ on* La, *we have*

$$H^i(G, i_*\mathscr{A}) \cong \begin{cases}\mathscr{A}(*) & \text{if } i=0,\\ 0 & \text{else.}\end{cases}$$

Proof. Since the left adjoint i^* is exact, i_* sends injectives to injectives. Since i_* is exact and $\Gamma\circ i_*(-) = \mathrm{Hom}_{BG}(*, i_*(-)) \cong \mathrm{Hom}_{\mathrm{Sh}(\mathrm{La})}(*, -)$ we see that

$$H^i(G, i_*\mathscr{A}) \cong H^i_{\mathrm{Sh}(\mathrm{La})}(*, \mathscr{A}) \cong \begin{cases}\mathscr{A}(*) & \text{if } i=0,\\ 0 & \text{else.}\end{cases} \qquad \square$$

We let $E_\bullet G$ be the simplicial manifold given in degree p by $E_pG := G^{p+1}$, and $\phi^* : E_qG \to E_pG$, for $\phi : \{0<\cdots<p\}\to\{0<\cdots<q\}$, given by

$$(g_0,\ldots,g_q)\mapsto(g_{\phi(0)},\ldots,g_{\phi(p)}).$$

The group G acts on $E_\bullet G$ via diagonal left multiplication. We denote the simplicial object of BG given by $y(E_\bullet G)$ equipped with diagonal $y(G)$-action by ${}_lE_\bullet G$.

For an abelian group object $\mathcal{A}$ in BG, the degree-wise sections over ${}_lE_\bullet G$ form a cosimplicial abelian group $\Gamma({}_lE_\bullet G, \mathcal{A})$.

Proposition 8. *Let $\mathcal{A}$ be an abelian group object of BG. Then*

$$H^*(G, \mathcal{A}) \cong H^*(\Gamma({}_lE_\bullet G, \mathcal{A})).$$

Proof. The projection ${}_ly(G) \to *$ is an epimorphism in BG. The Čech nerve of this morphism is precisely ${}_lE_\bullet G$. We thus have a quasi-isomorphism

$$\mathcal{A} \cong \mathcal{H}om_{BG}(*, \mathcal{A}) \xrightarrow{\sim} \mathcal{H}om_{BG}({}_lE_\bullet G, \mathcal{A}).$$

Using Lemma 4 and Corollary 7 we see that the complex on the right-hand side consists of Γ-acyclic objects. We conclude using $\Gamma(\mathcal{H}om_{BG}({}_lE_\bullet G, \mathcal{A})) = \mathrm{Hom}_{BG}(*, \mathcal{H}om_{BG}({}_lE_\bullet G, \mathcal{A})) \cong \mathrm{Hom}_{BG}({}_lE_\bullet G, \mathcal{A}) = \Gamma({}_lE_\bullet G, \mathcal{A})$. □

2. Locally analytic representations

In Example 3, we saw how to associate an abelian group object of BG to any finite-dimensional locally analytic representation of G. The goal of this section is to extend this to arbitrary locally analytic representations in the sense of [Schneider and Teitelbaum 2002], and to relate the cohomology groups thus defined to the cohomology groups defined in terms of analytic cochains (Proposition 12).

We begin by recalling some basic notions about analytic functions and locally analytic representations. References are [Féaux de Lacroix 1998] or [Schneider and Teitelbaum 2002, §§2,3].

If W is a Banach space, a function $f : X \to W$ from a manifold X to W is called analytic, if, in local charts, it is given by convergent power series with coefficients in W. The reader who is only interested in representations on Banach spaces can skip the following general definition and also all "admissibility" considerations later on. Let V be a locally convex separated K-vector space. A *BH-space*[2] for V is a continuous inclusion of a separated Banach space $W \hookrightarrow V$ [Féaux de Lacroix 1998, §1.2]. Let X be a manifold. A function $f : X \to V$ is called *analytic* if every $x \in X$ admits a neighborhood U and a BH-space $W \hookrightarrow V$ such that $f|_U$ factors through an analytic map $U \to W$. We denote the set of all analytic functions $X \to V$ by $C^{\mathrm{an}}(X, V)$. By [Féaux de Lacroix 1998, Korollar 2.4.4], $C^{\mathrm{an}}(X, V)$ is a module over the algebra of analytic functions $C^{\mathrm{an}}(X, K)$. For varying X, this is a sheaf on La denoted by $C^{\mathrm{an}}(-, V)$.

[2] Banach–Hausdorff

A *topological representation* of the Lie group G on V is an action of G on V by continuous automorphisms, i.e., a homomorphism $G \to \mathrm{Aut}(V)$ to the abstract group of continuous automorphisms $\mathrm{Aut}(V)$ of V. The topological representation is called *locally analytic* if all orbit maps $G \to V, g \mapsto gv$, are analytic (this is called a weakly analytic representation in [Féaux de Lacroix 1998, Definition 3.1.5]).

Example 9. Let W be a Banach space. Then $\mathrm{Aut}(W)$ is an open subset of the Banach space of continuous endomorphisms $\mathrm{End}(W)$. Féaux de Lacroix [1998, Korollar 3.1.9] showed that a topological representation of G on W is locally analytic if and only if the corresponding homomorphism $r : G \to \mathrm{Aut}(W) \subset \mathrm{End}(W)$ is analytic. Assume that this is the case. Let X be a manifold, and let $\rho : X \to G$ and $f : X \to W$ be analytic maps. Then the point-wise product $\rho f : X \to W$, $x \mapsto \rho(x) f(x)$, is again analytic. Indeed, ρf equals the composition

$$X \xrightarrow{(\rho,\mathrm{id}_X)} G \times X \xrightarrow{r\times f} \mathrm{Aut}(W) \times W \subseteq \mathrm{End}(W) \times W \xrightarrow{\mathrm{ev}} W,$$

where the first two maps are analytic by assumption and the evaluation map ev is continuous and bilinear. It follows that we get an action of $y(G)$ on the sheaf $C^{\mathrm{an}}(-, W)$, and $C^{\mathrm{an}}(-, W)$ can naturally be viewed as an object of BG. This generalizes Example 3.

For a general locally analytic representation of G on V, this need no longer be true. Instead, we have to consider the subsheaf of $C^{\mathrm{an}}(-, V)$ of admissible functions, as we explain in the next paragraph. The example above shows that for representations on Banach spaces every analytic function is admissible.

Let $G \to \mathrm{Aut}(V)$ be a topological representation. We call an analytic function $f : X \to V$ *admissible* if the map $\hat{f} : G \times X \to V$, $(g, x) \mapsto gf(x)$ is analytic. Note that $\hat{f}$ is analytic if and only if its restriction $\hat{f}|_{U\times X}$ for some open subset $U \subset G$ is analytic. Indeed, if this is the case, then for any $h \in G$ the restriction $\hat{f}|_{hU\times X}$ is equal to the composition

$$(hU \times X) \xrightarrow{(g,x)\mapsto(h^{-1}g,x)} U \times X \xrightarrow{\hat{f}|_{U\times X}} V \xrightarrow{h\cdot} V,$$

where the first two maps are analytic and the last is continuous and linear. We define

$$\underline{V}(X) := C^{\mathrm{ad}}(X, V) := \{f \in C^{\mathrm{an}}(X, V) \mid f \text{ is admissible}\}.$$

This is a $C^{\mathrm{an}}(X, K)$-submodule of $C^{\mathrm{an}}(X, V)$. We claim that $\underline{V}$ is a subsheaf of $C^{\mathrm{an}}(-, V)$ and that the point-wise multiplication by G-valued analytic maps defines an action of $y(G)$ on $\underline{V}$. We will henceforth view $\underline{V}$ as an abelian group object of BG.

Proof of the claim. If $f \in \underline{V}(X)$ and $\phi : Y \to X$ is an analytic map between manifolds, then $f \circ \phi$ is analytic. Moreover, $\widehat{f \circ \phi} = \hat{f} \circ (\mathrm{id}_G \times \phi)$ is analytic as well, hence $f \circ \phi$ is admissible. Thus $\underline{V}$ is a presheaf. Since admissibility is a local property, $\underline{V}$ is a sheaf.

Now let $\rho : X \to G$ be an analytic map. We define ρf by $(\rho f)(x) := \rho(x) f(x)$. We have to show that ρf is analytic and admissible. But this is clear since ρf equals the composition

$$X \xrightarrow{(\rho, \mathrm{id}_X)} G \times X \xrightarrow{\hat{f}} V$$

and $\widehat{\rho f}$ equals the composition

$$G \times X \xrightarrow{(g,x) \mapsto (g\rho(x),x)} G \times X \xrightarrow{\hat{f}} V. \qquad \square$$

Example 10. A topological representation of G on V is locally analytic if and only if $\underline{V}(*) = V$.

Definition 11. For a locally analytic representation of G on V and $i \geq 0$ we define the locally analytic group cohomology of G with coefficients in V as

$$H^i_{\mathrm{an}}(G, V) := H^i(G, \underline{V}).$$

Recall that a homogeneous analytic p-cochain of G with coefficients in V is an analytic function $f : E_p G \to V$ which is G-equivariant, i.e., satisfies $f(gg_0, \dots, gg_p) = gf(g_0, \dots g_p)$. We denote the complex of homogeneous analytic cochains by $C^{\mathrm{an}}_G(E_\bullet G, V)$. Its differential is induced by the simplicial structure of $E_\bullet G$.

Proposition 12. *The cohomology $H^*_{\mathrm{an}}(G, V)$ is isomorphic to the cohomology of the complex of homogeneous analytic cochains $C^{\mathrm{an}}_G(E_\bullet G, V)$.*

Proof. By Proposition 8 we have $H^*(G, \underline{V}) \cong H^*(\Gamma({}_lE_\bullet G, \underline{V}))$. Using the Yoneda lemma we see that a section in $\Gamma({}_lE_p G, \underline{V}) = \mathrm{Hom}_{BG}({}_lE_p G, \underline{V})$ is just an admissible function $f : E_p G \to V$ such that

$$\begin{tikzcd}
G \times E_p G \arrow[r, "\mathrm{id} \times f"] \arrow[d, "\text{diagonal multiplication}"'] & G \times V \arrow[d, "\text{action}"] \\
E_p G \arrow[r, "f"] & V
\end{tikzcd}$$

commutes, i.e., a G-equivariant admissible function $E_p G \to V$.

To prove the claim, it suffices to show that, vice versa, every G-equivariant analytic function $f : E_pG \to V$ is admissible. But, by the G-equivariance, $\hat{f}$ is equal to the composition

$$G \times E_pG \xrightarrow{(g,(g_0,\ldots,g_p)) \mapsto (gg_0,\ldots,gg_p)} E_pG \xrightarrow{f} V$$

and thus analytic. Thus f is admissible. $\square$

3. Differential forms and Lie algebra cohomology

In this section, we introduce sheaves of differential forms with coefficients in a locally analytic representation V of G as objects of BG. Again, unless V is a Banach space, we have to restrict to admissible forms. We show that the Lie algebra cohomology of the K-Lie algebra $\mathfrak{g}$ of G with coefficients in V can be computed as the cohomology in BG of the complex of forms on G with coefficients in V.

Let V be a separated locally convex K-vector space. For a submersion $Y \to X$ we denote by $\Omega^k(Y/X, V)$ the vector space of relative analytic k-forms with values in V. Here, a k-form ω is called analytic if for any k-tuple $\phi_1, \ldots, \phi_k$ of local sections of the vertical tangent bundle $T(Y/X)$ the function $Y \to V$, $y \mapsto \omega(y)(\phi_1(y), \ldots, \phi_k(y))$ is analytic. It suffices to check this for the local sections given by a local frame of $T(Y/X)$. In particular, every $y \in Y$ admits a neighborhood U and a BH-space $W \hookrightarrow V$ such that $\omega|_U$ is in the image of $\Omega^k(U/X, W) \hookrightarrow \Omega^k(U/X, V)$. It follows that the exterior derivative $d\omega$ is a well-defined form in $\Omega^{k+1}(Y/X, V)$. If V is finite-dimensional, this is the usual notion of analytic forms.

For a fixed manifold Y, we have a complex of sheaves $\underline{\Omega}^*(Y, V)$ on La defined by

$$\underline{\Omega}^*(Y, V)(X) := \Omega^*(X \times Y/X, V).$$

Let V be a locally analytic representation of G. We would like to equip this complex with a $y(G)$-action. As for functions, we have to restrict to a subcomplex of admissible forms in order to do this. Again, one can ignore this, if one is only interested in the case that V is a Banach space.

A form $\omega \in \Omega^k(Y/X, V)$ is called *admissible* if the form $\hat{\omega}$ on $G \times Y/G \times X$ given by

$$\hat{\omega}(g, y)(\mathrm{x}_1, \ldots, \mathrm{x}_k) := g \cdot (\omega(y)(\mathrm{x}_1, \ldots, \mathrm{x}_k)),$$

where $g \in G$, $y \in Y$ and $\mathrm{x}_i \in T_{(g,y)}(G \times Y/G \times X) \cong T_y(Y/X)$, is analytic. Equivalently, ω is admissible if and only if for any k-tuple of local sections $\phi_1, \ldots, \phi_k$ of $T(Y/K)$ as above the function $\omega(\phi_1, \ldots, \phi_k)$ is admissible. As in the case of functions, this is the case if and only if $\hat{\omega}|_{U \times Y}$ is analytic for some open subset $U \subseteq G$. The admissible k-forms form a $C^{\mathrm{an}}(Y, K)$-submodule of $\Omega^k(Y/X, V)$, which we denote by $\Omega^k_{\mathrm{ad}}(Y/X, V)$. They are also stable under the exterior derivative: Let ω be an admissible form. Since G acts on V by continuous linear automorphisms we have $\widehat{d\omega} = d\hat{\omega}$, and this form is analytic. Thus, the admissible forms $\Omega^*_{\mathrm{ad}}(Y/X, V)$ form a subcomplex of the de Rham complex $\Omega^*(Y/X, V)$.

Example 13. If V is a Banach space, it follows from Example 9 that any V-valued analytic k-form is admissible.

We now fix a manifold Y. For an analytic map between manifolds $X' \to X$, the pullback map $\Omega^k(X \times Y/X) \to \Omega^k(X' \times Y/X')$ preserves admissible forms. Since admissibility is a local condition, $\underline{\Omega}^k_{\mathrm{ad}}(Y, V)$, defined by

$$\underline{\Omega}^k_{\mathrm{ad}}(Y, V)(X) := \Omega^k_{\mathrm{ad}}(X \times Y/X, V),$$

is a subsheaf of $\underline{\Omega}^k(Y, V)$, and $\underline{\Omega}^*_{\mathrm{ad}}(Y, V) \subseteq \underline{\Omega}^*(Y, V)$ is a subcomplex.

We define an action of $y(G)$ on $\underline{\Omega}^k_{\mathrm{ad}}(G, V)$ as follows: Let $\rho : X \to G$ be an analytic map and $\omega \in \Omega^k_{\mathrm{ad}}(X \times G/X, V)$ an admissible form. For any $h \in G$, left translation by h^{-1} induces a map $T_{(x,g)}(X \times G/X) \to T_{(x,h^{-1}g)}(X \times G/X)$, written $\mathsf{x} \mapsto h^{-1}\mathsf{x}$. Using this notation, we define $\rho\omega$ by the formula

$$(\rho\omega)(x, g)(\mathsf{x}_1, \dots, \mathsf{x}_k) := \rho(x) \cdot \big(\omega(x, \rho(x)^{-1}g)(\rho(x)^{-1}\mathsf{x}_1, \dots, \rho(x)^{-1}\mathsf{x}_k)\big).$$

Lemma 14. *This is a well-defined $y(G)$-action.*

Proof. We have to show that $\rho\omega$ is analytic and admissible. Consider the analytic maps $\hat{\rho} : G \times X \times G \to G \times X \times G, (h, x, g) \mapsto (h\rho(x), x, \rho(x)^{-1}g)$ and $i_1 : X \times G \hookrightarrow G \times X \times G, (x, g) \mapsto (1, x, g)$. Then $\rho\omega = i_1^* \hat{\rho}^* \hat{\omega}$; hence $\rho\omega$ is analytic. Similarly, we have $\widehat{\rho\omega} = \hat{\rho}^* \hat{\omega}$, and hence $\rho\omega$ is admissible. □

We thus consider $\underline{\Omega}^k_{\mathrm{ad}}(G, V)$ as an abelian group object in BG. We want to show that it is acyclic. Write $\widetilde{V} := \mathrm{Hom}(\bigwedge^k \mathfrak{g}, V)$. The adjoint action of G on $\mathfrak{g}$ and the given action of G on V induce a natural G-action on $\widetilde{V}$.

Lemma 15. *This representation of G on $\widetilde{V}$ is locally analytic. We have an isomorphism*

$$\underline{\Omega}^k_{\mathrm{ad}}(G, V) \cong \mathcal{H}om_{BG}({}_l y(G), \underline{\widetilde{V}}) \cong i_* i^* \underline{\widetilde{V}}.$$

Proof. Let Y be a manifold. We claim that a function $f : Y \to \widetilde{V}$ is admissible if and only if the function $f_{\mathsf{x}} : Y \to V, y \mapsto f(y)(\mathsf{x})$ is admissible for every $\mathsf{x} \in \bigwedge^k \mathfrak{g}$. Taking Y to be a point this implies the first assertion of the lemma.

To prove the claim, assume first that f is admissible. We have $\widehat{f_{\mathsf{x}}}(g, y) = g(f(y)(\mathsf{x})) = (gf(y))(g\mathsf{x}) = \hat{f}(g, y)(g\mathsf{x})$. The function $\hat{f}$ is analytic by assumption and so is $g \mapsto g\mathsf{x}$. Since the evaluation $\widetilde{V} \times \bigwedge^k \mathfrak{g} \to V$ is continuous and bilinear, and since $\bigwedge^k \mathfrak{g}$ is finite dimensional, [Féaux de Lacroix 1998, Satz 2.4.3] implies that $\widehat{f_{\mathsf{x}}}$ is analytic.

To see the converse, let $\mathsf{x}_1, \dots, \mathsf{x}_N$ be a basis of $\bigwedge^k \mathfrak{g}$ and $\mathsf{x}_1^*, \dots, \mathsf{x}_N^*$ the dual basis of $(\bigwedge^k \mathfrak{g})^\vee$. We can write f as a sum $f = \sum_{i=1}^N f_{\mathsf{x}_i} \mathsf{x}_i^*$ with f_{x_i} admissible. Then $\hat{f}(g, y) = \sum_i \widehat{f_{\mathsf{x}_i}}(g, y) g(\mathsf{x}_i^*)$ and, by [Féaux de Lacroix 1998, Satz 2.4.3] again, $\hat{f}$ is analytic.

We now prove the second assertion of the lemma. For any manifold X, right translations by elements of G induce a trivialization of the vertical tangent bundle $T(X \times G/X) \cong (X \times G) \times \mathfrak{g}$. This gives a natural isomorphism of vector spaces

$$\Omega^k(X \times G/X, V) \cong C^{\mathrm{an}}(X \times G, \widetilde{V}).$$

Using the above claim one sees that this isomorphism restricts to an isomorphism

$$\Omega^k_{\mathrm{ad}}(X \times G/X, V) \cong C^{\mathrm{ad}}(X \times G, \widetilde{V}).$$

Under this isomorphism, the $y(G)(X)$-action on the left-hand side corresponds to the action on the right-hand side induced by left translations on $X \times G$ and the action on $\widetilde{V}$ mentioned above. Using the isomorphism $C^{\mathrm{ad}}(X \times G, \widetilde{V}) \cong \mathcal{H}om_{BG}({}_l y(G), \underline{\widetilde{V}})(X)$, this gives the first isomorphism stated in the Lemma. The second follows immediately from Lemma 4. □

Corollary 16. *We have*

$$H^i(G, \underline{\Omega}^k_{\mathrm{ad}}(G, V)) \cong \begin{cases} \mathrm{Hom}_K(\bigwedge^k \mathfrak{g}, V) & \textit{if } i = 0, \\ 0 & \textit{else}. \end{cases}$$

Proof. By Lemma 15 and Corollary 7 the higher cohomology groups vanish, and

$$H^0(G, \underline{\Omega}^k_{\mathrm{ad}}(G, V)) \cong \underline{\widetilde{V}}(*) = \mathrm{Hom}_K(\textstyle\bigwedge^k \mathfrak{g}, V), \tag{17}$$

concluding the proof. □

Explicitly, this isomorphism is given by evaluating a form at $1 \in G$.

The differential d of the complex $\underline{\Omega}^*_{\mathrm{ad}}(G, V)$ is compatible with the $y(G)$-action. Hence we can view $\underline{\Omega}^*_{\mathrm{ad}}(G, V)$ as a complex in BG and we can compute its hypercohomology.

We now assume that V is barreled, i.e., that every closed convex absorbing subset is open (see [Schneider 2002, §6]). For example, any complete metrizable locally convex space, in particular any Banach space, is barreled. Differentiating the orbit maps $g \mapsto gv$ then induces an action of the Lie algebra $\mathfrak{g}$ on V [Féaux de Lacroix 1998, Sätze 3.1.3, 3.1.7].

Corollary 18. *We have natural isomorphisms*

$$H^i(G, \underline{\Omega}^*_{\mathrm{ad}}(G, V)) \cong H^i(\mathfrak{g}, V),$$

where the right-hand side is Lie algebra cohomology.

Proof. Corollary 16 gives an isomorphism

$$H^i(G, \underline{\Omega}^*_{\mathrm{ad}}(G, V)) \cong H^i(\mathrm{Hom}_K(\textstyle\bigwedge^* \mathfrak{g}, V)),$$

where the differential on $\mathrm{Hom}_K(\bigwedge^* \mathfrak{g}, V)$ is induced from the de Rham differential via (17). This is precisely the Chevalley–Eilenberg complex computing Lie algebra cohomology. □

4. Differential forms and locally analytic group cohomology

As before, we fix a locally analytic representation $G \to \mathrm{Aut}(V)$. In this section we use the Poincaré lemma to compare the hypercohomology of the complex of V-valued admissible forms with locally analytic group cohomology, and we give the proof of the theorem announced on page 938 in the introduction.

Fix a manifold Y. A function $f : Y \times X \to V$ will be called locally constant along Y if, for every $(y, x) \in Y \times X$, there exist open neighborhoods $Y' \subseteq Y$ of y and $X' \subseteq X$ of x such that $f|_{Y' \times X'}$ factors through the projection $Y' \times X' \to X'$. We define

$$\underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(Y, V)(X) := \{f \in C^{\mathrm{ad}}(X \times Y, V) \mid f \text{ is locally constant along } Y\}.$$

It is easy to see that $X \mapsto \underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(Y, V)(X)$ defines a sheaf on La.

Proposition 19. *The inclusion in degree* 0

$$\underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(Y, V) \to \underline{\Omega}^*_{\mathrm{ad}}(Y, V)$$

is a quasi-isomorphism.

If V is a Banach space, this is just the Poincaré lemma, and its usual proof works. For general locally convex V, it is a little bit more complicated, since we have to prove admissibility of primitives.

Proof. The map clearly induces an isomorphism on H^0, and it remains to show that $H^k(\underline{\Omega}^*_{\mathrm{ad}}(Y, V)) = 0$ for $k > 0$.

Let X be a manifold, and let ω be a closed form in $\Omega^k_{\mathrm{ad}}(X \times Y/X, V)$. We will show that there is an $\eta \in \Omega^{k-1}_{\mathrm{ad}}(X \times Y/X, V)$ such that $d\eta = \omega$. Since all manifolds are strictly paracompact, it is enough to construct such an η locally on X and Y (see Remark 5).

The rest of the proof uses some results and notations from the Appendix. It can be skipped on first reading. Since $d\hat{\omega} = \widehat{d\omega} = 0$, the form $\hat{\omega} \in \Omega^k(G \times X \times Y/G \times X, V)$ is closed. Replacing G be a small open neighborhood of 1 and using local charts, we may assume that there are multiradii $\delta \in \mathbb{R}^m_+$, $\epsilon \in \mathbb{R}^n_+$ such that $G \times X \cong B_\delta(0) \subset K^m$, $Y \cong B_\epsilon(0) \subset K^n$, and a BH-space $W \hookrightarrow V$ such that $\hat{\omega}$ is given by a power series in $F_\delta(\Omega^k_\epsilon(W))$ (see (25)). Choose a multiradius $\epsilon' < \epsilon$. The homotopy operator $h : \Omega^k_\epsilon(W) \to \Omega^{k-1}_{\epsilon'}(W)$ given by Lemma 26 induces an operator

$h : F_\delta(\Omega^k_\epsilon(W)) \to F_\delta(\Omega^{k-1}_{\epsilon'}(W))$. We set $\widetilde{\eta} := h(\hat{\omega})$. Hence $\widetilde{\eta}$ represents a relative analytic $(k-1)$-form on $G \times X \times Y'/G \times X$ with an open subset $Y' \subset Y$. Since $\hat{\omega}$ is closed, we have $d\widetilde{\eta} = \hat{\omega}|_{G\times X\times Y'}$.

For $g \in G$, let $i_g : X \times Y \to G \times X \times Y$ (and similarly with Y replaced by Y') be the inclusion $(x, y) \mapsto (g, x, y)$. We set $\eta := i_1^*\widetilde{\eta}$. Clearly, $d\eta = i_1^* d\widetilde{\eta} = i_1^*\hat{\omega} = \omega$. To prove that η is admissible, we show that $\hat{\eta} = \widetilde{\eta}$. Let $\Phi_g : V \to V$ be the continuous automorphism given by the action of g. We have to check that $i_g^*\widetilde{\eta} = \Phi_g \circ \eta$. By restriction, Φ_g induces a continuous isomorphism of BH-spaces $W \to g(W)$ (more precisely, we view W as a linear subspace of V and let $g(W)$ be its image under the action of $g \in G$ with Banach space structure induced from W via the linear isomorphism $\Phi_g|_W : W \xrightarrow{\sim} g(W)$). We have

$$\begin{aligned}
i_g^*\widetilde{\eta} &= i_g^*(h(\hat{\omega})) && \text{(by definition)}\\
&= h(i_g^*\hat{\omega}) && \text{(using (24) with } \Phi = h : \Omega^q_\epsilon(W) \to \Omega^{q-1}_{\epsilon'}(W))\\
&= h(\Phi_g \circ \omega) && \text{(definition of } \hat{\omega})\\
&= \Phi_g \circ h(\omega) && \text{(Lemma 27 for } \Phi = \Phi_g : W \to g(W))\\
&= \Phi_g \circ \eta && \text{(since } h(\omega) = h(i_1^*\hat{\omega}) = i_1^* h(\hat{\omega}) = i_1^*\hat{\eta} = \eta). \qquad \square
\end{aligned}$$

Proof of the Theorem. The sheaf $\underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(G, V)$ carries a natural $y(G)$-action induced by left translations on G and the given action on V. By Proposition 19 and Corollary 18 we have isomorphisms

$$H^*(G, \underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(G, V)) \cong H^*(G, \underline{\Omega}^*_{\mathrm{ad}}(G, V)) \cong H^*(\mathfrak{g}, V). \tag{20}$$

As in the proof of Proposition 12, Proposition 8 implies that $H^*(G, \underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(G, V))$ is the cohomology of the complex $C^{\mathrm{lc}}_G(G \times E_\bullet G, V)$ of G-equivariant analytic functions $G \times E_p G \to V$ that are locally constant along the first factor.

Since the open subgroups $G' \subseteq G$ form a fundamental system of neighborhoods of $1 \in G$ (see [Schneider 2011, Lemma 18.7]), we have an isomorphism

$$\operatorname*{colim}_{G'\subset G \text{ open}} C^{\mathrm{an}}_{G'}(E_\bullet G', V) \cong \operatorname*{colim}_{G'\subset G \text{ open}} C^{\mathrm{lc}}_{G'}(G' \times E_\bullet G', V).$$

Because taking the colimit over a directed system is exact, we see that

$$\operatorname*{colim}_{G'\subset G} H^*_{\mathrm{an}}(G', V) \to \operatorname*{colim}_{G'\subset G} H^*(G', \underline{C}^{\mathrm{lc}}(G', V))$$

is an isomorphism. Since the isomorphisms (20) are compatible with the restriction to open subgroups, the claim follows. $\square$

There is an additional action of G on $\underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(G,V)$ and on $\underline{\Omega}^*_{\mathrm{ad}}(G,V)$ induced by right translations on G. This action is compatible with the given $y(G)$-action. It induces a G-action on the cohomology groups. Via the isomorphism (20) this corresponds to the G-action on $H^*(\mathfrak{g},V)$ induced by the adjoint action on $\mathfrak{g}$ and left multiplication on V.

Corollary 21. *If G is compact, there is a natural isomorphism*

$$H^*_{\mathrm{an}}(G,V) \cong H^*(\mathfrak{g},V)^G.$$

Proof. Since G is compact, every open subgroup is of finite index and contains an open normal subgroup. If X is a compact manifold, every function in $\underline{C}^{\mathrm{lc}}_{\mathrm{ad}}(G,V)(X)$ factors through $G/H \times X$ for some open normal subgroup $H \trianglelefteq G$. Thus — using the notation from the previous proof —

$$\underline{C}^{\mathrm{lc}}_G(G \times E_\bullet G, V) = \operatorname*{colim}_{H \trianglelefteq G \text{ open}} C^{\mathrm{an}}_G(G/H \times E_\bullet G, V).$$

Since the colimit over a directed system is exact, this induces an isomorphism $H^*(\mathfrak{g},V) \cong \operatorname{colim}_{H \trianglelefteq G} H^*(C^{\mathrm{an}}_G(G/H \times E_\bullet G, V))$. Since each quotient G/H is finite, and taking invariants under a finite group is an exact functor on K-vector spaces with an action by that group, we get

$$\begin{aligned} H^*(\mathfrak{g},V)^G &\cong \operatorname*{colim}_{H \trianglelefteq G} H^*(C^{\mathrm{an}}_G(G/H \times E_\bullet G, V))^{G/H} \\ &\cong \operatorname*{colim}_{H \trianglelefteq G} H^*(C^{\mathrm{an}}_G(G/H \times E_\bullet G, V)^{G/H}) \\ &\cong \operatorname*{colim}_{H \trianglelefteq G} H^*(C^{\mathrm{an}}_G(E_\bullet G, V)) \cong H^*_{\mathrm{an}}(G,V). \end{aligned}$$ □

5. Explicit description of the comparison map

We want to describe an explicit map of complexes which induces the comparison map $H^*_{\mathrm{an}}(G,V) \to H^*(\mathfrak{g},V)$. Recall that $H^*_{\mathrm{an}}(G,V)$ is computed by the complex of homogeneous locally analytic cochains $C^{\mathrm{an}}_G(E_\bullet G,V)$, and that $H^*(\mathfrak{g},V)$ is computed by the complex of G-invariant admissible differential forms $\Omega^*_{\mathrm{ad}}(G,V)^G$.

For integers $p \geq 0$ and $0 \leq i \leq p$, we denote by d_i the partial exterior derivative in the direction of the $(i+1)$-th factor of the product $E_pG = G^{p+1}$. We denote by $\Delta_p : G \to E_pG$ the diagonal map. For $f \in C^{\mathrm{an}}(E_pG,V)$ we set

$$\Psi(f) := \Delta_p^*(d_1 d_2 \dots d_p f) \in \Omega^p(G,V).$$

Proposition 22. *The map Ψ induces a morphism of complexes $C^{\mathrm{an}}_G(E_\bullet G,V) \to \Omega^*_{\mathrm{ad}}(G;V)^G$, which agrees with the comparison map $H^*_{\mathrm{an}}(G,V) \to H^*(\mathfrak{g},V)$ on cohomology groups.*

Remark 23. Let us consider the special case that K is $\mathbb{Q}_p$ and V is finite dimensional. We want to indicate how the method of [Huber and Kings 2011] allows one to compare our map with Lazard's one. The space of functions $C^{\mathrm{an}}(E_pG, V)$ has topological generators of the form $f_0 \otimes \cdots \otimes f_p \otimes v$ with $f_i \in C^{\mathrm{an}}(G, K)$ and $v \in V$. For such a function we have

$$\Psi(f_0 \otimes \cdots \otimes f_p \otimes v) = f_0\, df_1 \wedge \cdots \wedge df_p \otimes v,$$

and its image in $\mathrm{Hom}(\bigwedge^p \mathfrak{g}, V)$ is given by $f_0(1)\, df_1(1) \wedge \cdots \wedge df_p(1) \otimes v$.

There is another simplicial model $\widetilde{E}_\bullet G$ for the universal G-bundle (see [Huber and Kings 2011, §4.4]), given by $\widetilde{E}_pG = E_pG$, but with face maps

$$\tilde{\partial}_i(g_0, \ldots, g_n) = \begin{cases} (g_0, \ldots, g_{i-1}, g_i g_{i+1}, g_{i+2}, \ldots, g_p) & \text{if } i = 0, \ldots p-1, \\ (g_0, \ldots, g_{p-1}) & \text{if } i = p. \end{cases}$$

The G-action on $\widetilde{E}_\bullet G$ is given by left multiplication on the first factor. There is a natural G-equivariant isomorphism $\widetilde{E}_\bullet G \cong E_\bullet G$. Huber and Kings show that Lazard's isomorphism (for G small enough) is induced by the map

$$\Phi : C_G^{\mathrm{an}}(\widetilde{E}_\bullet G, V) \to \mathrm{Hom}\big(\textstyle\bigwedge^\bullet \mathfrak{g}, V\big),$$
$$\Phi(f_0 \otimes \cdots \otimes f_p \otimes v) = f_0(1) df_1(1) \wedge \cdots \wedge df_p(1) \otimes v$$

(see [Huber and Kings 2011, Proposition 4.6.1]; this is formulated in the case of trivial coefficients, but can easily be adapted to our setting). The argument of [Huber and Kings 2011, Theorem 4.7.1] shows that the composition of Φ with the isomorphism $C_G^{\mathrm{an}}(E_\bullet G, V) \cong C_G^{\mathrm{an}}(E_\bullet G, V)$ is homotopic to Ψ, hence both maps agree on cohomology groups.

Proof of Proposition 22. From the proof of Proposition 8 we have the acyclic resolution $\underline{V} \xrightarrow{\sim} \mathcal{H}om_{BG}({}_lE_\bullet G, \underline{V})$. For a manifold X we have

$$\mathcal{H}om_{BG}({}_lE_\bullet G, \underline{V})(X) = C^{\mathrm{ad}}(X \times E_\bullet G, V)$$

with $y(G)$-action induced from left translations on $E_\bullet G$ and the action on V. We define $\Psi : C^{\mathrm{ad}}(X \times E_pG, V) \to \Omega^p_{\mathrm{ad}}(X \times G/X, V)$ by the same formula as above. We claim that this gives a morphism of complexes $\Psi : \mathcal{H}om_{BG}({}_lE_\bullet G, \underline{V}) \to \underline{\Omega}^*_{\mathrm{ad}}(G, V)$ in BG.

Proof of the claim. One checks without difficulty that Ψ is equivariant for the $y(G)$-action. Now consider $f \in C^{\mathrm{ad}}(E_pG \times X, V)$. Recall the face maps

$$\partial_i : E_{p+1}G \to E_pG, \quad (g_0, \ldots, g_{p+1}) \mapsto (g_0, \ldots, \widehat{g_i}, \ldots, g_{p+1}).$$

The differential of the complex $C^{\mathrm{ad}}(E_\bullet G \times X, V)$ maps f to

$$\sum_{i=0}^{p+1} (-1)^i \partial_i^* f.$$

Since $\partial_i^* f$ is constant along the $(i+1)$-th factor G, we have $d_i(\partial_i^* f) = 0$. Since the partial derivatives commute up to sign, it follows that

$$\begin{aligned}
\Psi\left(\sum_{i=0}^{p+1} (-1)^i \partial_i^* f\right) &= \Psi(\partial_0^* f)\\
&= \Delta_{p+1}^*(d_1 \dots d_{p+1}(\partial_0^* f))\\
&= \Delta_{p+1}^*(\partial_0^*(d_0 \dots d_p f))\\
&= \Delta_p^*(d_0 \dots d_p f)\\
&= \Delta_p^*(d(d_1 \dots d_p f))\\
&= d(\Delta_p^*(d_1 \dots d_p f))\\
&= d(\Psi(f)).
\end{aligned}$$

□

We thus have a commutative diagram

$$\begin{tikzcd}
\underline{V} \arrow[r, "\simeq"] \arrow[dr] & \mathcal{H}om_{BG}({}_lE_\bullet G, \underline{V}) \arrow[d, "\Psi"] \\
& \underline{\Omega}^*_{\mathrm{ad}}(G, V)
\end{tikzcd}$$

where the complexes on the right-hand side consist of acyclic sheaves. The proposition now follows by taking global sections. □

Appendix: The Poincaré lemma

Let W be a K-Banach space with norm $\|\cdot\|$. For a multiradius $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}_+^n$ we denote the space of ϵ-convergent power series in n variables $x = (x_1, \dots, x_n)$ with coefficients in W by $F_\epsilon(W)$:

$$F_\epsilon(W) := \left\{ \sum_{I \in \mathbb{N}_0^n} a_I x^I \mid a_I \in W, \|a_I\| \epsilon^I \xrightarrow{I \to \infty} 0 \right\}.$$

Equipped with the norm $\left\| \sum_I a_I x^I \right\|_\epsilon := \max_I \|a_I\| \epsilon^I$, this is again a Banach space.

Let $\Phi : W \to W'$ be a continuous linear map between Banach spaces. It induces a continuous linear map $F_\epsilon(W) \to F_\epsilon(W')$. Let $B_\epsilon(0) \subset K^n$ be the closed ball of radius ϵ around 0. For any $x \in B_\epsilon(0)$ we have the evaluation at x, written

$i_x^*: F_\epsilon(W) \to W$, and similarly for W'. Since Φ is continuous the diagram

$$\begin{tikzcd} F_\epsilon(W) \arrow[r, "\Phi"] \arrow[d, "i_x^*"'] & F_\epsilon(W') \arrow[d, "i_x^*"] \\ W \arrow[r, "\Phi"] & W' \end{tikzcd} \tag{24}$$

commutes.

For $q \geq 0$ we denote by $\Omega^q_\epsilon(W)$ the space of ϵ-convergent W-valued q-forms in n variables:

$$\Omega^q_\epsilon(W) := \textstyle\bigwedge^q_K (K^n)^\vee \otimes_K F_\epsilon(W).$$

Since $\bigwedge^q_K (K^n)^\vee$ is a finite-dimensional K-vector space, this is again a Banach space. The usual differential defines a continuous linear map $d : \Omega^q_\epsilon(W) \to \Omega^{q+1}_\epsilon(W)$.

There is a natural injection $\Omega^q_\epsilon(W) \hookrightarrow \Omega^q(B_\epsilon(0), W)$ into the space of locally analytic W-valued q-forms. It is compatible with the differential. More generally, if $\delta \in \mathbb{R}^m_+$ is a second multiradius, we can identify δ-convergent power series with coefficients in $\Omega^q_\epsilon(W)$ with relative W-valued forms:

$$F_\delta(\Omega^q_\epsilon(W)) \hookrightarrow \Omega^q(B_\delta(0) \times B_\epsilon(0)/B_\delta(0), W). \tag{25}$$

On the other hand, every relative q-form is in the image of (25) after shrinking δ and ϵ appropriately.

Let $\epsilon' \in \mathbb{R}^n_+$ be a multiradius which is component-wise strictly smaller than ϵ, written $\epsilon' < \epsilon$. There is a continuous inclusion $i : \Omega^q_\epsilon(W) \hookrightarrow \Omega^q_{\epsilon'}(W)$.

Lemma 26 (Poincaré lemma). *Let $\epsilon' < \epsilon$ and $q > 0$. Then there exists a bounded linear map*

$$h : \Omega^q_\epsilon(W) \to \Omega^{q-1}_{\epsilon'}(W)$$

such that $d \circ h + h \circ d = i$.

Proof. We have

$$\Omega^q_\epsilon(W) = \bigoplus_{1 \leq k_1 < \cdots < k_q \leq n} F_\epsilon(W)\, dx_{k_1} \ldots dx_{k_q}.$$

Set $C := \max_i(\epsilon_i/\epsilon'_i)$. By assumption we have $C > 1$. Hence, for integers $N \gg 0$ we have $|1/(N+q)| \leq C^N$. We define

$$h(x^I dx_{k_1} \ldots dx_{k_q}) := \frac{1}{|I|+q} \sum_{\alpha=1}^{q} (-1)^{\alpha-1} x^{I+e_{k_\alpha}} dx_{k_1} \ldots \widehat{dx_{k_\alpha}} \ldots dx_{k_q}.$$

and

$$h\Bigl(\sum a_I x^I dx_{k_1} \ldots dx_{k_q}\Bigr) := \sum a_I h(x^I dx_{k_1} \ldots dx_{k_q}).$$

Since
$$\left\|\frac{a_I}{|I|+q}\right\|\epsilon'^I \le \|a_I\|C^{|I|}\epsilon'^I \le \|a_I\|\epsilon^I \quad \text{for } |I| \gg 0,$$
it follows that the power series $\sum_I a_I/(|I|+q)x^{I+e_{k_\alpha}}$ is ϵ'-convergent, whence that h is well-defined, and also that h is a bounded linear operator.

By continuity, it is now enough to check the equality $dh+hd=i$ on monomials $x^I dx_{k_1}\dots dx_{k_q}$. Relabeling the coordinates, we may moreover assume that $(k_1,\dots,k_q)=(1,\dots,q)$. We have

$$\begin{aligned}
&dh(x^I dx_1\dots dx_q)\\
&\quad= d\left(\frac{1}{|I|+q}\sum_{\alpha=1}^{q}(-1)^{\alpha-1}x^{I+e_\alpha}dx_1\dots\widehat{dx_\alpha}\dots dx_q\right)\\
&\quad= \left(\frac{1}{|I|+q}\sum_{\alpha=1}^{q}(i_\alpha+1)x^I dx_1\dots dx_q\right)\\
&\qquad+\frac{1}{|I|+q}\sum_{\alpha=1}^{q}\sum_{\beta=q+1}^{n}(-1)^{\alpha-1}(-1)^{q-1}i_\beta x^{I+e_\alpha-e_\beta}dx_1\dots\widehat{dx_\alpha}\dots dx_q dx_\beta\\
&\quad= \frac{\left(\sum_{\alpha=1}^{q} i_\alpha\right)+q}{|I|+q}x^I dx_1\dots dx_q\\
&\qquad+\frac{1}{|I|+q}\sum_{\alpha=1}^{q}\sum_{\beta=q+1}^{n}(-1)^{\alpha+q}i_\beta x^{I+e_\alpha-e_\beta}dx_1\dots\widehat{dx_\alpha}\dots dx_q dx_\beta
\end{aligned}$$

and

$$\begin{aligned}
&hd(x^I dx_1\dots dx_q)\\
&\qquad= h\left((-1)^q\sum_{\beta=q+1}^{n} i_\beta x^{I-e_\beta}dx_1\dots dx_q dx_\beta\right)\\
&\qquad= \frac{(-1)^q}{|I|+q}\sum_{\alpha=1}^{q}\sum_{\beta=q+1}^{n}(-1)^{\alpha-1}i_\beta x^{I+e_\alpha-e_\beta}dx_1\dots\widehat{dx_\alpha}\dots dx_q dx_\beta\\
&\qquad\quad+\frac{(-1)^q}{|I|+q}\sum_{\beta=q+1}^{n}(-1)^q i_\beta x^I dx_1\dots dx_q\\
&\qquad= \frac{1}{|I|+q}\sum_{\alpha=1}^{q}\sum_{\beta=q+1}^{n}(-1)^{\alpha+q-1}i_\beta x^{I+e_\alpha-e_\beta}dx_1\dots\widehat{dx_\alpha}\dots dx_q dx_\beta\\
&\qquad\quad+\frac{\left(\sum_{\beta=q+1}^{n} i_\beta\right)}{|I|+q}x^I dx_1\dots dx_q
\end{aligned}$$

Thus, $(dh + hd)(x^I dx_1 \dots dx_k) = x^I dx_1 \dots dx_k$. This finishes the proof. □

Lemma 27. *Let $\Phi : W \to W'$ be a bounded linear map between Banach spaces. It induces a map $\Omega^q_\epsilon(W) \to \Omega^q_\epsilon(W')$, denoted by the same symbol. For $q > 0$ and $\epsilon' < \epsilon$, the diagram*

$$\begin{tikzcd}
\Omega^q_\epsilon(W) \arrow[r, "h"] \arrow[d, "\Phi"'] & \Omega^{q-1}_{\epsilon'}(W) \arrow[d, "\Phi"] \\
\Omega^q_\epsilon(W') \arrow[r, "h"] & \Omega^{q-1}_{\epsilon'}(W')
\end{tikzcd}$$

commutes.

Proof. This follows directly from the definitions. □

References

[Féaux de Lacroix 1998] C. T. Féaux de Lacroix, *Einige Resultate über die topologischen Darstellungen p-adischer Liegruppen auf unendlich dimensionalen Vektorräumen über einem p-adischen Körper*, Ph.D. thesis, Univ. Münster, 1998. pp. 1–111 in Schriftenreihe Math. Inst. Univ. Münster (3) **23**, 1999. MR 2000k:22021 Zbl 0963.22009

[Flach 2008] M. Flach, "Cohomology of topological groups with applications to the Weil group", *Compos. Math.* **144**:3 (2008), 633–656. MR 2009f:14033 Zbl 1145.18006

[Huber and Kings 2011] A. Huber and G. Kings, "A p-adic analogue of the Borel regulator and the Bloch–Kato exponential map", *J. Inst. Math. Jussieu* **10**:1 (2011), 149–190. MR 2012a:19010 Zbl 1243.11113

[Huber et al. 2011] A. Huber, G. Kings, and N. Naumann, "Some complements to the Lazard isomorphism", *Compos. Math.* **147**:1 (2011), 235–262. MR 2012d:22016 Zbl 1268.20051

[Kohlhaase 2011] J. Kohlhaase, "The cohomology of locally analytic representations", *J. Reine Angew. Math.* **651** (2011), 187–240. MR 2774315 Zbl 1226.22020

[Lazard 1965] M. Lazard, "Groupes analytiques p-adiques", *Inst. Hautes Études Sci. Publ. Math.* 26 (1965), 389–603. MR 35 #188 Zbl 0139.02302

[Lechner 2012] S. Lechner, "A comparison of locally analytic group cohomology and Lie algebra cohomology for p-adic Lie groups", preprint, 2012. arXiv 1201.4550

[Schneider 2002] P. Schneider, *Nonarchimedean functional analysis*, Springer, Berlin, 2002. MR 2003a:46106 Zbl 0998.46044

[Schneider 2011] P. Schneider, *p-adic Lie groups*, Grundlehren der Mathematischen Wissenschaften **344**, Springer, Heidelberg, 2011. MR 2012h:22010 Zbl 1223.22008

[Schneider and Teitelbaum 2002] P. Schneider and J. Teitelbaum, "Locally analytic distributions and p-adic representation theory, with applications to GL_2", *J. Amer. Math. Soc.* **15**:2 (2002), 443–468. MR 2003b:11132 Zbl 1028.11071

[SGA 4_1 1972] M. Artin, A. Grothendieck, and J. L. Verdier (editors), *Théorie des topos et cohomologie étale des schémas, Tome 1: Théorie des topos, Exposés I–IV* (Séminaire de Géométrie Algébrique du Bois Marie 1963–1964), Lecture Notes in Math. **269**, Springer, Berlin, 1972. MR 50 #7130 Zbl 0234.00007

Communicated by Marie-France Vignéras
Received 2014-09-09 Revised 2015-02-17 Accepted 2015-04-06

georg.tamme@ur.de *Fakultät für Mathematik, Universität Regensburg, D-93040 Regensburg, Germany*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.957

Towards local-global compatibility for Hilbert modular forms of low weight

James Newton

We prove some new cases of local-global compatibility for the Galois representations associated to Hilbert modular forms of low weight. If $F/\mathbb{Q}$ is a totally real extension of degree d, we are interested in Hilbert modular forms for F of weight $(k_1, \dots, k_d, w)$, with the k_i and w odd integers and some but not all k_i equal to 1 (the partial weight-one case). Recall that a Hecke eigenform f with such a weight has an associated compatible system $\rho_{f,p}$ of two-dimensional p-adic representations of $\mathrm{Gal}(\overline{F}/F)$, first constructed by Jarvis using congruences to forms of cohomological weight ($k_i \geq 2$ for all i).

One expects that the restriction of the representation $\rho_{f,p}$ to a decomposition group D_v at a finite place $v \nmid p$ of F should correspond (under the local Langlands correspondence) to the local factor at v, $\pi_{f,v}$, of the automorphic representation π_f generated by f. This expectation is what we refer to as local-global compatibility. For forms of cohomological weight, the compatibility was in most cases verified by Carayol using geometric methods. Combining this result with Jarvis's construction of Galois representations establishes many cases of local-global compatibility in the partial weight-one situation. However, when $\pi_{f,v}$ is a twist of the Steinberg representation, this method establishes a statement weaker that local-global compatibility. The difficulty in this case is to show that the Weil–Deligne representation associated to $\rho_{f,p}|_{D_v}$ has a nonzero monodromy operator. In this paper, we verify local-global compatibility in many of these 'missing' cases, using methods from the p-adic Langlands programme (including analytic continuation of overconvergent Hilbert modular forms, maps between eigenvarieties encoding Jacquet–Langlands functoriality and Emerton's completed cohomology).

1. Introduction

In this paper, we study the problem of local-global compatibility for the Galois representations attached to Hilbert modular forms of low weight. These Galois representations were constructed by Jarvis [1997]. We begin by recalling the main

MSC2010: primary 11F41; secondary 11F33, 11F80.
Keywords: Hilbert modular forms, Galois representations, local-global compatibility.

theorem of that paper. Let F be a totally real finite extension of $\mathbb{Q}$. For a prime p, let ι be an isomorphism from a fixed algebraic closure $\overline{\mathbb{Q}}_p$ of $\mathbb{Q}_p$ to $\mathbb{C}$. If v is a finite place of F and π_v is an admissible smooth irreducible representation of $\mathrm{GL}_2(F_v)$ over $\mathbb{C}$, we denote the Frobenius-semisimple Weil–Deligne representation associated to π_v under the (Hecke normalised) local Langlands correspondence by $\sigma(\pi_v)$. We denote by $\sigma^\iota(\pi_v)$ the Weil–Deligne representation with coefficients in $\overline{\mathbb{Q}}_p$ obtained by composition with ι^{-1}. For a continuous $\overline{\mathbb{Q}}_p$-representation ρ of $\mathrm{Gal}(\overline{F}_v/F_v)$, with $v \nmid p$ a finite place of F, we denote by $\mathrm{WD}(\rho)$ the Weil–Deligne representation associated to ρ by Grothendieck's p-adic monodromy theorem, with Frobenius semisimplification $\mathrm{WD}(\rho)^{F\text{-ss}}$ and semisimplification $\mathrm{WD}(\rho)^{\mathrm{ss}}$ (i.e., we set the monodromy operator to zero to obtain $\mathrm{WD}(\rho)^{\mathrm{ss}}$).

Results of Carayol [1986b], Blasius and Rogawski [1993], Rogawski and Tunnell [1983], Taylor [1989] and Jarvis [1997] have shown the following:

Theorem 1.1. *Let π be a cuspidal algebraic[1] automorphic representation of $\mathrm{GL}_2(\mathbb{A}_F)$, such that for each infinite place τ of F the local factor π_τ is either discrete series or holomorphic limit of discrete series. Then there exists an irreducible representation*

$$r_{p,\iota}(\pi) : \mathrm{Gal}(\overline{F}/F) \to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$$

such that if v is a finite place of F, with $v \nmid p$, and one of the following holds:

- *π_τ is discrete series for all infinite places τ,*
- *π_v is not special (i.e., π_v is not a twist of the Steinberg representation),*

then

$$\mathrm{WD}(r_{p,\iota}(\pi)|_{\mathrm{Gal}(\overline{F}_v/F_v)})^{F\text{-ss}} \cong \sigma^\iota(\pi_v).$$

Remark 1.2. The irreducibility of the Galois representation appearing in this theorem is proved using an argument of Ribet [1977] (see [Taylor 1995, Proposition 3.1]).

Remark 1.3. In the excluded case, where π_τ is holomorphic limit of discrete series for some τ and π_v is special, then Jarvis [1997, Remark 7.3] proved that

$$\mathrm{WD}(r_{p,\iota}(\pi)|_{\mathrm{Gal}(\overline{F}_v/F_v)})^{\mathrm{ss}} \cong \sigma^\iota(\pi_v)^{\mathrm{ss}}.$$

The main result of this paper addresses some of the excluded cases in Theorem 1.1. We prove:

Theorem 1.4. *Let π be as in Theorem 1.1, suppose that π_τ is holomorphic limit of discrete series for some infinite place τ and let $v \nmid p$ be a finite place of F such that π_v is special.*

Suppose the following technical hypotheses hold:

[1] By algebraic, we just mean that the usual parity condition on the weights holds.

(1) *The prime p is absolutely unramified in F.*

(2) *For every place w of F with $w \mid p$, π_w is an unramified principal series representation.*

(3) *Moreover, for each $w \mid p$ we have $\pi_w \cong \mathrm{Ind}(\mu_1 \otimes \mu_2)$ (normalised parabolic induction from a Borel subgroup) with μ_1, μ_2 distinct unramified characters of $F_w^\times$.*

(4) *The residual representation*
$$\bar{r}_{p,\iota}(\pi) : \mathrm{Gal}(\overline{F}/F) \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$$
is irreducible.

Then
$$\mathrm{WD}(r_{p,\iota}(\pi)|_{\mathrm{Gal}(\overline{F}_v/F_v)})^{F\text{-ss}} \cong \sigma^\iota(\pi_v).$$

Remark 1.5. We will not discuss the question of local-global compatibility at places dividing p in this paper. However, under the 'p-distinguished' hypothesis (3) made above, an argument using analytic continuation of crystalline periods (as done in [Jorza 2010] for the case of low-weight Siegel modular forms) shows that the representations $r_{p,\iota}(\pi)$ are crystalline at places dividing p, with the expected associated Weil–Deligne representation.

The main novelty in this theorem is that we are able to establish nontriviality of the monodromy operator for a Galois representation which is only indirectly related (by congruences) to the cohomology of Shimura varieties.

1.6. ***The technical hypotheses.*** We will say a few words about the hypotheses in Theorem 1.4. Hypotheses (1) and (2) are satisfied for all but finitely many p. Hypothesis (1) appears because we use results of Kassaei [2013] on analytic continuation and gluing of overconvergent Hilbert modular forms. At least for forms of parallel weight, results have been announced by Pilloni and Stroh [2013] and Sasaki [2014] which apply without restriction on the ramification of p, so it seems reasonable to hope that this hypothesis could be removed in future. Once hypothesis (1) is removed, hypothesis (2) can be substantially relaxed by first making a base change to an extension F'/F in which p ramifies.

Hypothesis (3) is again present in [Kassaei 2013] and has been removed by Pilloni and Stroh [2013] and Sasaki [2014] in their setting, but we also make some independent use of this condition. In particular, we use (3) to keep track of the distinct p-stabilisations when we apply p-adic functoriality to move between different kinds of p-adic automorphic form. By contrast, in [Pilloni and Stroh 2013; Sasaki 2014], the situation is that the various p-stabilisations form a single generalised eigenspace for the Hecke operators. So extending our techniques to this setting would require more care when applying p-adic functoriality.

We assume hypothesis (4) in order to apply the results of [Newton 2013] (it simplifies the analysis of completed cohomology of Shimura curves). It should be possible to avoid this assumption by modifying the arguments of that paper to work with overconvergent cohomology [Ash and Stevens 2008].

To summarise, we believe that the techniques described in this paper could in principle prove a version of the above theorem with just one technical hypothesis: for each place $w \mid p$ of F, π_w is not a twist of the Steinberg representation. To handle the remaining cases seems to require a new idea or a different method.

Remark 1.7. Recent work of Kassaei, Sasaki and Tian [Kassaei et al. 2014] permits us to relax hypotheses (2) and (3) slightly (allowing tame ramification). We have kept to the 'unramified' setting for expository reasons, but here is the precise condition which can replace hypotheses (2) and (3) in Theorem 1.4:

(2′) For every place w of F with $w \mid p$, $\pi_w \cong \mathrm{Ind}(\mu_1 \otimes \mu_2)$ with μ_1, μ_2 distinct characters of $F_w^\times$ whose ratio is tamely ramified.

1.8. ***Sketch of proof.*** We now sketch the proof of Theorem 1.4. By twisting and base change we may assume that $[F:\mathbb{Q}]$ is even, that π_v is an unramified twist of the Steinberg representation, and that we have an auxiliary finite place w such that π_w is also an unramified twist of Steinberg. In light of Remark 1.3 we just need to show that the Weil–Deligne representation associated to the local representation

$$\rho_v = r_{p,\iota}(\pi)|_{\mathrm{Gal}(\overline{F}_v/F_v)}$$

has a nonzero monodromy operator. Equivalently, we must show that ρ_v is not an unramified representation. We will assume that ρ_v is unramified and obtain a contradiction.

The auxiliary Steinberg place w allows us to find systems of Hecke eigenvalues attached to π in the p-adically completed cohomology of Shimura curves associated to indefinite quaternion algebras. In this context, a p-adic analogue of Mazur's principle [Newton 2013, Theorem 4.33] allows us to show that, since the representation ρ_v is unramified, we can strip v from the level of π. More precisely we can produce overconvergent Hilbert modular forms, with level prime to v, which share the same system of Hecke eigenvalues (outside v) as π. Finally, a generalisation, due to Kassaei, of Buzzard and Taylor's gluing and analytic continuation of overconvergent eigenforms allows us to produce a *classical* Hilbert modular form, with level prime to v, contributing to π. But we assumed that π_v was (a twist of) Steinberg, so π contains no nonzero $\mathrm{GL}_2(\mathbb{O}_{F_v})$-invariant vectors. Therefore we obtain the desired contradiction.

Readers familiar with the theory of p-adic and overconvergent automorphic forms may find it amusing that we make use of three different avatars of overconvergent automorphic forms in this paper — firstly geometrically defined overconvergent Hilbert modular forms (sections of automorphic line bundles on strict neighbourhoods of the ordinary locus in Hilbert modular varieties as in, for example, [Kisin and Lai 2005]), secondly the overconvergent automorphic forms on definite quaternion algebras defined by Buzzard [2007] and finally the spaces obtained by applying Emerton's locally analytic Jacquet functor to the p-adically completed cohomology of Shimura curves [Emerton 2006b; Newton 2013].

The organisation of our paper is as follows: In Section 2 we reduce the statement of Theorem 1.4 to a special case. Then, in Section 3, we describe the various eigenvarieties we will make use of, together with their construction. We next explain how to find systems of Hecke eigenvalues arising from the automorphic representations π of interest in the completed cohomology of Shimura curves, where we apply the level optimisation results of [Newton 2013] (Section 4). Finally we return to overconvergent modular forms on Hilbert modular varieties and apply a crucial result of Kassaei [2013] (Section 5).

1.9. ***Other approaches.*** Luu [2015, §3.2] has also recently described an approach to proving some cases of local-global compatibility for Hilbert modular forms of low weight when π satisfies an ordinarity hypothesis at places dividing p. We will use the same notation as in the previous subsection. Using the assumption that ρ_v is unramified, Luu applied a modularity lifting theorem to produce an ordinary p-adic Hilbert modular form g with level prime to v and the same system of Hecke eigenvalues (outside v) as π, and then imposed a hypothesis that amounts to ruling out the existence of g. It may also be possible to show that g is classical (hence obtaining a contradiction without an additional hypothesis) using a variant of the methods applied in the parallel weight-one case (as in, for example, [Kassaei 2013]), but it is not obvious to the author that these methods can be easily applied. One obstacle is that, unlike the parallel weight-one case, not all the p-stabilisations of the newform generating π are ordinary. The advantage of our method is that these p-stabilisations still have finite slope, so we can apply the level optimisation results of [Newton 2013] to obtain several overconvergent forms and then apply [Kassaei 2013].

1.10. ***Examples.*** Until recently, not many examples of Hilbert modular forms of partial weight one were known — in particular, as far as the author knows, the only known examples were CM forms (and therefore not twists of Steinberg at any place). However, Moy and Specter [2015] have recently explicitly computed examples which are non-CM and Steinberg at a finite place.

2. Notation and reductions

We let $F/\mathbb{Q}$ be a totally real, finite extension of $\mathbb{Q}$, and fix a prime p such that p is absolutely unramified in F. We denote by Σ the set of Archimedean places of F.

We write $\mathbb{A}_F$ for the adeles of F and $\mathbb{A}_F^\infty$ for the finite adeles. For a finite place v of F we write $\mathbb{O}_v$ for the ring of integers in the local field F_v. For convenience, we fix once and for all a choice of uniformiser $\varpi_v \in \mathbb{O}_v$ for each finite place v.

We denote by Σ_p the set of embeddings $F \hookrightarrow \overline{\mathbb{Q}}_p$, and for a place $\mathfrak{p} \mid p$ denote by $\Sigma_\mathfrak{p}$ those embeddings which extend to $F_\mathfrak{p}$. Our fixed isomorphism $\iota : \overline{\mathbb{Q}}_p \cong \mathbb{C}$ induces a bijection between Σ and Σ_p.

We will need some notation for (limits of) discrete series representations of $\mathrm{GL}_2(\mathbb{R})$. For $k \geq 2$ and w integers of the same parity, we let $D_{k,w}$ denote the discrete series representation of $\mathrm{GL}_2(\mathbb{R})$ with central character $t \mapsto t^{-w}$ defined in [Carayol 1986b, Section 0.2]. For $k = 1$ and w an odd integer, we define $D_{1,w}$ to be the limit of discrete series $\mathrm{Ind}(\mu, \nu)$, where the induction is a normalised parabolic induction from a Borel subgroup and μ, ν are characters of $\mathbb{R}^\times$ defined by $\mu(t) = |t|^{-w/2}\,\mathrm{sgn}(t)$ and $\nu(t) = |t|^{-w/2}$.

We assume π is an automorphic representation of $\mathrm{GL}_2(\mathbb{A}_F)$ as in the statement of Theorem 1.4, and that v is a finite place of F with π_v special. First we note that by twisting and making a quadratic base change (to an extension of F in which v splits) we can assume without loss of generality that $[F : \mathbb{Q}]$ is even, that π_v is an unramified twist of the Steinberg representation and that there is another finite place v' with $\pi_{v'}$ also an unramified twist of Steinberg.

With this in mind, for the rest of the paper we fix a cuspidal automorphic representation π_0 of $\mathrm{GL}_2(\mathbb{A}_F)$ as in the statement of Theorem 1.4, and moreover assume that:

- $[F : \mathbb{Q}]$ is even.
- $\pi_{0,v}$ is an unramified twist of Steinberg.
- There is a finite place v' coprime to p and v such that $\pi_{0,v'}$ is an unramified twist of Steinberg.
- $\pi_{0,\tau} \cong D_{k_\tau,w}$ for each $\tau \in \Sigma$ with k_τ an odd positive integer and w an odd integer (independent of τ).

2.1. ***Hilbert modular forms and varieties.*** We now proceed to reduce Theorem 1.4 to a statement about Hilbert modular forms. Let $L \subset \overline{\mathbb{Q}}_p$ be a finite extension of $\mathbb{Q}_p$ which contains the image of every embedding from F into $\overline{\mathbb{Q}}_p$. Let $\mathfrak{n}$ and $\mathfrak{m}$ be two coprime ideals of $\mathbb{O}_F$, both coprime to p, and suppose that $\mathfrak{n}$ is divisible by a rational integer which is ≥ 4.

Definition 2.2. Let $\mathfrak{c}$ be a fractional ideal of F such that its absolute norm $\boldsymbol{N}\mathfrak{c}$ is coprime to $\mathfrak{m}\mathfrak{n}$. We denote by $\underline{\mathscr{M}}^\mathfrak{c}(\mathfrak{n}, \mathfrak{m})$ the functor from schemes over $\mathbb{O}_L$ to sets taking S to isomorphism classes of tuples $(A, \lambda, [\eta])$, where:

- A is a Hilbert–Blumenthal abelian scheme over S (in particular A is equipped with an action of $\mathbb{O}_F$ [Kisin and Lai 2005, 1.1]).

- λ is a $\mathfrak{c}$-polarisation of A [Kisin and Lai 2005, 1.2].
- η is an $\mathcal{O}_F$-equivariant closed immersion of S-group schemes

$$\eta : (\mathcal{O}_F/\mathfrak{m}\mathfrak{n})^\vee \hookrightarrow A[\mathfrak{m}\mathfrak{n}],$$

and $[\eta]$ is its equivalence class under the natural action of $(\mathcal{O}_F/\mathfrak{m})^\times$. Here $(\mathcal{O}_F/\mathfrak{m}\mathfrak{n})^\vee$ denotes the Cartier dual of the constant S-group scheme (with $\mathcal{O}_F$ action) $\mathcal{O}_F/\mathfrak{m}\mathfrak{n}$.

Under our assumptions, the functor $\underline{\mathscr{M}}^{\mathfrak{c}}(\mathfrak{n},\mathfrak{m})$ is representable by a smooth $\mathcal{O}_L$-scheme $\mathscr{M}^{\mathfrak{c}}(\mathfrak{n},\mathfrak{m})$, and there is a good theory of toroidal and minimal compactifications (see [loc. cit., 1.6, 1.8] — although we use a slightly modified level structure, everything in this reference goes through).

Having fixed a choice of representatives $\mathfrak{c}_1,\dots,\mathfrak{c}_h$ for the narrow class group of F (which we may assume all satisfy $\mathbf{N}\mathfrak{c}_i$ coprime to $p\mathfrak{n}\mathfrak{m}$), we define $\mathscr{M}(\mathfrak{n},\mathfrak{m})$ to be the disjoint union of the $\mathscr{M}^{\mathfrak{c}_i}(\mathfrak{n},\mathfrak{m})$.

For κ a homomorphism

$$\kappa : \mathrm{Res}_{\mathbb{Z}}^{\mathcal{O}_F}\mathbb{G}_{m/\mathcal{O}_L} \to \mathbb{G}_{m/\mathcal{O}_L}$$

we have line bundles $\underline{\omega}^\kappa$ as in [loc. cit., 1.4.2] on $\mathscr{M}(\mathfrak{n},\mathfrak{m})$, which extend to suitable toroidal compactifications. We assume that w is an integer such that the character $\mathbf{N}^w\cdot\kappa^{-1}$ admits a square root. Then, following [loc. cit., 1.11], we define Hecke operators $T_{\mathfrak{a}}$ for each prime ideal $\mathfrak{a}$ of $\mathcal{O}_F$ coprime to $\mathfrak{m}\mathfrak{n}$ and $U_{\mathfrak{a}}$ for each prime ideal $\mathfrak{a}$ dividing $\mathfrak{m}\mathfrak{n}$.

Denote by $\mathbb{T}(\mathfrak{n},\mathfrak{m})$ the polynomial algebra over $\mathbb{Z}$ generated by symbols $T_{\mathfrak{a}}$ and $U_{\mathfrak{a}}$ as above. This algebra naturally acts on $H^0(\mathscr{M}(\mathfrak{n},\mathfrak{m})_L,\underline{\omega}^\kappa)$, and we say that a nonzero element f of this L-vector space is a *Hecke eigenform* if the action of $\mathbb{T}(\mathfrak{n},\mathfrak{m})$ preserves the one-dimensional space $L\cdot f$. If f is a cuspidal Hecke eigenform, we denote by θ_f the character $\mathbb{T}(\mathfrak{n},\mathfrak{m})\to L$ giving the action of the Hecke algebra on f. After projecting to a classical Hilbert modular form (as described in [loc. cit., 1.11.8]) f generates (a model over L for) the finite part π_f^∞ of a cuspidal automorphic representation π_f of $\mathrm{GL}_2(\mathbb{A}_F)$. For $\tau\in\Sigma$ we have $\pi_\tau = D_{k_\tau,w}$, where the k_τ can be read off from κ (using the bijection between Σ and Σ_p). The central character of π_f is an algebraic Hecke character of the form $\chi_f\omega^{-w}$, where ω is the norm character and χ_f is a finite-order character. The Galois representation $r_{\pi_f,\iota}$ is the unique semisimple representation, unramified outside $p\mathfrak{m}\mathfrak{n}$, such that for $\mathfrak{q}\nmid p\mathfrak{m}\mathfrak{n}$ a geometric Frobenius element at $\mathfrak{q}$ has characteristic polynomial $X^2-\theta_f(T_{\mathfrak{q}})X+(\mathbf{N}\mathfrak{q})^{1-w}\iota^{-1}\chi_f(\varpi_{\mathfrak{q}})$

In the subsequent sections of the paper, we will prove the following 'level optimisation' result:

Proposition 2.3. *Suppose $\mathfrak{m} = \mathfrak{q}\mathfrak{l}$ is a product of two prime ideals. Let $f \in H^0(\mathscr{M}(\mathfrak{n}, \mathfrak{m})_L, \underline{\omega}^\kappa)$ be a Hecke eigenform. Suppose that:*

- *The local factor $\pi_{f,\mathfrak{q}}$ is an unramified twist of the Steinberg representation.*
- *The residual representation*
$$\bar{r}_{p,\iota}(\pi_f) : \mathrm{Gal}(\bar{F}/F) \to \mathrm{GL}_2(\bar{\mathbb{F}}_p)$$
is irreducible.
- *The polynomials $X^2 - \theta_f(T_\mathfrak{p})X + (N\mathfrak{p})^{1-w}\iota^{-1}\chi_f(\varpi_\mathfrak{p})$ have distinct roots for all prime ideals $\mathfrak{p} \mid p$.*
- *The local representation*
$$r_{p,\iota}(\pi_f)|_{\mathrm{Gal}(\bar{F}_\mathfrak{l}/F_\mathfrak{l})}$$
is unramified.

Then there exists a Hecke eigenform $g \in H^0(\mathscr{M}(\mathfrak{n}, \mathfrak{q})_L, \underline{\omega}^\kappa)$ such that the Hecke eigenvalues of f and g coincide outside $\mathfrak{l}$ and $\theta_f(U_\mathfrak{l})$ is one of the roots of $X^2 - \theta_g(T_\mathfrak{l})X + (N\mathfrak{l})^{1-w}\iota^{-1}\chi_f(\varpi_\mathfrak{l})$.

We now explain why this proposition is sufficient to deduce Theorem 1.4.

Lemma 2.4. *Proposition 2.3 implies Theorem 1.4.*

Proof. Recall that we have reduced the statement of Theorem 1.4 to the case of π_0 as described immediately before Section 2.1. We denote the finite part of π_0 by π_0^∞. For $\mathfrak{m}$ any nonzero ideal of $\mathbb{O}_F$ we define congruence subgroups $U_0(\mathfrak{m})$ and $U_1(\mathfrak{m})$ to be the subgroups of $\mathrm{GL}_2(\hat{\mathbb{Z}})$ given by

$$U_0(\mathfrak{m}) = \left\{ g \in \mathrm{GL}_2(\hat{\mathbb{Z}}) : g = \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \mod \mathfrak{m} \right\}$$
$$U_1(\mathfrak{m}) = \left\{ g \in \mathrm{GL}_2(\hat{\mathbb{Z}}) : g = \begin{pmatrix} * & * \\ 0 & 1 \end{pmatrix} \mod \mathfrak{m} \right\}$$

Now we denote by $\mathfrak{l}$ and $\mathfrak{q}$ the prime ideals of $\mathbb{O}_F$ corresponding to the places v and v' respectively, and suppose that $\mathfrak{n}$ is an ideal of $\mathbb{O}_F$, coprime to $p\mathfrak{q}\mathfrak{l}$ and divisible by a rational integer ≥ 4, such that $(\pi_0^\infty)^{U_1(\mathfrak{n})\cap U_0(\mathfrak{q}\mathfrak{l})} \neq 0$. The isomorphism $\iota : \bar{\mathbb{Q}}_p \to \mathbb{C}$ induces a bijection between Σ and the embeddings Σ_p from $F \hookrightarrow L$, so using this we can associate a character

$$\kappa : \mathrm{Res}_{\mathbb{Z}}^{\mathbb{O}_F} \mathbb{G}_{m/\mathbb{O}_L} \to \mathbb{G}_{m/\mathbb{O}_L}$$

to the Σ-tuple of integers $\{k_\tau\}_{\tau\in\Sigma}$ describing the Archimedean part of π_0. Now (possibly enlarging L) there is a Hecke eigenform $f \in H^0(\mathscr{M}(\mathfrak{n}, \mathfrak{q}\mathfrak{l})_L, \underline{\omega}^\kappa)$ with $\pi_f \cong \pi_0$, and so f satisfies the hypotheses of Proposition 2.3. This proposition gives us a Hecke eigenform $g \in H^0(\mathscr{M}(\mathfrak{n}, \mathfrak{q})_L, \underline{\omega}^\kappa)$ with $\pi_g \cong \pi_0$, and so the space

$(\pi_0^\infty)^{U_1(\mathfrak{n})\cap U_0(\mathfrak{q})}$ is nonzero, which contradicts the assumption that $\pi_{0,\mathfrak{l}}$ is a twist of Steinberg. $\square$

We end this section by discussing Hilbert modular varieties with Iwahori level at p and a definition of overconvergent Hilbert modular forms.

Definition 2.5. Let $\mathfrak{c}$ be a fractional ideal of F such that its absolute norm $N\mathfrak{c}$ is coprime to $p\mathfrak{m}\mathfrak{n}$. We denote by $\underline{\mathscr{M}}^{\mathfrak{c},\mathrm{Iw}}(\mathfrak{n},\mathfrak{m})$ the functor from schemes over $\mathbb{O}_L$ to sets taking S to isomorphism classes of tuples $(A,\lambda,[\eta],H)$ up to isomorphism, where $(A,\lambda,[\eta])$ are as in Definition 2.2 and H is a finite flat subgroup scheme of $A[p]$, stable under the action of $\mathbb{O}_F$, of rank $p^{[F:\mathbb{Q}]}$ and isotropic with respect to the λ-Weil pairing.

The functor $\underline{\mathscr{M}}^{\mathfrak{c},\mathrm{Iw}}(\mathfrak{n},\mathfrak{m})$ is represented by an $\mathbb{O}_L$-scheme $\mathscr{M}^{\mathfrak{c},\mathrm{Iw}}(\mathfrak{n},\mathfrak{m})$. We denote by $\mathscr{M}^{\mathrm{Iw}}(\mathfrak{n},\mathfrak{m})$ the disjoint union of these over suitable representatives $\mathfrak{c}_i$ of the narrow class group, as before.

Denoting the rigid generic fibre of $\mathscr{M}^{\mathrm{Iw}}(\mathfrak{n},\mathfrak{m})$ by $M^{\mathrm{Iw}}(\mathfrak{n},\mathfrak{m})_L$, we can consider sections of the line bundles $\underline{\omega}^\kappa$ over strict neighbourhoods of the locus in $M^{\mathrm{Iw}}(\mathfrak{n},\mathfrak{m})_L$, where H is Cartier dual to $\mathbb{O}_F/p$ (i.e., the multiplicative ordinary locus), to obtain a space $M_\kappa^\dagger(\mathfrak{n},\mathfrak{m})$ of overconvergent modular forms of weight κ (see [Kassaei 2013, §4]).

3. Eigenvarieties

We will need to make use of eigenvarieties constructed in different contexts. To clarify the relationship between these eigenvarieties, we are going to follow the abstract approach of [Bellaïche and Chenevier 2009, 7.2]. First we need to discuss the weight spaces over which our eigenvarieties will live.

3.1. ***Weight spaces.*** We set $G=\mathrm{Res}_{F/\mathbb{Q}}(\mathrm{GL}_2)(\mathbb{Q}_p)=\prod_{\mathfrak{p}\mid p}G_\mathfrak{p}$, where $G_\mathfrak{p}=\mathrm{GL}_2(F_\mathfrak{p})$. We denote by $B=\prod_{\mathfrak{p}\mid p}B_\mathfrak{p}\subset G$ the Borel subgroup comprising upper-triangular matrices and by $T=\prod_{\mathfrak{p}\mid p}T_\mathfrak{p}$ the maximal torus comprising diagonal matrices. We denote by $N_\mathfrak{p}$ the subgroup of $B_\mathfrak{p}$ whose elements have 1 on the diagonal. Finally, $T_0=\prod_{\mathfrak{p}\mid p}T_{0,\mathfrak{p}}\subset T$ is the compact subgroup given by elements with integral entries.

Fix a finite extension $L\subset\overline{\mathbb{Q}}_p$ of $\mathbb{Q}_p$, which we assume contains a normal closure of F. The functor taking an L-affinoid $\mathrm{Sp}\,A$ to the set of continuous $A^\times$-valued characters of

$$T_0=(\mathbb{O}_F\otimes_\mathbb{Z}\mathbb{Z}_p)^\times\times(\mathbb{O}_F\otimes_\mathbb{Z}\mathbb{Z}_p)^\times$$

is representable by a rigid analytic space $\widehat{T}_0$ over L [Buzzard 2004, §2]. Likewise, we have a rigid space $\widehat{T}$ representing continuous characters of T.

Definition 3.2. Denote by $\mathscr{W}$ the subspace of $\widehat{T}_0$ whose points correspond to continuous characters of T_0 which are trivial on a finite-index subgroup of $\mathbb{O}_F^\times$ (embedded diagonally in T_0).

Suppose we have an algebraic character $\kappa : \operatorname{Res}_{\mathbb{Z}}^{\mathbb{O}_F} \mathbb{G}_m/\mathbb{O}_L \to \mathbb{G}_m/\mathbb{O}_L$ such that $N^w \cdot \kappa^{-1}$ admits a square root. Recall that Σ_p denotes the set of embeddings from F to L, so κ corresponds to a Σ-tuple of integers $\{k_\tau\}_{\tau\in\Sigma}$, with the same parity as w.

The character

$$(t_1, t_2) \mapsto \prod_{\tau\in\Sigma_p} \tau(t_1)^{\frac{k_\tau-2-w}{2}} \tau(t_2)^{\frac{-w-k_\tau+2}{2}}$$

is a point in $\mathscr{W}(L)$, which we also denote by κ.

Definition 3.3. Denote by $\mathscr{W}_\kappa$ the subspace of $\mathscr{W}$ such that maps from an L-affinoid $\operatorname{Sp} A$ to $\mathscr{W}_\kappa$ correspond to characters $(\chi_1, \chi_2) : (\mathbb{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p)^\times \times (\mathbb{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p)^\times \to A^\times$ such that $\chi_1\chi_2 = \mathrm{Nm}^{-w}$ and $\chi_1\chi_2^{-1} = \kappa\, \mathrm{Nm}^{-2} \cdot (\epsilon \circ \mathrm{Nm})$, where ϵ is a continuous character $\epsilon : \mathbb{Z}_p^\times \to A^\times$ with

$$v_p(1-\epsilon(a)) > \frac{1}{p-1}$$

for all $a \in \mathbb{Z}_p^\times$. Here Nm is the natural extension of the norm map to a continuous map $(\mathbb{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p)^\times \to \mathbb{Z}_p^\times$.

Taking ϵ to be the trivial character, we see that κ is a point of $\mathscr{W}_\kappa$. In fact, $\mathscr{W}_\kappa$ is a (one-dimensional) disc in $\mathscr{W}$, with centre κ.

Remark 3.4. The weight space $\mathscr{W}$ is isomorphic to the weight space (also denoted by $\mathscr{W}$) defined in [Buzzard 2007, §8]. Our character (χ_1, χ_2) of $(\mathbb{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p)^\times \times (\mathbb{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p)^\times$ corresponds to the character $(n, v) = (\chi_1\chi_2^{-1}, \chi_2)$ in Buzzard's weight space.

The weight space $\mathscr{W}_\kappa$ is isomorphic to the weight space (also denoted by $\mathscr{W}_\kappa$) defined in [Kisin and Lai 2005, §4.5].

We are going to build eigenvarieties interpolating classical Hilbert modular forms over the weight space $\mathscr{W}_\kappa$. These eigenvarieties will be constructed using three different notions of 'overconvergent automorphic form'.

3.5. ***Refinements.*** For $\mathfrak{p} \mid p$, denote by $I_\mathfrak{p}$ the Iwahori subgroup of $G_\mathfrak{p}$, comprising matrices which reduce to upper-triangular matrices mod $\varpi_\mathfrak{p}$. We write I for the product $\prod_{\mathfrak{p}\mid p} I_\mathfrak{p} \subset G$.

Definition 3.6. Let $\pi_\mathfrak{p}$ be an irreducible smooth representation of $G_\mathfrak{p}$ on a complex vector space which is either an irreducible principal series or a twist of the Steinberg representation. An *accessible refinement* of $\pi_\mathfrak{p}$ is a character χ of $T_\mathfrak{p}$ such that there is a $G_\mathfrak{p}$-equivariant embedding

$$\pi_\mathfrak{p} \hookrightarrow \operatorname{Ind}_{B_\mathfrak{p}}^{G_\mathfrak{p}} \chi.$$

Remark 3.7. We have the following explicit description of accessible refinements: we have $\pi_\mathfrak{p} \cong \operatorname{Ind}_{B_\mathfrak{p}}^{G_\mathfrak{p}} \mu_1 \otimes \mu_2$ or $\pi_\mathfrak{p} \cong \operatorname{St} \otimes \mu$, where μ, μ_i are characters of $F_\mathfrak{p}^\times$.

- Suppose $\pi_\mathfrak{p} \cong \operatorname{Ind}_{B_\mathfrak{p}}^{G_\mathfrak{p}} \mu_1 \otimes \mu_2$. Then the accessible refinements of $\pi_\mathfrak{p}$ are the characters $\mu_1 \otimes \mu_2$ and $\mu_2 \otimes \mu_1$. Note that if $\mu_1 = \mu_2$ these refinements are the same.
- Suppose $\pi_\mathfrak{p} \cong \operatorname{St} \otimes \mu$. Then $\pi_\mathfrak{p}$ is isomorphic to the unique irreducible subrepresentation of the normalised induction $\operatorname{Ind}_{B_\mathfrak{p}}^{G_\mathfrak{p}} \mu|\cdot|_\mathfrak{p}^{1/2} \otimes \mu|\cdot|_\mathfrak{p}^{-1/2}$ and the unique accessible refinement of $\pi_\mathfrak{p}$ is $\mu|\cdot|_\mathfrak{p}^{1/2} \otimes \mu|\cdot|_\mathfrak{p}^{-1/2}$.

For an ideal $\mathfrak{n}$ of $\mathcal{O}_F$, coprime to p, we denote by $\mathcal{H}(\mathfrak{n})$ the free commutative polynomial ring over $\mathbb{Z}$ on generators labelled T_v and S_v for places v of F not dividing $p\mathfrak{n}$ and U_v for places $v \mid \mathfrak{n}$.

For $\mathfrak{m}, \mathfrak{n}$ ideals of $\mathcal{O}_F$ (coprime to each other and to p), suppose π is a cuspidal automorphic representation of $\mathrm{GL}_2(\mathbb{A}_F)$ with $(\pi^{\infty p})^{U_1(\mathfrak{n}) \cap U_0(\mathfrak{m})} \neq 0$ and $\pi_\tau = D_{k_\tau, w}$ for each $\tau \in \Sigma$, with $k_\tau \geq 2$ and $k_\tau = w \bmod 2$ for each τ. We say that such a π is a classical automorphic representation of tame level $U_1(\mathfrak{n}) \cap U_0(\mathfrak{m})$ and weight (k, w). We moreover say that π has finite slope if the smooth Jacquet modules $J_{B_\mathfrak{p}}(\pi_\mathfrak{p})$ are nonzero for all $\mathfrak{p} \mid p$. Equivalently, the $\pi_\mathfrak{p}$ are all irreducible principal series or twists of Steinberg (since local factors of a cuspidal automorphic representation are generic).

Definition 3.8. Suppose π is a finite-slope representation. An *accessible refinement* of π is a character

$$\chi = \bigotimes_{\mathfrak{p} \mid p} \chi_\mathfrak{p} : T \to \mathbb{C}^\times$$

such that each $\chi_\mathfrak{p}$ is an accessible refinement of $\pi_\mathfrak{p}$.

Remark 3.9. The possible accessible refinements of finite-slope automorphic representations are completely classified by Remark 3.7.

Definition 3.10. We say that a finite-slope representation is *unramified* if $\pi_\mathfrak{p}$ is either an unramified principal series or an unramified twist of Steinberg for all $\mathfrak{p} \mid p$.

Remark 3.11. The following two conditions are easily seen to be equivalent to π being unramified:

- For every $\mathfrak{p} \mid p$, $\pi_\mathfrak{p}$ has nonzero invariants for the Iwahori subgroup $I_\mathfrak{p}$.

- Every accessible refinement of π factors through T/T_0.

Definition 3.12. Suppose π is a finite-slope representation of weight (k, w) and χ is an accessible refinement of π. We define a continuous character

$$\nu(\pi,\chi) : T \to \overline{\mathbb{Q}}_p^\times$$

by

$$\nu(\pi,\chi) = \bigotimes_{\mathfrak{p}\mid p} \iota^{-1}\chi_{\mathfrak{p}}|\cdot|_{\mathfrak{p}}^{1/2} \otimes |\cdot|_{\mathfrak{p}}^{-1/2} \prod_{\tau\in\Sigma_{\mathfrak{p}}} \tau^{(k_\tau-2-w)/2} \otimes \tau^{(-w-k_\tau+2)/2}.$$

Remark 3.13. Our discussion of accessible refinements and the definition of the character $\nu(\pi,\chi)$ is parallel to that of [Chenevier 2009, §1.4]. We will see in Section 4.2 that these definitions are completely natural when constructing eigenvarieties using completed cohomology and Emerton's locally analytic Jacquet functor.

3.14. *Abstract eigenvarieties.*

Definition 3.15. Let $\mathscr{W}'$ be any subspace of $\mathscr{W}$ which is an admissible increasing union of affinoids (in practice it will be $\mathscr{W}_\kappa$). Let $\mathscr{H}$ be a commutative $\mathbb{Z}$-algebra and let $\mathscr{Z}$ be a subset of $\mathrm{Hom}(\mathscr{H}, \overline{\mathbb{Q}}_p) \times \widehat{T}(\overline{\mathbb{Q}}_p)$ whose image in $\widehat{T}_0(\overline{\mathbb{Q}}_p)$ is an accumulation[2] and Zariski-dense subset of $\mathscr{W}'$. Denote by Y the fibre product of $\widehat{T}$ and $\mathscr{W}'$ over $\widehat{T}_0$. Then an *eigenvariety* for the triple $(\mathscr{H}, \mathscr{Z}, \mathscr{W}')$ is a reduced rigid space X over L equipped with:

- a ring homomorphism $\psi : \mathscr{H} \to \mathbb{O}(X)$,
- a finite morphism $\nu : X \to Y$,
- an accumulation and Zariski-dense subset $Z \subset X(\overline{\mathbb{Q}}_p)$ (which we refer to as the 'classical subset'),

such that the following are satisfied:

(1) For all open affinoids $V \subset Y$ the natural map

$$\psi \otimes \nu^* : \mathscr{H} \otimes \mathbb{O}(V) \to \mathbb{O}(\nu^{-1}(V))$$

is surjective.

(2) The natural evaluation map

$$\begin{aligned} X(\overline{\mathbb{Q}}_p) &\to \mathrm{Hom}(\mathscr{H}, \overline{\mathbb{Q}}_p), \\ x &\mapsto \psi_x := (h \mapsto \psi(h)(x)) \end{aligned}$$

induces a bijection $z \mapsto (\psi_z, \nu(z))$ from Z to $\mathscr{Z}$.

[2] A set is accumulation if each point $z \in Z$ has a basis of affinoid neighbourhoods U such that $Z \cap U$ is Zariski-dense in U.

The key property of the above definition is that an eigenvariety is unique up to unique isomorphism, by [Bellaïche and Chenevier 2009, Proposition 7.2.8].

Remark 3.16. An alternative way to abstractly characterise eigenvarieties in our context is as Zariski closures of sets of classical points in the rigid space given by the product of the rigid generic fibre of a Galois (pseudo)deformation ring and some affine spaces or copies of $\mathbb{G}_m$ (to keep track of additional Hecke eigenvalues).

For $\mathfrak{m}, \mathfrak{n}$ ideals of $\mathcal{O}_F$ (coprime to each other and to p), suppose π is an unramified representation of tame level $U_1(\mathfrak{n}) \cap U_0(\mathfrak{m})$.

There is a natural action of $\mathscr{H}(\mathfrak{mn})$ on $(\pi^\infty)^{U_1(\mathfrak{n})\cap U_0(p\mathfrak{m})}$, where we let T_v and U_v act by double coset operators

$$\left[U\begin{pmatrix}\varpi_v & 0\\ 0 & 1\end{pmatrix}U\right]$$

and let S_v act by the double coset operator

$$\left[U\begin{pmatrix}\varpi_v & 0\\ 0 & \varpi_v\end{pmatrix}U\right],$$

where $U = U_1(\mathfrak{n}) \cap U_0(p\mathfrak{m})$.

Given such a π, we obtain a subset $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_\pi$ of $\mathrm{Hom}(\mathscr{H}(\mathfrak{mn}), \overline{\mathbb{Q}}_p) \times \widehat{T}(\overline{\mathbb{Q}}_p)$ by taking pairs

$$(\psi, \nu(\pi,\chi)),$$

where χ is an accessible refinement of π and ψ is a character corresponding (via ι) to a Hecke eigenform in $(\pi^\infty)^{U_1(\mathfrak{n})\cap U_0(p\mathfrak{m})}$. Note that any element of the set $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_\pi$ determines π, by strong multiplicity one. The choice of accessible refinement χ corresponds to a choice of $U_\mathfrak{p}$-eigenvalue in the space $(\pi_\mathfrak{p})^{I_\mathfrak{p}}$ for each $\mathfrak{p} \mid p$. The character $\kappa_\pi := \nu(\pi,\chi)|_{T_0}$ is independent of the refinement χ.

Definition 3.17. Let $\mathfrak{m}$ and $\mathfrak{n}$ be a pair of coprime ideals in $\mathcal{O}_F$, both coprime to p.

(1) Denote by $\mathscr{Z}(\mathfrak{n},\mathfrak{m})$ the union of the $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_\pi$ obtained from unramified π with tame level $U_1(\mathfrak{n}) \cap U_0(\mathfrak{m})$ such that $\kappa_\pi \in \mathscr{W}_K$.

(2) For $\mathfrak{q}$ a prime divisor of $\mathfrak{m}$, we write $\mathscr{Z}(\mathfrak{n},\mathfrak{m})^{\mathfrak{q}\text{-sp}}$ for the subset of $\mathscr{Z}(\mathfrak{n},\mathfrak{m})$ arising from π with $\pi^{U_1(\mathfrak{n})\cap U_0(p\mathfrak{m}/\mathfrak{q})} = 0$ (equivalently, the local factor $\pi_\mathfrak{q}$ is an unramified twist of the Steinberg representation).

(3) Similarly, we write $\mathscr{Z}(\mathfrak{n},\mathfrak{m})^{\mathfrak{q}\text{-ps}}$ for the subset of $\mathscr{Z}(\mathfrak{n},\mathfrak{m})$ arising from π with $\pi^{U_1(\mathfrak{n})\cap U_0(p\mathfrak{m}/\mathfrak{q})} \neq 0$ (equivalently, the local factor $\pi_\mathfrak{q}$ is an unramified principal series representation).

(4) Now we consider the irreducible mod p Galois representation $\bar{\rho} = \bar{r}_{p,\iota}(\pi_f)$. We can define $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$, $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-sp}}$ and $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-ps}}$ to be the subsets of $\mathscr{Z}(\mathfrak{n},\mathfrak{m})$, etc. arising from those π with residual Galois representation $\bar{r}_{p,\iota}(\pi)$ isomorphic to $\bar{\rho}$.

3.18. ***Buzzard's eigenvarieties.***

Theorem 3.19. *There exist eigenvarieties $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$, $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-sp}}$ and $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-ps}}$ for the triples*

$$(\mathscr{H}(\mathfrak{mn}), \mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}, \mathscr{W}_\kappa),$$
$$(\mathscr{H}(\mathfrak{mn}), \mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-sp}}, \mathscr{W}_\kappa),$$
$$(\mathscr{H}(\mathfrak{mn}), \mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-ps}}, \mathscr{W}_\kappa).$$

We denote the classical subsets of these eigenvarieties by $Z(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$, $Z(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-sp}}$ and $Z(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-ps}}$. The following properties are satisfied by these eigenvarieties:

- *There are closed immersions*

$$\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-sp}} \hookrightarrow \mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}} \quad \textit{and} \quad \mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-ps}} \hookrightarrow \mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$$

  commuting with the maps to weight space and respecting the homomorphisms ψ from $\mathscr{H}(\mathfrak{mn})$, with images X, Y, respectively, given by unions of irreducible components (in the sense of [Conrad 1999]).

- *Each irreducible component of $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$ is contained in precisely one of X and Y.*

- *We have $Z(\mathfrak{n},\mathfrak{m})_{\bar{\rho}} \cap X = Z(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-sp}}$ and $Z(\mathfrak{n},\mathfrak{m})_{\bar{\rho}} \cap Y = Z(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-ps}}$.*

- *There is a map*

$$\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\mathfrak{q}\text{-ps}} \to \mathscr{E}(\mathfrak{n},\mathfrak{m}/\mathfrak{q})_{\bar{\rho}},$$

  surjective on closed points, for which the preimage of a closed point $x \in \mathscr{E}(\mathfrak{n},\mathfrak{m}/\mathfrak{q})_{\bar{\rho}}$ is indexed by the roots of the Hecke polynomial $X^2 - \psi(T_\mathfrak{q})(x)X + N\mathfrak{q}\psi(S_\mathfrak{q})(x)$.

Proof. We fix a definite quaternion algebra D/F, ramified precisely at the infinite places of F, and an isomorphism $(D \otimes_F \mathbb{A}_F^\infty)^\times \cong \mathrm{GL}_2(\mathbb{A}_F^\infty)$. Buzzard's definition [2007, Part III] of overconvergent automorphic forms on D, with tame level $U_1(\mathfrak{n}) \cap U_0(\mathfrak{m})$, allows us to construct an eigenvariety $\mathscr{E}(\mathfrak{n},\mathfrak{m})$ for the triple $(\mathscr{H}(\mathfrak{mn}), \mathscr{Z}(\mathfrak{n},\mathfrak{m}), \mathscr{W}_\kappa)$. The Zariski density and accumulation properties for the classical points follow from a special case of the classicality criterion of [Loeffler 2011, Theorem 3.9.6]. To obtain the eigenvariety for $(\mathscr{H}(\mathfrak{mn}), \mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}, \mathscr{W}_\kappa)$ we just take the union of connected components in $\mathscr{E}(\mathfrak{n},\mathfrak{m})$ whose closed points have associated residual Galois representation isomorphic to $\bar{\rho}$.

We then define $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$ to be the Zariski closure in $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$ of the subset $Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}} \subset Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}$ corresponding to systems of Hecke eigenvalues in $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$. Similarly, we define $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}^{\text{q-ps}}$ to be the Zariski closure in $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$ of the subset $Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-ps}}$. We now need to check that $Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$ and $Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-ps}}$ are accumulation subsets in their Zariski closures, together with the rest of the assertions of the theorem.

We can deduce everything we need by applying the results of [Bellaïche and Chenevier 2009, 7.8] on the family of Weil–Deligne representations carried by an eigenvariety (see also [Paulin 2011]). We proceed as follows: Denote by $X \subset \mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$ the reduced closed subspace given by the union of irreducible components where the monodromy operator in the associated family of Weil–Deligne representations is generically nonzero — we call such a component 'generically special'. More precisely, in the notation of [Bellaïche and Chenevier 2009, 7.8], a generically special component W is one in which for all closed points $x \in W$ we have $N_{s(x)}^{\text{gen}}$ nonzero for $s(x)$ any germ of an irreducible component at x which is contained in W. Then we claim that

$$X \cap Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}} = Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}.$$

Indeed, if $x \in Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$, then, by Proposition 7.8.19(iii) of [Bellaïche and Chenevier 2009] and local-global compatibility at $\mathfrak{q}$ for the automorphic representation π_x [Carayol 1986b], every irreducible component passing through x is generically special.

Conversely, if $x \in X \cap Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}$, then the Weil–Deligne representation at $\mathfrak{q}$ associated to ρ_x is forced to have the form $(W \oplus W(1), N)$, where W is a one-dimensional $\overline{k(x)}$-vector space with an unramified action of the Weil group $W_{\mathfrak{q}}$ and $W(1)$ denotes the twist of W by the p-adic cyclotomic character (N could be zero or nonzero). This means that either $x \in Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$ or the local factor $\pi_{\mathfrak{q}}$ of the automorphic representation π associated to x is one-dimensional. The latter situation cannot occur, since π is a cuspidal automorphic representation of $\mathrm{GL}_2(\mathbb{A}_F)$ and therefore its local factors are infinite-dimensional (this follows from the existence of a global Whittaker model; for example, see the proof of Theorem 11.1 of [Jacquet and Langlands 1970]).

Now it is easy to deduce the accumulation property for $Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$ from the accumulation property for $Z(\mathfrak{n}\mathfrak{m})_{\bar{\rho}}$.

Similar arguments apply if we take $Y \subset \mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar{\rho}}$ to be the reduced closed subspace given by the union of irreducible components where the monodromy operator in the associated family of Weil–Deligne representations is generically zero — we call such components 'generically principal series'.

Finally we need to construct the map

$$\mathscr{E}(\mathfrak{n},\mathfrak{m})^{\mathfrak{q}\text{-ps}}_{\bar\rho} \to \mathscr{E}(\mathfrak{n},\mathfrak{m}/\mathfrak{q})_{\bar\rho}.$$

This can be done by giving an alternative construction of $\mathscr{E}(\mathfrak{n},\mathfrak{m})^{\mathfrak{q}\text{-ps}}_{\bar\rho}$ — indeed, by the uniqueness of abstract eigenvarieties, $\mathscr{E}(\mathfrak{n},\mathfrak{m})^{\mathfrak{q}\text{-ps}}_{\bar\rho}$ coincides with the nilreduction of the covering of $\mathscr{E}(\mathfrak{n},\mathfrak{m}/\mathfrak{q})_{\bar\rho}$ given by the roots of $X^2 - \psi(T_\mathfrak{q})X + N\mathfrak{q}\psi(S_\mathfrak{q})(x)$ (this has a natural interpretation as a relative spectrum over $\mathscr{E}(\mathfrak{n},\mathfrak{m}/\mathfrak{q})_{\bar\rho}$). □

3.20. ***Kisin and Lai's eigenvarieties.*** Now we let f be a Hecke eigenform as in the statement of Proposition 2.3. For each prime ideal $\mathfrak{p} \mid p$, we denote the two distinct roots of $X^2 - \theta_f(T_\mathfrak{p})X + (N\mathfrak{p})^{1-w}\iota^{-1}\chi_f(\varpi_\mathfrak{p})$ by $\alpha_\mathfrak{p}$ and $\beta_\mathfrak{p}$. Then for each subset $S \subset \{\mathfrak{p} \mid p\}$ there is a unique Hecke eigenform

$$f_S \in H^0(\mathscr{M}^{\mathrm{Iw}}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_L, \underline{\omega}^\kappa),$$

whose Hecke eigenvalues away from p are the same as f, and for $\mathfrak{p} \mid p$ we have $\theta_{f_S}(U_\mathfrak{p}) = \alpha_\mathfrak{p}$ if $\mathfrak{p} \in S$ and $\theta_{f_S}(U_\mathfrak{p}) = \beta_\mathfrak{p}$ if $\mathfrak{p} \notin S$.

We moreover define a point $\nu_S \in \widehat{T}(\overline{\mathbb{Q}}_p)$ to be given by the character

$$\bigotimes_{\mathfrak{p} \mid p} \chi_{\mathfrak{p},1} \otimes \chi_{\mathfrak{p},2} \prod_{\tau \in \Sigma_\mathfrak{p}} \tau^{(k_\tau - 2 - w)/2} \otimes \tau^{(-w-k_\tau+2)/2},$$

where the $\chi_{\mathfrak{p},i}$ are characters of $F_\mathfrak{p}^\times/\mathbb{O}_\mathfrak{p}^\times$ defined by:

- $\chi_{\mathfrak{p},1}(\varpi_\mathfrak{p}) = \alpha_\mathfrak{p}(N\mathfrak{p})^{-1}$ if $\mathfrak{p} \in T$ and $\chi_{\mathfrak{p},1}(\varpi_\mathfrak{p}) = \beta_\mathfrak{p}(N\mathfrak{p})^{-1}$ otherwise.
- $\chi_{\mathfrak{p},2}(\varpi_\mathfrak{p}) = \beta_\mathfrak{p}$ if $\mathfrak{p} \in T$ and $\chi_{\mathfrak{p},1}(\varpi_\mathfrak{p}) = \alpha_\mathfrak{p}$ otherwise.

Proposition 3.21. *For each $S \subset \{\mathfrak{p} \mid p\}$ there is a point x_S of $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}$ such that*

$$\nu(x_S) = \nu_S,$$

and the character

$$\psi_{x_S} : \mathscr{H}(\mathfrak{n}\mathfrak{q}\mathfrak{l}) \to \overline{\mathbb{Q}}_p$$

induced by ψ is compatible with θ_{f_S}.

Proof. For this result, we need to use an alternative construction of the eigenvariety $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}$. This is given by the space $\mathscr{C}_\kappa(\mathfrak{m})$ of [Kisin and Lai 2005, Theorem 4.5.4] (with modified level structures). Here the $\mathfrak{m}$ corresponds to our choice of residual Galois representation $\bar\rho$. To show that $\mathscr{C}_\kappa(\mathfrak{m})$ coincides with $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}$, we need to verify that the subset of $\mathscr{C}_\kappa(\mathfrak{m})$ corresponding to the 'classical points' $\mathscr{Z}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}$ is Zariski-dense and accumulation. This follows from a classicality criterion for overconvergent Hilbert modular forms, which has recently been proved in two different ways — by Pilloni and Stroh [2011, Théorème 1.2] and by Tian and Xiao [2013, Proposition 6.3]. Note that Theorem 6.5 of the latter is the statement that the classical points are Zariski-dense in the Kisin–Lai eigenvarieties, but the accumulation property also follows immediately from the proof. The existence of the points x_S is now immediate from the construction of the Kisin–Lai eigenvariety.

Alternatively, one can avoid the appeal to difficult classicality theorems and instead show that $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}$ is isomorphic (with its additional structures) to the Zariski closure of the classical points in $\mathscr{C}_\kappa(\mathfrak{m})$ by working with the set-up described in Remark 3.16. Applying [Kisin and Lai 2005, Theorem 4.5.6] then concludes the proof. □

Since the local factor $\pi_{f,\mathfrak{q}}$ is an unramified twist of Steinberg, one naturally expects that the points x_S lie in the eigenvariety $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$. Proving that this is the case is a little delicate: the first thing to observe is that since we do not yet know local-global compatibility at $\mathfrak{q}$ for π_f, it is also possible that x_S could lie in $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-ps}}$.

One can build $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$ using modules given by taking a kernel of trace maps to prime-to-$\mathfrak{q}$ level from the appropriate Banach modules of overconvergent forms. However, forming this kernel (for modules over the ring of functions on an affinoid open U in weight space) does not necessarily commute with base change from U to a point in weight space (the cokernel of the trace maps may not be flat). So the fact that $\pi_{f,\mathfrak{q}}$ is an unramified twist of Steinberg does not immediate imply that x_S lies in $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$. One needs to show that x_S lies in a *family* of forms which are '$\mathfrak{q}$-new'. This problem is solved in [Newton 2014]:

Proposition 3.22. *The points x_S of Proposition 3.21 all lie in $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$.*

Proof. Take a point x_S and suppose that it does not lie in $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$. Then x_S is a point of $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-ps}}$, so we may consider the image of x_S in $\mathscr{E}(\mathfrak{n},\mathfrak{l})_{\bar\rho}$. This point satisfies the assumptions of [Newton 2014, Theorem 4.3], but the conclusion of this theorem tells us that x_S is indeed in $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$. □

4. Completed cohomology and level optimisation

We give another construction of the eigenvariety $\mathscr{E}(\mathfrak{n},\mathfrak{m})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$, using the completed cohomology of Shimura curves. We will then apply [Newton 2013, Theorem 4.33] to deduce the following result:

Theorem 4.1. *Let $S \subset \{\mathfrak{p} \mid p\}$ and let x_S be the point of $\mathscr{E}(\mathfrak{n},\mathfrak{q}\mathfrak{l})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$ obtained from Propositions 3.21 and 3.22. Then there is a point y_S of $\mathscr{E}(\mathfrak{n},\mathfrak{q})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$ such that:*

- *The Hecke eigenvalues outside $\mathfrak{l}$ coincide with those of x_S.*
- *$\psi_{x_S}(U_\mathfrak{l})$ is one of the roots of $X^2-\psi_{y_S}(T_\mathfrak{l})X+(N\mathfrak{l})^{1-w}\iota^{-1}\chi_f(\varpi_\mathfrak{l})$.*

The proof of this theorem occupies the rest of the section.

4.2. ***Completed cohomology of Shimura curves.*** For this section, we fix a quaternion algebra B/F such that B is nonsplit at precisely one infinite place, denoted τ_1, and one finite place, denoted $\mathfrak{q}$ (recall that $[F:\mathbb{Q}]$ is assumed to be even, so such quaternion algebras exist). Denote by G_B the reductive algebraic group over $\mathbb{Q}$ arising from the unit group of B. Note that G_B is an inner form of $\mathrm{Res}_{F/\mathbb{Q}}(\mathrm{GL}_2)$. For U a compact open subgroup of $G_B(\mathbb{A}_f)$, we have a complex (disconnected) Shimura curve

$$M(U)(\mathbb{C}) = G_B(\mathbb{Q})\backslash G_B(\mathbb{A}_f)\times(\mathbb{C}-\mathbb{R})/U,$$

where $G_B(\mathbb{Q})$ acts on $\mathbb{C}-\mathbb{R}$ via the τ_1 factor of $G_B(\mathbb{R})$.

These curves have canonical models over F, which we denote by $M(U)$. We follow the conventions of [Carayol 1986a] to define this canonical model.

Definition 4.3. We define $\widetilde{H}^1(U^p, L)$ to be

$$\Big(\varprojlim_n \varinjlim_{U_p} H^1_{\text{ét}}(M(U_pU^p)_{\overline{F}}, \mathbb{O}_L/\mathfrak{m}_L^n)\Big)\otimes_{\mathbb{O}_L} L,$$

where U^p is any compact open subgroup of $G_B(\mathbb{A}^{\infty,p})$ and U_p runs over the compact open subgroups of $G_B(\mathbb{Q}_p)$.

The L-vector space $\widetilde{H}^1(U^p, L)$ is naturally an L-Banach space with an admissible continuous action of $G_B(\mathbb{Q}_p)\cong\prod_{\mathfrak{p}\mid p}\mathrm{GL}_2(F_\mathfrak{p})$. Moreover, there is a direct summand $\widetilde{H}^1(U^p, L)_{\bar\rho}$ such that all the systems of Hecke eigenvalues arising from $\widetilde{H}^1(U^p, L)_{\bar\rho}$ correspond to Galois representations whose residual representations are isomorphic to $\bar\rho$.

We now explain how the systems of Hecke eigenvalues parametrised by the set $\mathscr{Z}(\mathfrak{n},\mathfrak{m})_{\bar\rho}^{\mathfrak{q}\text{-sp}}$ can be seen in the space $\widetilde{H}^1(U^p, L)_{\bar\rho}$.

Suppose we have a Σ-tuple of integers $k=(k_\tau)_{\tau\in\Sigma}$ with each $k_\tau\geq 2$ and an integer w with $k_\tau = w \bmod 2$ for all τ.

We denote by $W_{k,w}$ the L-representation of $\mathrm{GL}_2(F_p)$ defined by

$$\otimes_{\tau\in\Sigma_p}(\tau\circ\det)^{(w-k_\tau+2)/2}\,\mathrm{Sym}^{k_\tau-2}\,V_\tau,$$

where V_τ is the representation of $\mathrm{GL}_2(F_p)$ acting via τ and the standard representation of $\mathrm{GL}_2(L)$. These representations then give rise to lisse étale L-sheaves $\mathscr{F}_{k,w}$ on the curves $M(U)$ (see, for example, [Newton 2013, §3.2]).

We set U^p to be the prime-to-p part of the compact open subgroup of $G_B(\mathbb{A}^\infty)$ given by $U_1(\mathfrak{n})\cap U_0(\mathfrak{m}/\mathfrak{q})$. Now the Hecke algebra $\mathscr{H}(\mathfrak{m}\mathfrak{n})$ acts on $\widetilde{H}^1(U^p, L)$ as follows: For places v prime to $\mathfrak{q}$, we have a standard action by double coset operators associated to our fixed uniformisers ϖ_v. For the place $\mathfrak{q}$, we choose a uniformiser $\varpi_{D_\mathfrak{q}}$ of the order $\mathbb{O}_{D_\mathfrak{q}}$ whose reduced norm is equal to the fixed uniformiser $\varpi_{\mathfrak{q}}$ of $\mathcal{O}_{\mathfrak{q}}$, and let $U_{\mathfrak{q}}$ act on $\widetilde{H}^1(U^p, L)$ via the action of $\varpi_{D_{\mathfrak{q}}}$. This definition is explained by the following:

Lemma 4.4. *Let $\pi_{\mathfrak{q}} = \mathrm{St} \otimes \mu$ be an unramified twist of the Steinberg representation of $\mathrm{GL}_2(F_{\mathfrak{q}})$. The local Jacquet–Langlands correspondent $\mathrm{JL}(\pi_{\mathfrak{q}})$ of $\pi_{\mathfrak{q}}$ is the one-dimensional representation of $D_{\mathfrak{q}}^\times$ given by $\mu \circ \mathrm{Nrd}$, where Nrd denotes the reduced norm. The $U_{\mathfrak{q}}$-eigenvalue of the space of Iwahori-invariants in $\pi_{\mathfrak{q}}$ is equal to $\mu(\varpi_{\mathfrak{q}})$, and is therefore equal to the eigenvalue for the action of $\varpi_{D_{\mathfrak{q}}}$ on $\mathrm{JL}(\pi_{\mathfrak{q}})$.*

Proof. This follows from the standard computation of the $U_{\mathfrak{q}}$-eigenvalue of the space of Iwahori-invariants in $\pi_{\mathfrak{q}}$. $\square$

We have the following proposition, which is proved exactly the same way as [Newton 2013, Theorem 5.2]:

Proposition 4.5. *There is a $G_B(\mathbb{Q}_p)$-, $\mathrm{Gal}(\overline{F}/F)$- and Hecke-equivariant isomorphism*

$$\bigoplus_{(k,w)} \varinjlim_{U_p} H^1_{\text{ét}}(M(U_pU^p)_{\overline{F}}, \mathscr{F}_{k,w})_{\bar{\rho}} \otimes_L W^\vee_{k,w} \cong \widetilde{H}^1(U^p, L)^{\mathrm{alg}}_{\bar{\rho}},$$

where the right-hand side is the space of locally algebraic vectors (in the sense of [Emerton 2011, 4.2.6]) *in the L-Banach space representation $\widetilde{H}^1(U^p, L)_{\bar{\rho}}$.*

The above proposition allows us to determine the contribution of classical automorphic representations to the Jacquet module (in the sense of [Emerton 2006a]) $J_B(\widetilde{H}^1(U^p, L)_{\bar{\rho}})$ of the ($\mathbb{Q}_p$-)locally analytic vectors in $\widetilde{H}^1(U^p, L)_{\bar{\rho}}$:

Lemma 4.6. *There is a T-, $\mathrm{Gal}(\overline{F}/F)$- and Hecke-equivariant embedding*

$$\bigoplus_{(k,w)} \left(\varinjlim_{U_p} H^1_{\text{ét}}(M(U_pU^p)_{\overline{F}}, \mathscr{F}_{k,w})_{\bar{\rho}} \right)_N \otimes_L \chi_{k,w} \hookrightarrow J_B(\widetilde{H}^1(U^p, L)_{\bar{\rho}}),$$

where the subscript N denotes coinvariants (*i.e.*, *the classical Jacquet module*) *and $\chi_{k,w}$ is the character of T given by*

$$\begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix} \mapsto \prod_{\tau \in \Sigma_p} \tau(s_1)^{(k_\tau - 2 - w)/2} \tau(s_2)^{(-w-k_\tau+2)/2}$$

Proof. This follows from left-exactness of the Jacquet module functor and [Emerton 2006a, Proposition 4.3.6], since the highest weight space $(W^\vee_{k,w})^N$ has T-action given by $\chi_{k,w}$. $\square$

The following lemma is a standard result in the smooth representation theory of the groups $\mathrm{GL}_2(F_{\mathfrak{p}})$:

Lemma 4.7. *Let μ, μ_1, μ_2 be smooth complex characters of $F_{\mathfrak{p}}^{\times}$.*

(1) *The Jacquet module $\pi(\mu_1, \mu_2)_{N(\mathfrak{p})}$ is isomorphic as a $T_{\mathfrak{p}}$-representation to*
$$\mu_1|\cdot|_{\mathfrak{p}}^{1/2} \otimes \mu_2|\cdot|_{\mathfrak{p}}^{-1/2} \oplus \mu_2|\cdot|_{\mathfrak{p}}^{1/2} \otimes \mu_1|\cdot|_{\mathfrak{p}}^{-1/2}.$$

(2) *The Jacquet module $(\mathrm{St} \otimes \mu)_{N(\mathfrak{p})}$ is isomorphic as a $T_{\mathfrak{p}}$-representation to*
$$\mu|\cdot|_{\mathfrak{p}} \otimes \mu|\cdot|_{\mathfrak{p}}^{-1}.$$

Proof. See for example [Goldfeld and Hundley 2011, Theorem 8.12.15]. □

As a consequence of Lemma 4.7, together with the Jacquet–Langlands correspondence and the contribution of automorphic representations of $G_B(\mathbb{A})$ to the cohomology of the curves $M(U)$, we obtain the following proposition:

Lemma 4.8. *Suppose $(\psi, \nu) \in \mathscr{Z}(\mathfrak{n}, \mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$. Then there is a nonzero element*
$$v \in J_B(\widetilde{H}^1(U^p, L)_{\bar{\rho}}) \otimes_L \overline{\mathbb{Q}}_p$$
on which the Hecke operators away from p act via the character ψ and on which the torus T acts via the character ν.

The above lemma tells us that the 'classical set' $\mathscr{Z}(\mathfrak{n}, \mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$ can be seen in the locally analytic T-representations $J_B(\widetilde{H}^1(U^p, L)_{\bar{\rho}})$. We now summarise Emerton's construction of an eigenvariety from this representation, and show that it is an eigenvariety for the triple $(\mathscr{H}(\mathfrak{m}\mathfrak{n}), \mathscr{Z}(\mathfrak{n}, \mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}, \mathscr{W}_{\kappa})$.

The T-representation $J_B(\widetilde{H}^1(U^p, L)_{\bar{\rho}})$ is naturally dual to a coherent sheaf $\mathscr{M}$ on $\widehat{T}$ (see [Emerton 2006b, Proposition 2.3.2]). Denote by Y the fibre product $\widehat{T} \times_{\widehat{T}_0} \mathscr{W}_{\kappa}$, and let $\mathscr{M}_Y$ denote the pullback of $\mathscr{M}$ to a coherent sheaf on Y.

Taking the relative spectrum of the commutative subalgebra of endomorphisms of this sheaf generated by the Hecke algebra $\mathscr{H}(\mathfrak{m}\mathfrak{n})$ gives a rigid space with a finite map to Y. Passing to the nilreduction gives a reduced rigid space which we denote by $\mathscr{E}_{\kappa}$. By the above lemma, we have a subset $Z \subset \mathscr{E}_{\kappa}$ of classical points corresponding to the elements of $\mathscr{Z}(\mathfrak{n}, \mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}$.

Lemma 4.9. *The space $\mathscr{E}_{\kappa}$, together with the classical subset Z, is an eigenvariety for the triple*
$$(\mathscr{H}(\mathfrak{m}\mathfrak{n}), \mathscr{Z}(\mathfrak{n}, \mathfrak{m})_{\bar{\rho}}^{\text{q-sp}}, \mathscr{W}_{\kappa}).$$

Proof. The only condition we have to check is that Z is an accumulation and Zariski-dense subset of $\mathscr{E}_{\kappa}$ — everything else follows from the construction of $\mathscr{E}_{\kappa}$. To prove this, we have to interpret $\mathscr{E}_{\kappa}$ as part of (the nilreduction of) an eigenvariety constructed as in [Buzzard 2007], following [Emerton 2006a, Proposition 4.2.36; Newton 2013, Lemma 5.13]. The proof is slightly involved, the main reason being that we can show that eigenvarieties constructed with completed cohomology have nice properties only (a priori) after composing the map to weight space $\widehat{T}_0$ with a map corresponding to restriction to a finite-index subgroup of T_0. This comes about because [Emerton 2006a, Proposition 4.2.36] applies to 'cofree' modules over an Iwasawa algebra, not 'coprojective' modules.

Choose $U_p \subset G_B(\mathbb{Q}_p)$ a sufficiently small compact open subgroup such that $\widetilde{H}^1(U^p, L)_{\bar{\rho}}$ is a cofree representation of $U_p/\overline{F^\times \cap U_p U^p}$, in the sense of [Newton 2013, Definition 5.7]. Such a U_p exists by [Newton 2013, Corollary 5.8] — in fact, it suffices to take U_p pro-p, since our assumptions on the tame level already ensure that $U_p U^p$ is neat.

Denoting the closed subgroup $\overline{F^\times \cap U_p U^p}$ of U_p by X, we define a compact commutative p-adic analytic group by $S := T_0 \cap U_p/X$, and denote the rigid space parametrising its continuous characters by $\widehat{S}$. The characters corresponding to points of $\mathscr{W}_\kappa$ are trivial on the units of F with norm 1, so (possibly shrinking U_p if $p = 2$) these characters are trivial on X. Therefore restriction to $T_0 \cap U_p$ gives a map $\mathscr{W}_\kappa \to \widehat{S}$. In fact, the definition of $\mathscr{W}_\kappa$ implies that this map is an isomorphism onto its image, which we denote by $\mathscr{W}_S$.

We denote by $\widetilde{\mathscr{W}}_\kappa$ the preimage of $\mathscr{W}_S$ in $\mathscr{W}$. The space $\widetilde{\mathscr{W}}_\kappa$ is a finite disjoint union of open discs, whose components are indexed by characters of the finite group $T_0/T_0 \cap U_p$. We denote by $\widetilde{Y}$ the fibre product $\widehat{T} \times_{\widehat{T}_0} \widetilde{\mathscr{W}}_\kappa$, and let $\mathscr{M}_{\widetilde{Y}}$ denote the pullback of $\mathscr{M}$ to a coherent sheaf on $\widetilde{Y}$. Mimicking the construction of $\mathscr{E}_\kappa$, we obtain a rigid space $\widetilde{\mathscr{E}}_\kappa$ with a finite map to $\widetilde{Y}$, such that $\mathscr{E}_\kappa$ is the open and closed subspace of $\widetilde{\mathscr{E}}_\kappa$ lying over $\mathscr{W}_\kappa$.

Consider an increasing sequence $X_n = \mathrm{Sp}(A_n)$ of admissible affinoid opens covering $\widehat{S}$, and write M for the space of global sections of $\mathscr{M}$ and M_n for the base change $M \widehat{\otimes}_{\mathscr{C}^{\mathrm{an}}(\widehat{S},L)} A_n$. It follows from [Emerton 2006a, Proposition 4.2.36] that M_n is the finite-slope part of an orthonormalisable Banach A_n-module with the action of a compact operator (coming from the action of an element $z \in T$). We may choose the X_n such that their inverse images in $\widetilde{\mathscr{W}}_\kappa$ are admissible affinoid opens $\widetilde{Y}_n = \mathrm{Sp}(B_n)$ (e.g., closed discs). It follows from [Buzzard 2007, Corollary 2.10] that the modules $M \widehat{\otimes}_{\mathscr{C}^{\mathrm{an}}(\widehat{S},L)} B_n$ are likewise finite-slope parts of orthonormalisable Banach B_n-modules with the action of a compact operator.

It now follows, as in [Chenevier 2004, Corollaire 6.4.4], that the image of each irreducible component of $\widetilde{\mathscr{E}}_\kappa$ in $\mathscr{W}_S$ is the image of a Fredholm hypersurface, and is therefore Zariski-open in $\mathscr{W}_S$. For each irreducible component of $\widetilde{\mathscr{E}}_\kappa$, the map to $\mathscr{W}_S$ factors through one of the connected components of $\widetilde{\mathscr{W}}_\kappa$, so each irreducible component of $\widetilde{\mathscr{E}}_\kappa$ has Zariski-open image in this connected component. In particular, each irreducible component of $\mathscr{E}_\kappa$ has Zariski-open image in $\mathscr{W}_\kappa$.

The classicality criterion of [Emerton 2006a, Theorem 4.4.5] now shows that Z is Zariski-dense and accumulation in $\mathscr{E}_\kappa$. $\square$

Now that we know that $\mathscr{E}(\mathfrak{n},\mathfrak{m})^{\text{q-sp}}_{\bar{\rho}}$ and $\mathscr{E}_\kappa$ are eigenvarieties for the same triple, we have the following:

Corollary 4.10. *The closed points $x \in \mathscr{E}(\mathfrak{n},\mathfrak{m})^{\text{q-sp}}_{\bar{\rho}}$ with $\nu(x) = \lambda \in \widehat{T}(\overline{\mathbb{Q}}_p)$ correspond bijectively with systems of Hecke eigenvalues appearing in the (finite-dimensional) $\overline{\mathbb{Q}}_p$-vector space*

$$J_B(\widetilde{H}^1(U^p, L)_{\bar{\rho}}) \otimes_L \overline{\mathbb{Q}}_p[\lambda],$$

defined to be the subspace where T acts via the character λ.

Proof. This follows from the construction of $\mathscr{E}_\kappa$ and Lemma 4.9. □

Now Theorem 4.1 follows from combining Corollary 4.10 and [Newton 2013, Theorem 4.33].

5. Proof of Theorem 1.4

We can also construct $\mathscr{E}(\mathfrak{n},\mathfrak{q})_{\bar{\rho}}$ using the overconvergent Hilbert modular forms of [Kisin and Lai 2005]. Therefore we conclude from Theorem 4.1 that there are overconvergent Hilbert modular eigenforms $g_S \in M_\kappa^\dagger(\mathfrak{n},\mathfrak{q})$ whose systems of Hecke eigenvalues correspond to those of the points y_S.

Applying [Kassaei 2013, Theorem 7.10], we glue the g_S into a classical Hilbert modular eigenform

$$g \in H^0(\mathscr{M}(\mathfrak{n},\mathfrak{q})_L, \underline{\omega}^\kappa)$$

as in the statement of Proposition 2.3 (we are using more general tame levels than Kassaei, but this presents no problem). This completes the proof of Proposition 2.3, and hence of Theorem 1.4.

Remark 5.1. Note that the situation in the above proof is that we have an overconvergent form g_S with level prime to $\mathfrak{l}$ such that one of its $\mathfrak{l}$-stabilisations is equal to the classical form f_S. One might guess that g_S can be obtained from f_S by applying a trace map, in which case it is immediate that g_S is classical. However, in the case we are considering all the trace maps vanish on f_S.

A simple example illustrating that some argument is required here is given by the classical weight-two Eisenstein series of level $\Gamma_0(l)$ (note that our level-lowering result [Newton 2013, Theorem 4.33] does not apply to this form). It is in the kernel of the trace maps, but it is also the l-stabilisation of a (nonoverconvergent) p-adic modular form.

Acknowledgements

During the period in which this research was conducted, the author was supported by Trinity College, Cambridge and an EPSRC postdoctoral fellowship. The author would also like to thank the anonymous referees for their helpful comments.

References

[Ash and Stevens 2008] A. Ash and G. Stevens, "p-adic deformations of arithmetic cohomology", preprint, 2008, http://www2.bc.edu/~ashav/Papers/Ash-Stevens-Oct-07-DRAFT-copy.pdf.

[Bellaïche and Chenevier 2009] J. Bellaïche and G. Chenevier, *Families of Galois representations and Selmer groups*, Astérisque **324**, Société Mathématique de France, Paris, 2009. MR 2011m:11105 Zbl 1192.11035

[Blasius and Rogawski 1993] D. Blasius and J. D. Rogawski, "Motives for Hilbert modular forms", *Invent. Math.* **114**:1 (1993), 55–87. MR 94i:11033 Zbl 0829.11028

[Buzzard 2004] K. Buzzard, "On p-adic families of automorphic forms", pp. 23–44 in *Modular curves and abelian varieties*, edited by J. Cremona et al., Progr. Math. **224**, Birkhäuser, Basel, 2004. MR 2005g:11069 Zbl 1166.11322

[Buzzard 2007] K. Buzzard, "Eigenvarieties", pp. 59–120 in *L-functions and Galois representations*, edited by D. Burns et al., London Math. Soc. Lecture Note Ser. **320**, Cambridge University Press, 2007. MR 2010g:11076 Zbl 1230.11054

[Carayol 1986a] H. Carayol, "Sur la mauvaise réduction des courbes de Shimura", *Compositio Math.* **59**:2 (1986), 151–230. MR 88a:11058 Zbl 0607.14021

[Carayol 1986b] H. Carayol, "Sur les représentations l-adiques associées aux formes modulaires de Hilbert", *Ann. Sci. École Norm. Sup.* (4) **19**:3 (1986), 409–468. MR 89c:11083 Zbl 0616.10025

[Chenevier 2004] G. Chenevier, "Familles p-adiques de formes automorphes pour GL_n", *J. Reine Angew. Math.* **570** (2004), 143–217. MR 2006b:11046 Zbl 1093.11036

[Chenevier 2009] G. Chenevier, "Une application des variétés de Hecke des groups unitaires", preprint, 2009, http://gaetan.chenevier.perso.math.cnrs.fr/articles/famgal.pdf.

[Conrad 1999] B. Conrad, "Irreducible components of rigid spaces", *Ann. Inst. Fourier* (*Grenoble*) **49**:2 (1999), 473–541. MR 2001c:14045 Zbl 0928.32011

[Emerton 2006a] M. Emerton, "Jacquet modules of locally analytic representations of p-adic reductive groups, I: Construction and first properties", *Ann. Sci. École Norm. Sup.* (4) **39**:5 (2006), 775–839. MR 2008c:22013 Zbl 1117.22008

[Emerton 2006b] M. Emerton, "On the interpolation of systems of eigenvalues attached to automorphic Hecke eigenforms", *Invent. Math.* **164**:1 (2006), 1–84. MR 2007k:22018 Zbl 1090.22008

[Emerton 2011] M. Emerton, "Locally analytic vectors in representations of locally p-adic analytic groups", preprint, 2011, http://www.math.uchicago.edu/~emerton/pdffiles/analytic.pdf. To appear in Memoirs of the AMS.

[Goldfeld and Hundley 2011] D. Goldfeld and J. Hundley, *Automorphic representations and L-functions for the general linear group, I*, Cambridge Studies in Advanced Mathematics **129**, Cambridge University Press, 2011. MR 2012i:11054 Zbl 1247.11070

[Jacquet and Langlands 1970] H. Jacquet and R. P. Langlands, *Automorphic forms on* GL(2), Lecture Notes in Mathematics **114**, Springer, Berlin-New York, 1970. MR 53 #5481 Zbl 0236.12010

[Jarvis 1997] F. Jarvis, "On Galois representations associated to Hilbert modular forms", *J. Reine Angew. Math.* **491** (1997), 199–216. MR 98h:11067 Zbl 0914.11025

[Jorza 2010] A. Jorza, *Crystalline representations for GL(2) over quadratic imaginary fields*, Ph.D. thesis, Princeton University, 2010, http://search.proquest.com/docview/522118786. MR 2753218

[Kassaei 2013] P. L. Kassaei, "Modularity lifting in parallel weight one", *J. Amer. Math. Soc.* **26**:1 (2013), 199–225. MR 2983010 Zbl 1296.11052

[Kassaei et al. 2014] P. L. Kassaei, S. Sasaki, and Y. Tian, "Modularity lifting results in parallel weight one and applications to the Artin conjecture: the tamely ramified case", *Forum Math. Sigma* **2** (2014), 58 (electronic). MR 3294624 Zbl 1296.11053

[Kisin and Lai 2005] M. Kisin and K. F. Lai, "Overconvergent Hilbert modular forms", *Amer. J. Math.* **127**:4 (2005), 735–783. MR 2006d:11045 Zbl 1129.11020

[Loeffler 2011] D. Loeffler, "Overconvergent algebraic automorphic forms", *Proc. Lond. Math. Soc.* (3) **102**:2 (2011), 193–228. MR 2012f:11098 Zbl 1232.11056

[Luu 2015] M. Luu, *Deformation theory and local-global compatibility of Langlands correspondences*, Memoirs of the AMS **1123**, Amer. Math. Soc., Providence, RI, online publication February 2015.

[Moy and Specter 2015] R. Moy and J. Specter, "There exist non-CM Hilbert modular forms of partial weight 1", preprint, 2015. arXiv 1407.3872

[Newton 2013] J. Newton, "Completed cohomology of Shimura curves and a p-adic Jacquet–Langlands correspondence", *Math. Ann.* **355**:2 (2013), 729–763. MR 3010145 Zbl 1304.22025

[Newton 2014] J. Newton, "Level raising for p-adic Hilbert modular forms", preprint, 2014. arXiv 1409.6533

[Paulin 2011] A. G. M. Paulin, "Local to global compatibility on the eigencurve", *Proc. Lond. Math. Soc.* (3) **103**:3 (2011), 405–440. MR 2012i:11047 Zbl 1290.11080

[Pilloni and Stroh 2011] V. Pilloni and B. Stroh, "Surconvergence et classicité: le cas Hilbert", preprint, 2011, http://perso.ens-lyon.fr/vincent.pilloni/surconv_hilbert.pdf.

[Pilloni and Stroh 2013] V. Pilloni and B. Stroh, "Surconvergence, ramification et modularité", preprint, 2013, http://perso.ens-lyon.fr/vincent.pilloni/Artinfinal.pdf.

[Ribet 1977] K. A. Ribet, "Galois representations attached to eigenforms with Nebentypus", pp. 17–51 in *Modular functions of one variable, V* (Univ. Bonn, 1976), edited by J.-P. Serre and D. Zagier, Lecture Notes in Mathematics **601**, Springer, Berlin, 1977. MR 56 #11907 Zbl 0363.10015

[Rogawski and Tunnell 1983] J. D. Rogawski and J. B. Tunnell, "On Artin L-functions associated to Hilbert modular forms of weight one", *Invent. Math.* **74**:1 (1983), 1–42. MR 85i:11044 Zbl 0523.12009

[Sasaki 2014] S. Sasaki, "Integral models of Hilbert modular varieties in the ramified case, deformations of modular Galois representations, and weight one forms", preprint, 2014, http://www.cantabgold.net/users/s.sasaki.03/hmv1-3-14.pdf.

[Taylor 1989] R. Taylor, "On Galois representations associated to Hilbert modular forms", *Invent. Math.* **98**:2 (1989), 265–280. MR 90m:11176 Zbl 0705.11031

[Taylor 1995] R. Taylor, "On Galois representations associated to Hilbert modular forms, II", pp. 185–191 in *Elliptic curves, modular forms & Fermat's last theorem* (Hong Kong, 1993), edited by J. Coates, Ser. Number Theory **I**, Int. Press, Cambridge, MA, 1995. MR 96j:11073 Zbl 0836.11017

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

Communicated by Brian Conrad
Received 2014-09-24 Revised 2015-02-06 Accepted 2015-03-27

j.newton@imperial.ac.uk *Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.981

Horrocks correspondence on arithmetically Cohen–Macaulay varieties

Francesco Malaspina and A. Prabhakar Rao

We describe a vector bundle $\mathscr{E}$ on a smooth n-dimensional arithmetically Cohen–Macaulay variety in terms of its cohomological invariants $H^i_*(\mathscr{E})$, $1 \le i \le n-1$, and certain graded modules of "socle elements" built from $\mathscr{E}$. In this way we give a generalization of the Horrocks correspondence. We prove existence theorems, where we construct vector bundles from these invariants, and uniqueness theorems, where we show that these data determine a bundle up to isomorphism. The cases of the quadric hypersurface in $\mathbb{P}^{n+1}$ and the Veronese surface in $\mathbb{P}^5$ are considered in more detail.

Introduction

In a fundamental paper, Horrocks [1964] described all vector bundles on projective space $\mathbb{P}^n$ in terms of their intermediate cohomology modules. He described these cohomology modules using what he called a $\mathfrak{Z}$-complex, and showed that the category of vector bundles modulo stable equivalence was equivalent to the category of all $\mathfrak{Z}$-complexes modulo exact free complexes. In particular, this gives the well-known Horrocks criterion for a vector bundle to be a sum of line bundles in terms of the vanishing of its intermediate cohomology. His results were reformulated by Walters [1996] in the language of derived categories, and extended to sheaves by Coandă [2010]. Beĭlinson [1978] described the derived category of sheaves on a projective space using complexes built from an "exceptional sequence" $\{\mathbb{O}_{\mathbb{P}^n}(-n), \dots, \mathbb{O}_{\mathbb{P}^n}(-1), \mathbb{O}_{\mathbb{P}^n}\}$ of line bundles on $\mathbb{P}^n$, and Kapranov [1988] gave a similar description for smooth quadric hypersurfaces by enlarging the sequence to include the spinor bundles Σ of the quadric. Ancona and Ottaviani [1991] used these methods to extend the Horrocks splitting criterion to quadrics, with a theorem that a vector bundle $\mathscr{E}$ on a quadric $\mathfrak{Q}_n$ (of dimension n) is a sum of line bundles if and only if $\mathscr{E}$ has its intermediate cohomology modules $H^i_*(\mathscr{E})$ all zero for $1 \le i \le n-1$ and also $H^{n-1}_*(\mathscr{E} \otimes \Sigma) = 0$ for the spinor bundles Σ.

MSC2010: primary 14F05; secondary 14J60.
Keywords: vector bundles, cohomology modules, Horrocks correspondence, smooth ACM varieties.

In this paper, we copy Horrocks' method on a smooth arithmetically Cohen–Macaulay (ACM) subvariety X of projective space. Given a vector bundle $\mathscr{E}$ on X, we construct a $\mathfrak{Z}$-complex of free A-modules (where A is the coordinate ring of X). This complex, when sheafified, gives rise to a vector bundle $\mathscr{F}$ on X which we call a Horrocks data bundle for $\mathscr{E}$, since it comes with a map $\beta : \mathscr{F} \to \mathscr{E}$ which is an isomorphism on intermediate cohomology modules. When $H^0_*(\beta)$ is surjective, the kernel of β is some ACM bundle on X.

These methods of Horrocks provide for ACM varieties a vector-bundle version of results of Auslander and Bridger [1969, Proposition 4.26, Corollary 4.27], who gave a structure theorem for a module M of finite Gorenstein dimension n over a commutative ring, showing that $M \oplus P$ for some projective module P can be expressed as an extension of a module $H_n(M)$ of projective dimension n by a module of zero Gorenstein dimension, where the map $M \to H_n(M)$ satisfies a universal property. In an unpublished preprint, Buchweitz [1986] proved a similar result for finitely generated modules over strongly Gorenstein (noncommutative) rings. We will see that the graded A-module F of global sections of the Horrocks data bundle $\mathscr{F}$ will have $F^\vee$ of finite projective dimension.

With this natural extension of Horrocks' arguments to an ACM variety, we give a generalization of the Horrocks correspondence in Section 1. Our goal in looking at a Horrocks correspondence on X is to look for cohomological invariants that determine $\mathscr{E}$. We will take the Horrocks data bundle as encoding all the intermediate cohomology for $\mathscr{E}$, and view it as one of the invariants. So we will study the bundles $\mathscr{E}$ with a fixed (minimal) Horrocks data bundle $\mathscr{F}$. While for the map $\mathscr{F} \to \mathscr{E}$ the induced map of first cohomology modules $H^1_*(\mathscr{F}) \to H^1_*(\mathscr{E})$ is an isomorphism, for various irreducible ACM bundles $\mathscr{B}$ on X, the map $H^1_*(\mathscr{F} \otimes \mathscr{B}^\vee) \to H^1_*(\mathscr{E} \otimes \mathscr{B}^\vee)$ may have a kernel. These kernels will give more cohomological invariants and we will call them modules of $\mathscr{B}$-socle elements. In Theorems 1.10 and 1.11, we see how these invariants determine $\mathscr{E}$ up to direct sums of ACM bundles. We also give a splitting criterion for the bundle $\mathscr{E}$ to be a sum of line bundles restricted from projective space. What is lacking in Section 1 is an understanding of which modules of socle elements are obtained from a vector bundle for a general ACM variety.

In Section 2 we describe the case of quadrics, on which ACM bundles are well understood due to Knörrer [1987]. In particular, for the spinor bundles Σ_i on a quadric $\mathscr{Q}_n$, modules of Σ_i-socle elements of a Horrocks data bundle $\mathscr{F}$ are just graded vector spaces. We show that a vector bundle $\mathscr{E}$ exists for each choice of Horrocks data bundle $\mathscr{F}$ and vector spaces V_i of Σ_i-socle elements of $\mathscr{F}$, and that two vector bundles with the same data of $\mathscr{F}$, V_i (up to obvious isomorphisms) are isomorphic up to direct sums of ACM bundles. In this way we generalize the results obtained in [Malaspina and Rao 2014] on $\mathscr{Q}_2$.

In Section 3 we deal with the Veronese surface $\mathcal{V} \subset \mathbb{P}^5$. The study of vector bundles on $\mathcal{V}$ is trivial by Horrocks if we view $\mathcal{V}$ as $\mathbb{P}^2$. But, as another illustration of the methods, it is an interesting example of an arithmetically Cohen–Macaulay embedding which is not arithmetically Gorenstein and for which the ACM bundles are easy to handle.

1. Horrocks data bundles on ACM varieties

Let X be a smooth ACM variety of dimension n in $\mathbb{P}^{n+r}$ over a field k. For any sheaf $\mathcal{B}$ on X, $H^i_*(\mathcal{B})$ will denote $\bigoplus_{l\in\mathbb{Z}} H^i(X, \mathcal{B}(l))$. The coordinate ring of X, $A = H^0_*(\mathcal{O}_X)$, is a noetherian Cohen–Macaulay graded k-algebra. $H^i_*(\mathcal{B})$ is a graded module over A. Let M be the category of graded, finitely generated A-modules and graded homomorphisms. Any finitely generated projective graded A module has the form $\bigoplus_i A(a_i)$ for some shifts $a_i \in \mathbb{Z}$ in grading, and will be called a free A-module. Let $\mathfrak{P} \subset$ M be the full subcategory of finitely generated free A-modules. $\mathcal{C}^-(\mathrm{M})$ and $\mathcal{C}^-(\mathfrak{P})$ will denote the categories of all complexes, bounded above, of objects in M and $\mathfrak{P}$ respectively, where morphisms are maps between two complexes. Since M has enough projectives, given a complex $C^\bullet$ of objects in M, bounded above, one can find a free resolution, i.e., a complex $P^\bullet$ in $\mathcal{C}^-(\mathfrak{P})$ with a quasi-isomorphism $P^\bullet \to C^\bullet$.

Let $\mathcal{E} \in$ VB be an object in the category of finite-rank vector bundles on X. $H^i_*(\mathcal{E})$ is an A-module of finite length for $1 \le i \le n-1$. A vector bundle will be called free if it has the form $\bigoplus_i \mathcal{O}_X(a_i)$. A vector bundle $\mathcal{E}$ will be called ACM (arithmetically Cohen–Macaulay) if $H^i_*(\mathcal{E}) = 0$ for all $1 \le i \le n-1$. Since X is ACM, every free bundle is ACM. By Serre duality, the line bundle ω_X is an ACM line bundle.

Given $\mathcal{E}$, let E denote the graded A-module $H^0_*(\mathcal{E})$. Denoting duals by $^\vee$ in the categories VB and M, we have $H^0_*(\mathcal{E}^\vee) \cong (H^0_*(\mathcal{E}))^\vee$. Following Horrocks, we choose a resolution of $H^0_*(\mathcal{E}^\vee)$ by finitely generated free modules

$$\ldots \longrightarrow C^{3\vee} \longrightarrow C^{2\vee} \longrightarrow C^{1\vee} \longrightarrow C^{0\vee} \longrightarrow H^0_*(\mathcal{E}^\vee) \longrightarrow 0. \tag{1}$$

In [Horrocks 1964], this could be chosen as a finite resolution, but in our case it may be infinite. However, if $K = \ker(C^{n-2\vee} \to C^{n-3\vee})$, then $\mathcal{K}$ is an ACM vector bundle on X, where $\mathcal{K} = \widetilde{K}$ is the sheaf obtained from K. Replacing the terms up to and including $C^{n-1\vee}$ by K and dualizing, we get the complex

$$C^\bullet_{\{0,n\}} : 0 \longrightarrow C^0 \xrightarrow{\delta^1_{C^\bullet}} C^1 \xrightarrow{\delta^2_{C^\bullet}} C^2 \xrightarrow{\delta^3_{C^\bullet}} \cdots \xrightarrow{\delta^{n-2}_{C^\bullet}} C^{n-2} \longrightarrow K^\vee \longrightarrow 0. \tag{2}$$

The exact sequence (1), when sheafified, gives an exact sequence of vector bundles, and its dual gives the exact sequence of vector bundles

$$0 \longrightarrow \mathcal{E} \longrightarrow \widetilde{C}^0 \xrightarrow{\delta^1_{C^\bullet}} \widetilde{C}^1 \xrightarrow{\delta^2_{C^\bullet}} \widetilde{C}^2 \xrightarrow{\delta^3_{C^\bullet}} \cdots \xrightarrow{\delta^{n-2}_{C^\bullet}} \widetilde{C}^{n-2} \longrightarrow \mathcal{K}^\vee \longrightarrow 0. \tag{3}$$

From this it becomes evident that $E = H^0_*(\mathscr{E})$ is given as $H^0(C^\bullet_{\{0,n\}})$, and $H^i_*(\mathscr{E}) = H^i(C^\bullet_{\{0,n\}})$ for $i = 1, \dots n-1$ (where $C^{n-1}_{\{0,n\}}$ is understood to refer to $K^\vee$).

E itself has a free resolution (again possibly infinite). Splice $C^\bullet_{\{0,n\}}$ with a free resolution $L^\bullet$ of E and call the resulting complex $C^\bullet$. The complex $C^\bullet$ is bounded above and has the property that $H^i(C^\bullet) = H^i_*(\mathscr{E})$ for $i = 1, \dots n-1$ and equals 0 for other values of i.

Choose a free resolution $P^\bullet$ in $\mathscr{C}^-(\mathfrak{P})$ of $C^\bullet$:

$$\begin{array}{ccccccccccccccc}
P^\bullet : \cdots & \longrightarrow & P^{-2} & \longrightarrow & P^{-1} & \longrightarrow & P^0 & \xrightarrow{\delta^1_{P^\bullet}} & P^1 & \xrightarrow{\delta^2_{P^\bullet}} & \cdots & \xrightarrow{\delta^{n-2}_{P^\bullet}} & P^{n-2} & \longrightarrow & P^{n-1} \longrightarrow 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
C^\bullet : \cdots & \longrightarrow & L^{-2} & \longrightarrow & L^{-1} & \longrightarrow & C^0 & \xrightarrow{\delta^1_{C^\bullet}} & C^1 & \xrightarrow{\delta^2_{C^\bullet}} & \cdots & \xrightarrow{\delta^{n-2}_{C^\bullet}} & C^{n-2} & \longrightarrow & K^\vee \longrightarrow 0
\end{array}$$

Then $P^\bullet$ is an element in $\mathscr{C}^-(\mathfrak{P})$ with the property that $H^i(P^\bullet)$ is an A-module of at most finite length for $1 \le i \le n-1$ and is zero for other i. In [Horrocks 1964] the bounded version of such a free complex was called a $\mathfrak{Z}$-complex, while Walters [1996] called the category of such complexes $\mathrm{FinL}(\mathfrak{P})$. In our setting, we will call it a Horrocks data complex and use the notation of [Walter 1996]. We also define a "Horrocks data bundle" for each such Horrocks data complex:

Definition 1.1. $\mathrm{FinL}^-(\mathfrak{P})$ is the full subcategory of all complexes $P^\bullet$ in $\mathscr{C}^-(\mathfrak{P})$ with the property that $H^i(P^\bullet)$ is an A-module of at most finite length for $1 \le i \le n-1$ and is zero for other i. A complex $P^\bullet$ in $\mathrm{FinL}^-(\mathfrak{P})$ will be called a Horrocks data complex. For such a complex, let $F = \ker(\delta^1_{P^\bullet} : P^0 \to P^1)$. Then the sheaf $\mathscr{F} = \widetilde{F}$ will be called a Horrocks data bundle on X.

It should be clear that the above $\mathscr{F}$ is a vector bundle on X with the property that $H^i_*(\mathscr{F}) = H^i(P^\bullet)$ for $1 \le i \le n-1$.

Lemma 1.2 [Horrocks 1964, Theorem 7.2]. *$F^\vee$ has a finite free resolution.*

Proof. Horrocks' proof cited above is when A is a regular ring, but remains valid when A is Cohen–Macaulay. Another proof (indicated by the referee) is: $0 \to (P^{n-1})^\vee \to (P^{n-2})^\vee \to \cdots \to (P^0)^\vee \to F^\vee \to 0$ is a complex in M, locally free and exact away from the maximal ideal for the vertex of the cone over X, and hence is exact by the Peskine–Szpiro acyclicity lemma. □

Since the modules of global sections of a nonfree ACM bundle and of its dual bundle on X have infinite projective dimension over A, it follows that a Horrocks data bundle $\mathscr{F}$ can have no nonfree ACM bundle or its dual as a summand.

Since any $P^\bullet$ in $\mathscr{C}^-(\mathfrak{P})$ decomposes as $M^\bullet \oplus L^\bullet$, where $M^\bullet$ is a minimal free complex and $L^\bullet$ is an acyclic free complex, we get $\mathscr{F} = \mathscr{F}_{\min} \oplus \mathscr{L}$, where $\mathscr{F}$, $\mathscr{F}_{\min}$, and $\mathscr{L}$ are the Horrocks data bundles corresponding to $P^\bullet$, $M^\bullet$, and $L^\bullet$ respectively.

$\mathscr{L}$ is a free bundle and $\mathscr{F}_{\min}$ will be called a "minimal" Horrocks data bundle. The projective space version of the following isomorphism theorem can be found in [Horrocks 1964, Theorem 7.5, Proposition 9.5] or [Walter 1996, Lemma 2.11].

Proposition 1.3. *Let $\sigma : \mathscr{F} \to \mathscr{F}'$ be a homomorphism between two minimal Horrocks data bundles on X such that σ induces isomorphisms $H^i_*(\mathscr{F}) \to H^i_*(\mathscr{F}')$ for $1 \le i \le n-1$. Then σ is an isomorphism.*

Proof. The proofs of the results cited above work in our ACM setting as well. □

Returning to the vector bundle $\mathscr{E}$, let $P^\bullet$ be a free resolution of $C^\bullet$ as described above. Let $P^\bullet_{\ge 0}$ denote the naive truncation of $P^\bullet$ at the zeroth term. We get the induced homomorphism of complexes

$$P^\bullet_{\ge 0} \to C^\bullet_{\{0,n\}}.$$

For F defined as $\ker \delta^1_{P^\bullet}$, there is an induced homomorphism $F \to E$. For the Horrocks data bundle $\mathscr{F} = \widetilde{F}$, we get a homomorphism $\beta : \mathscr{F} \to \mathscr{E}$ which induces isomorphisms $H^i_*(\mathscr{F}) \to H^i_*(\mathscr{E})$ for $1 \le i \le n-1$. Hence any vector bundle $\mathscr{E}$ has a "Horrocks datum", as we now define:

Definition 1.4. Let $\mathscr{E}$ be a vector bundle on X. A pair $(\mathscr{F}, \beta)$ will be called a Horrocks datum for $\mathscr{E}$ if $\mathscr{F}$ is a Horrocks data bundle and β is a homomorphism $\beta : \mathscr{F} \to \mathscr{E}$ which induces isomorphisms $H^i_*(\mathscr{F}) \to H^i_*(\mathscr{E})$ for $1 \le i \le n-1$.

A point on terminology: Auslander's approximation theorem [Auslander and Bridger 1969, Proposition 4.26, Corollary 4.27] quoted in the introduction states that, given a module M of finite Gorenstein dimension n, there exist a projective module P, a module $H_n(M)$ of projective dimension n, a module M_n of zero Gorenstein dimension and an exact sequence $0 \to M_n \to M \oplus P \to H_n(M) \to 0$. Following Auslander's suggestion, Buchweitz [1986, Corollary 5.3.3] called $H_n(M)$ (with the map $M \to H_n(M)$) a "hull of finite projective dimension" for M, and M_n the maximal Cohen–Macaulay approximation to M.

In the case where our variety X is arithmetically Gorenstein, Auslander's sequence can be seen as coming from the dual of the η-sequence of Theorem 1.7 below: given E, the η-sequence $0 \to K \to F \to E \to 0$ dualized gives $0 \to E^\vee \to F^\vee \to K^\vee$, where $F^\vee$ has finite projective dimension. When X is arithmetically Gorenstein, $K^\vee$ is a maximal Cohen–Macaulay module and $F^\vee \to K^\vee$ is surjective. Pull back the exact sequence by a surjection $L \to K^\vee \to 0$, with L projective. It splits. This induces an exact sequence $0 \to N \to E^\vee \oplus L \to F^\vee \to 0$, where N (the kernel of $L \to K^\vee$) is a maximal Cohen–Macaulay module. This fits the above approximation theorem for $E^\vee$.

However, we have chosen the notation "$\mathscr{F}$ is the Horrocks data bundle for $\mathscr{E}$" since $\mathscr{F}$ encodes all the intermediate cohomology data of $\mathscr{E}$.

Theorem 1.5. *Let $\mathscr{E}_1, \mathscr{E}_2$ be vector bundles on X with Horrocks data $(\mathscr{F}_1, \beta_1)$, $(\mathscr{F}_2, \beta_2)$ respectively. Let $\sigma : \mathscr{E}_1 \to \mathscr{E}_2$ be a homomorphism.*

(1) *There is a free bundle $\mathscr{L}$ and a commuting square*

$$\begin{tikzcd} \mathscr{F}_1 \arrow[r] \arrow[d, "\beta_1"] & \mathscr{F}_2 \oplus \mathscr{L} \arrow[d, "(\beta_2,*)"] \\ \mathscr{E}_1 \arrow[r, "\sigma"'] & \mathscr{E}_2 \end{tikzcd}$$

(2) *If $H^0_*(\beta_2) : H^0_*(\mathscr{F}_2) \to H^0_*(\mathscr{E}_2)$ is surjective, the free bundle $\mathscr{L}$ can be chosen to be zero.*

Proof. It is straightforward to see that the construction of the complex $C^\bullet$ out of the vector bundle $\mathscr{E}$ is functorial in the sense that, given $\sigma : \mathscr{E}_1 \to \mathscr{E}_2$, there is an induced morphism from $C_1^\bullet \to C_2^\bullet$ with the property that the homomorphisms $H^i(C_1^\bullet) \to H^i(C_2^\bullet)$ coincide with $H^i(\sigma) : H^i_*(\mathscr{E}_1) \to H^i_*(\mathscr{E}_2)$ for $1 \le i \le n-1$. In the special case of $\beta_k : \mathscr{F}_k \to \mathscr{E}_k$, a Horrocks datum, we get a quasi-isomorphism $P_k^\bullet \to \mathscr{C}_k^\bullet$, where $P_k^\bullet$ is the Horrocks data complex associated to $\mathscr{F}_k$, so that $P_k^\bullet \to \mathscr{C}_k^\bullet$ is a free resolution of $\mathscr{C}_k^\bullet$. Now given a morphism of complexes $C_1^\bullet \to C_2^\bullet$, we can lift the morphism to their free resolutions, after adding a free acyclic complex to $P_2^\bullet$. This gives the commuting square of part (1). The proof of part (2) is elementary. $\square$

The following theorems (Theorems 1.6 and 1.7) are to be found in more general form in [Buchweitz 1986] as the "syzygy theorem for Gorenstein rings". The diagram in Theorem 1.8 below is Buchweitz's octahedron [1986, (5.3.1)].

Theorem 1.6 (γ sequence for $\mathscr{E}$). *Let $\mathscr{E}$ be a vector bundle on X and $(\mathscr{F}, \beta)$ a Horrocks datum for $\mathscr{E}$. From the Horrocks data complex $P^\bullet$ for $\mathscr{F}$, consider the exact sequence $\Psi : 0 \to \mathscr{F} \to \mathscr{P}^0 \to \mathscr{G} \to 0$, where $\mathscr{P}^0 = \widetilde{P}^0$ and $\mathscr{G} = \widetilde{G}$ with $G = \ker \delta^2_{P^\bullet}$. We define γ as the pushout of Ψ by β*

$$\begin{tikzcd} \Psi : 0 \arrow[r] & \mathscr{F} \arrow[r] \arrow[d, "\beta"] & \mathscr{P}^0 \arrow[r] \arrow[d] & \mathscr{G} \arrow[r] \arrow[d, equal] & 0 \\ \gamma : 0 \arrow[r] & \mathscr{E} \arrow[r, "f"'] & \mathscr{A} \arrow[r, "g"'] & \mathscr{G} \arrow[r] & 0 \end{tikzcd}$$

(1) *Given two bundles $\mathscr{E}_1, \mathscr{E}_2$, a morphism $\sigma : \mathscr{E}_1 \to \mathscr{E}_2$, and Horrocks data $(\mathscr{F}_1, \beta_1)$, $(\mathscr{F}_2, \beta_2)$ for each bundle, we obtain a commuting box of short exact*

sequences (using obvious notation)

$$\begin{array}{ccc} \Psi_1 & \longrightarrow & \Psi_2 \oplus \lambda \\ \downarrow{\scriptstyle \beta_1} & & \downarrow{\scriptstyle (\beta_2,*)} \\ \gamma_1 & \xrightarrow[\sigma]{} & \gamma_2 \end{array}$$

where λ is a short exact sequence $0 \to \mathscr{L} \to \mathscr{L} \to 0 \to 0$ of free bundles. If $H^0_*(\beta_2)$ is surjective onto $H^0_*(\mathscr{E}_2)$, λ may be taken to be zero.

(2) $H^{n-1}_*(\mathscr{G}) = 0$, *and $\mathscr{A}$ is an ACM bundle on X.*

(3) *Up to a short exact sequence $0 \to 0 \to \mathscr{L} \to \mathscr{L} \to 0$ of free bundles, the sequence γ depends only on $\mathscr{E}$ and not on the choice of Horrocks datum.*

Proof. (1) σ lifts to a map $\mathscr{F}_1 \to \mathscr{F}_2 \oplus \mathscr{L}$ to give a commuting square, by Theorem 1.5. $\mathscr{F}_2 \oplus \mathscr{L}$ is a Horrocks data bundle for the Horrocks data complex, where P^0 is replaced by $P^0 \oplus \mathscr{L}$ but with the same bundle $\mathscr{G}_2$. It is easy to see that the map $\mathscr{F}_1 \to \mathscr{F}_2 \oplus \mathscr{L}$ extends to a map of sequences $\Psi_1 \to \Psi_2 \oplus \lambda$. The pushouts of Ψ_2 and $\Psi_2 \oplus \lambda$ give the same sequence γ_2. Lastly, since we have a commuting square from the first line of the proof, the pushouts of Ψ_1 and $\Psi_2 \oplus \lambda$ give a commuting box of exact sequences.

(2) By construction, $H^{n-1}_*(\mathscr{G}) = H^n(P^\bullet) = 0$. Since we have isomorphisms $H^i_*(\mathscr{G}) \cong H^{i+1}_*(\mathscr{F}) \cong H^{i+1}_*(\mathscr{E})$ for $1 \le i \le n-2$ and $H^0_*(\mathscr{G}) \twoheadrightarrow H^1_*(\mathscr{F}) \cong H^1_*(\mathscr{E})$, we conclude that $\mathscr{A}$ is ACM.

(3) This follows from the first part when we apply the previous theorem to the identity morphism from $\mathscr{E}$ to $\mathscr{E}$. Indeed, the theorem, together with Proposition 1.3, shows that any two Horrocks data bundles for $\mathscr{E}$ are stably equivalent. □

Theorem 1.7 (η sequence for $\mathscr{E}$). *Let $(\mathscr{F}, \beta)$ be a Horrocks datum for the bundle $\mathscr{E}$ such that $H^0_*(\beta)$ is surjective. We define the η sequence for $\mathscr{E}$ to be*

$$0 \longrightarrow \mathscr{K} \longrightarrow \mathscr{F} \xrightarrow{\beta} \mathscr{E} \longrightarrow 0,$$

where $\mathscr{K}$ is the kernel bundle.

(1) *$\mathscr{K}$ is an ACM bundle.*

(2) *η is determined by $\mathscr{E}$ up to a short exact sequence $0 \to \mathscr{L} \to \mathscr{L} \to 0 \to 0$ of free bundles.*

(3) *Given a morphism $\sigma : \mathscr{E}_1 \to \mathscr{E}_2$, there is an induced morphism of short exact sequences $\eta_1 \to \eta_2$.*

Proof. The proof is easy. We just mention that the induced map $\eta_1 \to \eta_2$ depends on the choice of a map from $\mathscr{F}_1$ to $\mathscr{F}_2$ that lifts σ (as obtained from Theorem 1.5). □

Theorem 1.8 (diagram of $\mathscr{E}$). *Let $(\mathscr{F}, \beta)$ be a Horrocks datum for the bundle $\mathscr{E}$ such that $H^0_*(\beta)$ is surjective. The γ and η sequences of $\mathscr{E}$ fit into a diagram for $\mathscr{E}$*

$$\begin{tikzcd}
 & & 0 \arrow[d] & 0 \arrow[d] & & \\
 & & \mathscr{K} \arrow[r, equal] \arrow[d, "\alpha"] & \mathscr{K} \arrow[d] & & \\
\Psi: & 0 \arrow[r] & \mathscr{F} \arrow[r, "g"'] \arrow[d, "\beta"] & \mathscr{P}^0 \arrow[r] \arrow[d] & \mathscr{G} \arrow[r] \arrow[d, equal] & 0 \\
\gamma: & 0 \arrow[r] & \mathscr{E} \arrow[r, "f"'] \arrow[d] & \mathscr{A} \arrow[r] \arrow[d] & \mathscr{G} \arrow[r] & 0 \\
 & & 0 & 0 & & \\
 & & \eta & \Delta & &
\end{tikzcd}$$

Given a morphism $\sigma : \mathscr{E}_1 \to \mathscr{E}_2$, there is an induced map from the diagram of $\mathscr{E}_1$ to the diagram of $\mathscr{E}_2$.

Proof. While the existence of the diagram is clear, the map from diagram of $\mathscr{E}_1$ to the diagram of $\mathscr{E}_2$ with appropriate commuting boxes exists because the choice of a map from $\mathscr{F}_1$ to $\mathscr{F}_2$ that lifts σ will determine $\eta_1 \to \eta_2$ and then allows a choice of a map $\Psi_1 \to \Psi_2$. This now gives the commuting box of short exact sequences of Theorem 1.6. □

The following is a criterion for obtaining a map between two γ-sequences:

Proposition 1.9. *Let $\mathscr{E}, \mathscr{E}'$ be two vector bundles with the same (minimal) Horrocks data bundle $\mathscr{F}_{\min}$ and Horrocks data $(\mathscr{F}_{\min}, \beta)$, $(\mathscr{F}_{\min}, \beta')$. Let $\mathscr{B}_1, \mathscr{B}_2, \dots, \mathscr{B}_k$ be the distinct nonfree irreducible ACM bundles (up to twists by $\mathscr{O}_X(a)$) that appear as summands in the middle term $\mathscr{A}_{\mathscr{E}}$ of the γ-sequence of $\mathscr{E}$. For each $\mathscr{B}_i$, let V_i be the kernel of the map $H^1_*(\beta \otimes 1_{\mathscr{B}_i^\vee})$ from $H^1_*(\mathscr{F}_{\min} \otimes \mathscr{B}_i^\vee)$ to $H^1_*(\mathscr{E} \otimes \mathscr{B}_i^\vee)$, and let V_i' be the same with β replaced by β'. If $V_i \subseteq V_i'$ for all i, then there exists a map $\phi : \mathscr{E} \to \mathscr{E}'$ such that the γ-sequence of $\mathscr{E}'$ is the pushout by ϕ of the γ-sequence for $\mathscr{E}$.*

Proof. Since the γ-sequences γ, γ' are pushouts by β, β' of the Ψ-sequence for $\mathscr{F}_{\min}$

$$\Psi : 0 \longrightarrow \mathscr{F}_{\min} \longrightarrow \mathscr{P}^0 \longrightarrow \mathscr{G}_{\min} \longrightarrow 0$$

in the commutative diagram

$$\begin{tikzcd}
\mathrm{Hom}(\mathscr{P}^0, \mathscr{E}') \arrow[r] & \mathrm{Hom}(\mathscr{F}_{\min}, \mathscr{E}') \arrow[r, "\delta(\Psi)"'] & \mathrm{Ext}^1(\mathscr{G}_{\min}, \mathscr{E}') \arrow[r] & \mathrm{Ext}^1(\mathscr{P}^0, \mathscr{E}') \\
 & \mathrm{Hom}(\mathscr{E}, \mathscr{E}') \arrow[u, "\beta"'] \arrow[r, "\delta(\gamma)"'] & \mathrm{Ext}^1(\mathscr{G}_{\min}, \mathscr{E}') \arrow[u, equal] \arrow[r] & \mathrm{Ext}^1(\mathscr{A}_{\mathscr{E}}, \mathscr{E}') \arrow[u]
\end{tikzcd}$$

it suffices to show that $\gamma' \in \mathrm{Ext}^1(\mathscr{G}_{\min}, \mathscr{E}')$ maps to zero in $\mathrm{Ext}^1(\mathscr{A}_{\mathscr{E}}, \mathscr{E}')$, for then there is an element $\sigma \in \mathrm{Hom}(\mathscr{E}, \mathscr{E}')$ such that $\sigma \circ \beta$ differs from β' by a map that factors through $\mathscr{P}^0$.

Let $\rho : \mathscr{A}_{\mathscr{E}} \to \mathscr{G}_{\min}$ be the map occurring in the γ-sequence of $\mathscr{E}$. Then under the connecting homomorphism for $\gamma \otimes \mathscr{A}_{\mathscr{E}}^{\vee}$, ρ maps to zero under $H^0_*(\mathscr{G}_{\min} \otimes \mathscr{A}_{\mathscr{E}}^{\vee}) \to H^1_*(\mathscr{E} \otimes \mathscr{A}_{\mathscr{E}}^{\vee})$. Hence, under the connecting homomorphism of $\Psi \otimes \mathscr{A}_{\mathscr{E}}^{\vee}$, ρ maps to the kernel of $H^1_*(\mathscr{F}_{\min} \otimes \mathscr{A}_{\mathscr{E}}^{\vee}) \to H^1_*(\mathscr{E} \otimes \mathscr{A}_{\mathscr{E}}^{\vee})$. By the assumption $V_i \subseteq V_i'$ for all i, ρ also maps to the kernel of $H^1_*(\mathscr{F}_{\min} \otimes \mathscr{A}_{\mathscr{E}}^{\vee}) \to H^1_*(\mathscr{E}' \otimes \mathscr{A}_{\mathscr{E}}^{\vee})$. It follows that the pullback of γ' by ρ splits, which was the desired result. $\square$

This criterion leads to an isomorphism theorem on X:

Theorem 1.10 (isomorphism theorem). *Let $\mathscr{E}, \mathscr{E}'$ be two vector bundles on X, with the same minimal Horrocks data bundle $\mathscr{F}_{\min}$ and Horrocks data $(\mathscr{F}_{\min}, \beta)$, $(\mathscr{F}_{\min}, \beta')$. Let $\mathscr{B}_1, \mathscr{B}_2, \dots, \mathscr{B}_k$ be the distinct nonfree irreducible ACM bundles (up to twists by $\mathscr{O}_X(a)$) that appear as summands in either of the middle terms $\mathscr{A}_{\mathscr{E}}, \mathscr{A}_{\mathscr{E}'}$ of the γ-sequences of $\mathscr{E}, \mathscr{E}'$. If for each i the kernel of $H^1_*(\beta \otimes 1_{\mathscr{B}_i^{\vee}})$ equals the kernel of $H^1_*(\beta' \otimes 1_{\mathscr{B}_i^{\vee}})$ and if $\mathscr{E}$ and $\mathscr{E}'$ have no ACM summands, then $\mathscr{E} \cong \mathscr{E}'$.*

Proof. If $\mathscr{F}$ is free, $\mathscr{E}, \mathscr{E}'$ are ACM and the theorem does not apply. So we will assume that $\mathscr{F}_{\min}$ is a nonfree minimal Horrocks data bundle. By applying Proposition 1.9, there exists a homomorphism $\sigma : \mathscr{E} \to \mathscr{E}'$ and a commutative diagram of γ-sequences

$$\begin{tikzcd}
0 \arrow[r] & \mathscr{E} \arrow[r] \arrow[d, "\sigma"] & \mathscr{A}_{\mathscr{E}} \arrow[r] \arrow[d, "\sigma_1"] & \mathscr{G}_{\min} \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & \mathscr{E}' \arrow[r] & \mathscr{A}_{\mathscr{E}'} \arrow[r] & \mathscr{G}_{\min} \arrow[r] & 0
\end{tikzcd}$$

Tensor the diagram by $\mathscr{B}^{\vee}$, where $\mathscr{B}$ will stand for any of the distinct irreducible ACM bundles (up to twists by $\mathscr{O}_X(a)$) that appear as summands in $\mathscr{A}_{\mathscr{E}'}$, including the possible free line bundle $\mathcal{O}_X$. In the induced diagram of cohomology, we get

$$\begin{array}{ccc}
0 & & 0 \\
\downarrow & & \downarrow \\
H^0_*(\mathcal{E}\otimes\mathcal{B}^\vee) & \xrightarrow[\sigma]{} & H^0_*(\mathcal{E}'\otimes\mathcal{B}^\vee) \\
\downarrow & & \downarrow \\
H^0_*(\mathcal{A}_{\mathcal{E}}\otimes\mathcal{B}^\vee) & \xrightarrow[\sigma_1]{} & H^0_*(\mathcal{A}_{\mathcal{E}'}\otimes\mathcal{B}^\vee) \\
\downarrow & & \downarrow \\
H^0_*(\mathcal{G}_{\min}\otimes\mathcal{B}^\vee) & = & H^0_*(\mathcal{G}_{\min}\otimes\mathcal{B}^\vee) \\
\downarrow & & \downarrow \\
H^1_*(\mathcal{E}\otimes\mathcal{B}^\vee) & \xrightarrow[\sigma]{} & H^1_*(\mathcal{E}'\otimes\mathcal{B}^\vee) \\
\downarrow & & \downarrow \\
H^1_*(\mathcal{A}_{\mathcal{E}}\otimes\mathcal{B}^\vee) & \xrightarrow[\sigma_1]{} & H^1_*(\mathcal{A}_{\mathcal{E}'}\otimes\mathcal{B}^\vee)
\end{array}$$

The map $H^0_*(\mathcal{G}_{\min}\otimes\mathcal{B}^\vee)\to H^1_*(\mathcal{E}\otimes\mathcal{B}^\vee)$ factors through $H^1_*(\mathcal{F}\otimes\mathcal{B}^\vee)$, since γ is the pushout of Ψ by β. The condition of equality of kernels for $H^1_*(\beta\otimes 1_{\mathcal{B}^\vee})$ and $H^1_*(\beta'\otimes 1_{\mathcal{B}^\vee})$ implies that the kernel in $H^0_*(\mathcal{G}_{\min}\otimes\mathcal{B}^\vee)$ is the same for $\mathcal{E}$ and $\mathcal{E}'$. Therefore the mapping cone map $H^0_*(\mathcal{E}'\otimes\mathcal{B}^\vee)\oplus H^0_*(\mathcal{A}_{\mathcal{E}}\otimes\mathcal{B}^\vee)\to H^0_*(\mathcal{A}_{\mathcal{E}'}\otimes\mathcal{B}^\vee)$ is surjective. Viewing each summand $\mathcal{B}$ of $\mathcal{A}_{\mathcal{E}'}$, the identity global section in $H^0(\mathcal{B}\otimes\mathcal{B}^\vee)$ is in the image of this surjection. It cannot be in the image of $H^0_*(\mathcal{E}'\otimes\mathcal{B}^\vee)$ since $\mathcal{E}'$ does not have $\mathcal{B}$ as a summand. Hence it is in the image of some $\mathcal{B}'$ term in $\mathcal{A}_{\mathcal{E}}$. This forces $\mathcal{B}'$ to equal $\mathcal{B}$, and the map $\sigma_1:\mathcal{A}_{\mathcal{E}}\to\mathcal{A}_{\mathcal{E}'}$ has to split over this $\mathcal{B}$ term in $\mathcal{A}_{\mathcal{E}'}$.

It follows that σ_1 is a (split) surjection. Hence $\sigma:\mathcal{E}\to\mathcal{E}'$ is onto. The roles of $\mathcal{E},\mathcal{E}'$ can be interchanged, showing that they are bundles of the same rank. Hence $\sigma:\mathcal{E}\cong\mathcal{E}'$. □

The following theorem is in the same vein, and extends Proposition 1.3:

Theorem 1.11. *Let $\sigma:\mathcal{E}\to\mathcal{E}'$ be a sheaf homomorphism between two vector bundles on X, where $\mathcal{E}'$ has no ACM summands. Suppose that σ induces isomorphisms $H^i_*(\mathcal{E})\to H^i_*(\mathcal{E}')$ for $1\le i\le n-1$, and also, for each nonfree irreducible ACM bundle $\mathcal{B}$ appearing in $\mathcal{A}_{\mathcal{E}'}$, suppose that the induced map $H^1_*(\mathcal{E}\otimes\mathcal{B}^\vee)\to H^1_*(\mathcal{E}'\otimes\mathcal{B}^\vee)$ is an isomorphism. Then σ is a split surjection decomposing $\mathcal{E}$ into $\mathcal{E}'\oplus\mathcal{C}$, where $\mathcal{C}$ is an ACM bundle.*

Proof. By Theorem 1.5, σ can be lifted to a map $\tilde{\sigma} : \mathcal{F}_{\min} \to \mathcal{F}'_{\min}$ of minimal Horrocks data bundles. Since $H^i_*(\tilde{\sigma})$ is an isomorphism for $1 \le i \le n-1$, $\tilde{\sigma}$ is an isomorphism. So, for convenience, we may assume that $\mathcal{F}_{\min} = \mathcal{F}'_{\min}$, and, according to Theorem 1.6, σ induces a map of γ-sequences

$$\begin{tikzcd}
0 \arrow[r] & \mathcal{E} \arrow[r] \arrow[d, "\sigma"] & \mathcal{A}_{\mathcal{E}} \arrow[r] \arrow[d, "\sigma_1"] & \mathcal{G}_{\min} \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & \mathcal{E}' \arrow[r] & \mathcal{A}_{\mathcal{E}'} \arrow[r] & \mathcal{G}_{\min} \arrow[r] & 0
\end{tikzcd}$$

For each $\mathcal{B}$ appearing in $\mathcal{A}_{\mathcal{E}'}$, as in the proof of the previous theorem after tensoring by $\mathcal{B}^\vee$ we can look at the diagram of cohomology. Since $H^1_*(\mathcal{E} \otimes \mathcal{B}^\vee) \to H^1_*(\mathcal{E}' \otimes \mathcal{B}^\vee)$ is an isomorphism, the kernel in $H^0_*(\mathcal{G}_{\min} \otimes \mathcal{B}^\vee)$ is the same for $\mathcal{E}$ and $\mathcal{E}'$. The previous argument repeats to show that the homomorphism $\sigma_1 : \mathcal{A}_{\mathcal{E}} \to \mathcal{A}_{\mathcal{E}'}$ is a split surjection, with a kernel $\mathcal{C}$ which is ACM. Hence $\sigma : \mathcal{E} \to \mathcal{E}'$ is also a split surjection with kernel equal to $\mathcal{C}$. $\square$

Since the A-submodules $V_i = \ker(H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}_i^\vee) \to H^1_*(\mathcal{E} \otimes \mathcal{B}_i^\vee))$ play such an important role in the above description of a bundle $\mathcal{E}$, it is worthwhile to make the following definition describing its properties:

Definition 1.12. Let $\mathcal{F}$ be a sheaf on X and $\mathcal{B}$ an ACM bundle on X with a minimal set of generators for $H^0_*(\mathcal{B})$ given by $\bigoplus_j \mathcal{O}_X(a_j) \to \mathcal{B} \to 0$. The kernel of $H^1_*(\mathcal{F} \otimes \mathcal{B}^\vee) \to H^1_*(\mathcal{F} \otimes \bigoplus_j \mathcal{O}_X(-a_j))$ will be called the A-module of $\mathcal{B}$-socle elements for $\mathcal{F}$ and denoted by $H^1_*(\mathcal{F} \otimes \mathcal{B}^\vee)_{\mathrm{soc}}$. A homogeneous element in this kernel in degree d will be a $\mathcal{B}$-socle element in $H^1(\mathcal{F}(d) \otimes \mathcal{B}^\vee)$.

Remark 1.13. (1) For a vector bundle $\mathcal{F}$, the module of $\mathcal{B}$-socle elements for $\mathcal{F}$ has finite length over the field k.

(2) Suppose $\mathcal{B}^\vee \to \mathcal{O}_X(b)$ is any map. Then, for any sheaf $\mathcal{F}$, a $\mathcal{B}$-socle element in $H^1_*(\mathcal{F} \otimes \mathcal{B}^\vee)$ maps to zero in $H^1_*(\mathcal{F}(b))$, since $\mathcal{B}^\vee \to \mathcal{O}_X(b)$ factors through $\bigoplus_j \mathcal{O}_X(-a_j)$.

(3) Suppose $\mathcal{E}$ is a bundle on X with Horrocks datum $(\mathcal{F}_{\min}, \beta)$. Then, for any ACM bundle $\mathcal{B}$, the module $V = \ker(H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}^\vee) \to H^1_*(\mathcal{E} \otimes \mathcal{B}^\vee))$ consists of $\mathcal{B}$-socle elements for $\mathcal{F}_{\min}$. Indeed, the map $H^1_*\big(\mathcal{F}_{\min} \otimes \bigoplus_j \mathcal{O}_X(-a_j)\big) \to H^1_*\big(\mathcal{E} \otimes \bigoplus_j \mathcal{O}_X(-a_j)\big)$ is an isomorphism.

Example 1.14. Any ACM variety X with a nondegenerate embedding into $\mathbb{P}^N$ has a Horrocks data bundle given by $\Omega^1_{\mathbb{P}}|_X$ with $H^1_*(\Omega^1_{\mathbb{P}}|_X) = k$ and with an exact sequence

$$0 \longrightarrow \Omega^1_{\mathbb{P}}|_X \longrightarrow \mathcal{O}_X(-1)^{\oplus N+1} \longrightarrow \mathcal{O}_X \longrightarrow 0.$$

For any ACM bundle $\mathcal{B}$ on X, without free summands and with $\mathcal{B}^\vee \hookrightarrow \bigoplus_j \mathcal{O}_X(-a_j)$, consider the diagram

$$\begin{tikzcd} H^0_*(\mathcal{O}_X \otimes \mathcal{B}^\vee) \arrow[r] \arrow[d] & H^1_*(\Omega^1_{\mathbb{P}}|_X \otimes \mathcal{B}^\vee) \arrow[d] \\ H^0_*\big(\mathcal{O}_X \otimes \bigoplus_j \mathcal{O}_X(-a_j)\big) \arrow[r] & H^1_*\big(\Omega^1_{\mathbb{P}}|_X \otimes \bigoplus_j \mathcal{O}_X(-a_j)\big) \end{tikzcd}$$

Then any minimal generator of the module $H^0_*(\mathcal{O}_X \otimes \mathcal{B}^\vee)$ maps to a nongenerator in $H^0_*\big(\mathcal{O}_X \otimes \bigoplus_j \mathcal{O}_X(-a_j)\big)$, and hence maps to zero in $H^1_*\big(\Omega^1_{\mathbb{P}}|_X \otimes \bigoplus_j \mathcal{O}_X(-a_j)\big) = \bigoplus_j k(-a_j)$. Thus the image of $H^0_*(\mathcal{O}_X \otimes \mathcal{B}^\vee)$ in $H^1_*(\Omega^1_{\mathbb{P}}|_X \otimes \mathcal{B}^\vee)$ is nonzero and consists of $\mathcal{B}$-socle elements for $\Omega^1_{\mathbb{P}}|_X$. So, for any ACM bundle $\mathcal{B}$ on X, without free summands, the Horrocks data bundle $\Omega^1_{\mathbb{P}}|_X$ will have $\mathcal{B}$-socle elements.

For a general ACM variety X, one would expect infinitely many families of nonisomorphic and irreducible ACM bundles; hence this shows that even for a fixed Horrocks data bundle $\mathcal{F}_{\min}$, the number of bundles $\mathcal{E}$ with Horrocks datum $(\mathcal{F}_{\min}, \beta_{\mathcal{E}})$ would get out of control, especially with the construction given below. In later sections, we will limit our attention to the quadric hypersurface and the Veronese surface, where there are only finitely many ACM bundles. In these sections, we will also be able to deal with arbitrary submodules of $\mathcal{B}$-socle elements, instead of the entire $\mathcal{B}$-socle module of the rather crude theorem below.

Theorem 1.15 (existence). *Let $\mathcal{F}_{\min}$ be a minimal Horrocks data bundle on X, and let $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_k$ be a finite collection of irreducible, nonfree ACM bundles on X. Then there is a vector bundle $\mathcal{E}$ on X with Horrocks datum $(\mathcal{F}_{\min}, \beta)$ and with* $\ker H^1_*(\beta \otimes 1_{\mathcal{B}_i^\vee}) = H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}_i^\vee)_{\mathrm{soc}}$ *for $1 \le i \le k$.*

Proof. Each $H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}_i^\vee)_{\mathrm{soc}}$ is an A-module, and we can pick a collection of minimal generators for the module. Let K_i be the vector subspace spanned by this collection inside $H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}_i^\vee)_{\mathrm{soc}}$. Let $\mathcal{B} = \bigoplus (K_i \otimes_k \mathcal{B}_i)$. The data $K_i, 1 \le i \le k$, can be viewed as a $\mathcal{B}$-socle element in $H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}^\vee)$, and hence gives an extension (that defines a bundle $\mathcal{E}$)

$$0 \longrightarrow \mathcal{F}_{\min} \xrightarrow{\beta} \mathcal{E} \xrightarrow{\rho} \mathcal{B} \longrightarrow 0.$$

Since the element is a socle element, the pullback of the sequence under any map $\mathcal{O}_X(b) \to \mathcal{B}$ will split. Hence $H^0_*(\rho)$ is surjective, giving $(\mathcal{F}_{\min}, \beta)$ the Horrocks datum for $\mathcal{E}$.

By construction, the subspace $K_i \cdot I_{\mathcal{B}_i}$ in $H^0_*(\mathcal{B} \otimes \mathcal{B}_i^\vee)$ maps isomorphically to $K_i \subseteq H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}_i^\vee)_{\mathrm{soc}}$. Hence the image of the map of A-modules $H^0_*(\mathcal{B} \otimes \mathcal{B}_i^\vee) \to H^1_*(\mathcal{F}_{\min} \otimes \mathcal{B}_i^\vee)_{\mathrm{soc}}$ is onto. □

Remark 1.16. (1) The same construction can be done for an arbitrary A-submodule V_i of $H^1_*(\mathscr{F}_{\min} \otimes \mathscr{B}_i^\vee)_{\mathrm{soc}}$. We would choose K_i to be the subspace spanned by a set of minimal generators for V_i. In the last step of the above proof, we find that image of the map of A-modules $H^0_*(\mathscr{B} \otimes \mathscr{B}_i^\vee) \to H^1_*(\mathscr{F}_{\min} \otimes \mathscr{B}_i^\vee)_{\mathrm{soc}}$ contains V_i, and could possibly be larger. Hence the Horrocks invariants of $\mathscr{E}$, $\ker H^1_*(\beta \otimes 1_{\mathscr{B}_i^\vee})$, may not be precisely recognizable in this case.

(2) In the above theorem, for the $\mathscr{E}$ so constructed, it is possible to identify $A_{\mathscr{E}}$ in the case when X is arithmetically Gorenstein, or when the dual of each of the ACM bundles $\mathscr{B}_i$, $1 \le i \le k$, is also ACM: since the γ-sequence of $\mathscr{E}$ is the pushforward of the Ψ-sequence for $\mathscr{F}_{\min}$, we get the exact sequence $0 \to \mathscr{P}^0 \to \mathscr{A}_{\mathscr{E}} \to \mathscr{B} \to 0$, which is forced to split by the extra hypotheses. Once the ACM bundles in $\mathscr{A}_{\mathscr{E}}$ are identified, it is possible to compare $\mathscr{E}$ with other bundles via the uniqueness theorems (Theorems 1.10, 1.11).

(3) However, in the non-arithmetically Gorenstein case, a clear description of $\mathscr{A}_{\mathscr{E}}$ may not be apparent at the end of the construction of the theorem. We will give an example later (Example 3.3) where an identification of $\mathscr{A}_{\mathscr{E}}$ requires more work.

It is easy to obtain a splitting criterion for a vector bundle $\mathscr{E}$ on X to be free, which gives for example the criterion for quadrics in [Ancona and Ottaviani 1991] that was cited in the introduction. Once again, in the theorem below, note that the condition invoking any ACM bundle is not very useful when there are too many ACM bundles on X. It is more interesting (see the proof below) in the case where the choices for $\mathscr{B}$ are limited, for example, if one could limit the possible ACM bundles that might appear as a summand in the diagram of $\mathscr{E}$.

Theorem 1.17 (a splitting criterion). *Let $\mathscr{E}$ be a vector bundle of rank $\le r$ on X, a smooth ACM variety of dimension n, such that $H^i_*(\mathscr{E}^\vee) = 0$ for $1 \le i \le \min\{r-1, n-1\}$ and also $H^1_*(\mathscr{E}^\vee \otimes \mathscr{B}) = 0$ for any ACM bundle $\mathscr{B}$ on X. Then $\mathscr{E}$ is free.*

Proof. Now the η-sequence (Theorem 1.7) of $\mathscr{E}$, $0 \to \mathscr{K} \to \mathscr{F} \to \mathscr{E} \to 0$, gives an element in $H^1(\mathscr{E}^\vee \otimes \mathscr{K})$ which is zero by hypothesis. Hence $\mathscr{K}$ and $\mathscr{E}$ are summands of $\mathscr{F}$. Since $\mathscr{F}$ is a Horrocks data bundle, it can have no nonfree ACM summand, so $\mathscr{K}$ must be free. Thus $\mathscr{E}$ itself is a Horrocks data bundle.

If $r \ge n$, $\mathscr{E}^\vee$ is ACM. But the dual of a Horrocks data bundle has finite resolution, so $\mathscr{E}^\vee$ must be free.

If $r < n$, consider the sequence (3) with $\mathscr{E}$ replaced by $\mathscr{E}^\vee$. From the vanishing of cohomologies of $\mathscr{E}^\vee$, when we look at the complex of global sections of the sequence, we conclude that the module $E^\vee$ is an $(r+1)$-th syzygy, and $E^\vee$ has finite projective dimension since $\mathscr{E}$ is a Horrocks data bundle. By the Evans–Griffith syzygy theorem [1981], $\mathscr{E}^\vee$ is free. □

Remark 1.18. If X is a smooth quadric hypersurface, the above splitting criterion is also equivalent to Corollary 4.3 of [Ballico and Malaspina 2009]. Splitting criteria have been established on other varieties. For a Grassmannian of lines $G(1,n)$, which supports infinitely many irreducible ACM bundles when $n \geq 4$, it is possible to prove a splitting criterion (see Theorem 2.6 of [Arrondo and Malaspina 2010]) with a finite number of cohomological vanishing conditions involving only the ACM bundles $S^i Q$, where $i = 1, \dots, n-2$ and Q is the tautological rank-two bundle. Similarly, on multiprojective spaces $\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_s}$, there is a splitting criterion (see Theorem 3.9 of [Ballico and Malaspina 2011]) with a finite number of cohomological vanishing conditions involving only the ACM bundles $\mathbb{O}(k_1, \dots, k_s)$, where $-n_j \leq k_j \leq 0$. These results are much stronger than Theorem 1.17. Due to the generality of our setting, we are unable to prove a splitting criterion with conditions involving only a finite number of ACM bundles.

However, when there is additional analysis of the ACM bundles, more can be said. For example, Arrondo and Graña [1999] identified a list of six specific ACM bundles on $G(1,4)$, and showed that any other ACM bundle $\mathcal{B}$ is a summand of a bundle that appears in the middle of a short exact sequence of bundles, where the bundles on either side are built from direct sums of twists of these six bundles. Hence in our Theorem 1.17, applied to $G(1,4)$, it suffices to consider only these six specific bundles for $\mathcal{B}$. It is now straightforward to check that Ottaviani's splitting criterion on $G(1,4)$ (which is just one case of [Ottaviani 1987, Théorème 1]) follows from Theorem 1.17. (He assumed that $H^i_*(\mathscr{E}^\vee) = 0$ for $1 \leq i \leq 5$ and his other hypotheses imply that $H^1_*(\mathscr{E}^\vee \otimes \mathcal{B}) = 0$ for these six ACM bundles.)

2. Quadric hypersurfaces

Let $\mathcal{Q}_n \subset \mathbb{P}^{n+1}$ be a smooth quadric hypersurface. We will work over a field of characteristic not two. The quadratic form defining $\mathcal{Q}_n$ descends to a quadratic form on the tangent bundle of $\mathcal{Q}_n$. Hence one can define spinor bundles on $\mathcal{Q}_n$ [Karrer 1973]. Set $l := \lfloor (n+1)/2 \rfloor$. If n is even, then $\mathcal{Q}_n$ has two distinct spinor bundles Σ_1 and Σ_2 of rank 2^{l-1}. If n is odd, then $\mathcal{Q}_n$ has a unique spinor bundle, which we denote Σ_1, of rank 2^{l-1}. Algebraic properties of these bundles were studied by Ottaviani [1988], who obtained them using the geometry of the variety of all maximal linear subspaces of $\mathcal{Q}_n$ to construct morphisms from $\mathcal{Q}_n$ to $G(2^{l-1}, 2^l)$. He shows that these spinor bundles on $\mathcal{Q}_n$ are ACM bundles. Kapranov [1988] showed how these bundles were crucial in describing the derived category of sheaves on the quadric. Meanwhile, Knörrer [1987], classifying maximal Cohen–Macaulay modules over isolated quadratic hypersurface singularities, described these bundles as the fundamental ACM bundles on $\mathcal{Q}_n$ (see [Buchweitz et al. 1987] for the interpretation of Knörrer's results in terms of bundles). Knörrer's classification of ACM bundles on $\mathcal{Q}_n$ was proved also in [Ancona and Ottaviani 1991].

We use a unified notation Σ_i for spinor bundles on $\mathcal{Q}_n$, where for even n, i can take on the values $1, 2$, while if n is odd, i can be only 1. We follow the notation of [Kapranov 1988], whose spinor bundles differ from those in [Ottaviani 1988] by a twist of 1. Hence Σ_i is generated by its global sections and $\Sigma_i(-1)$ has no sections.

We will call a bundle of the form $\Sigma_i(a)$ a twisted spinor bundle on $\mathcal{Q}_n$. The fundamental theorem of [Knörrer 1987] is:

Theorem 2.1. *Any ACM bundle on $\mathcal{Q}_n$ is a direct sum of line bundles and twisted spinor bundles.*

The spinor bundles on $\mathcal{Q}_n$ satisfy some dualities [Ottaviani 1988]: when n is odd or $n \equiv 0 \pmod 4$, $\Sigma_i^\vee \cong \Sigma_i(-1)$, while if $n \equiv 2 \pmod 4$, $\Sigma_i^\vee \cong \Sigma_j(-1)$, where $j \neq i$.

In addition, the spinor bundles on $\mathcal{Q}_n$ satisfy canonical sequences. To further unify the notation, when n is odd or when $n \equiv 2 \pmod 4$, define $i \mapsto \bar{i}$ to be the identity on indices, and when $n \equiv 0 \pmod 4$, define $i \mapsto \bar{i}$ to be the transposition of the indices 1 and 2. With this notation, we have the canonical sequences

$$0 \longrightarrow \Sigma_{\bar{i}}^\vee \xrightarrow{u_i} \mathcal{O}^{\oplus 2^l} \xrightarrow{v_i} \Sigma_i \longrightarrow 0 \tag{4}$$

(see [Ottaviani 1988, Theorem 2.8]).

Ottaviani [1988, Lemma 2.7] proved that, for any spinor bundle Σ_i, $\operatorname{End}(\Sigma_i) = H^0(\Sigma_i \otimes \Sigma_i^\vee) = k$ and $\operatorname{Hom}(\Sigma_i, \Sigma_j) = 0$ for $i \neq j$. Using this, and tensoring the sequence above with $\Sigma_i^\vee$, we get $H^1(\Sigma_{\bar{i}}^\vee \otimes \Sigma_i^\vee) = k$, where Id_{Σ_i} maps to a generator of $H^1(\Sigma_{\bar{i}}^\vee \otimes \Sigma_i^\vee)$. For completeness, the following lemma is also easy to prove:

Lemma 2.2.

$$H^1_*(\Sigma_{\bar{i}}^\vee \otimes \Sigma_i^\vee) = k, \tag{5}$$

$$H^1_*(\Sigma_j^\vee \otimes \Sigma_i^\vee) = 0 \quad \textit{if } j \neq \bar{i}. \tag{6}$$

Recall the definition of socle elements.

Definition 2.3. Let $\mathcal{F}$ be a sheaf on $\mathcal{Q}_n$. The sequence dual to (4) tensored by $\mathcal{F}$ gives

$$0 \longrightarrow \mathcal{F} \otimes \Sigma_i^\vee \longrightarrow \mathcal{F} \otimes \mathcal{O}^{\oplus 2^l} \longrightarrow \mathcal{F} \otimes \Sigma_{\bar{i}} \longrightarrow 0$$

and a natural map $H^1_*(\mathcal{F} \otimes \Sigma_i^\vee) \to H^1_*(\mathcal{F} \otimes \mathcal{O}^{\oplus 2^l})$.

An element in $H^1(\mathcal{F}(d) \otimes \Sigma_i^\vee)$ will be called a Σ_i-socle element for $\mathcal{F}$ in degree d if it is annihilated by the map $H^1(\mathcal{F}(d) \otimes \Sigma_i^\vee) \to H^1_*(\mathcal{F} \otimes \mathcal{O}^{\oplus 2^l})$.

The terminology "socle" comes from the case of a quadric surface studied in [Malaspina and Rao 2014], where socle elements were annihilated by multiplication by the forms lifted from one of the $\mathbb{P}^1$ factors of $\mathcal{Q}_2$. We have extended this terminology to all ACM bundles in Section 1.

Lemma 2.4. *Let $\mathcal{F}$ be a sheaf on $\mathcal{Q}_n$. Let V be a finite-dimensional graded subspace consisting of Σ_i-socle elements in $H^1_*(\mathcal{F} \otimes \Sigma_i^\vee)$. Then there is a homomorphism $\alpha : V \otimes \Sigma_{\bar{i}}^\vee \to \mathcal{F}$ such that $H^1_*(\alpha \otimes 1_{\Sigma_i^\vee})$ has image V.*

Proof. Consider the dual canonical sequence (4) tensored by $\mathcal{F}$

$$0 \longrightarrow \mathcal{F} \otimes \Sigma_i^\vee \longrightarrow \mathcal{F} \otimes \mathcal{O}^{\oplus 2^l} \longrightarrow \mathcal{F} \otimes \Sigma_{\bar{i}} \longrightarrow 0.$$

We get

$$H^0(\mathcal{F} \otimes \Sigma_{\bar{i}}) \to H^1(\mathcal{F} \otimes \Sigma_i^\vee) \to H^1(\mathcal{F} \otimes \mathcal{O}^{\oplus 2^l}).$$

There is a graded subspace V' of $H^0_*(\mathcal{F} \otimes \Sigma_{\bar{i}})$ which is mapped isomorphically to $V \subset H^1_*(\mathcal{F} \otimes \Sigma_i^\vee)$. This induces a map $\alpha : V' \otimes_k \Sigma_{\bar{i}}^\vee \to \mathcal{F}$. Thus we can construct the commuting diagram

$$\begin{tikzcd}
0 \arrow[d] & 0 \arrow[d] \\
\mathcal{F} \otimes \Sigma_i^\vee \arrow[d, "1\otimes v_i^\vee"] & \left(V' \otimes_k \Sigma_{\bar{i}}^\vee\right) \otimes \Sigma_i^\vee \arrow[l, "\alpha\otimes 1"'] \arrow[d, "1\otimes v_i^\vee"] \\
\mathcal{F} \otimes \mathcal{O}^{\oplus 2^l} \arrow[d, "1\otimes u_i^\vee"] & \left(V' \otimes_k \Sigma_{\bar{i}}^\vee\right) \otimes \mathcal{O}^{\oplus 2^l} \arrow[l, "\alpha\otimes 1"'] \arrow[d, "1\otimes u_i^\vee"] \\
\mathcal{F} \otimes \Sigma_{\bar{i}} \arrow[d] & \left(V' \otimes_k \Sigma_{\bar{i}}^\vee\right) \otimes \Sigma_{\bar{i}} \arrow[l, "\alpha\otimes 1"'] \arrow[d] \\
0 & 0
\end{tikzcd}$$

Then $H^1_*(\alpha \otimes 1) : H^1_*((V' \otimes_k \Sigma_{\bar{i}}^\vee) \otimes \Sigma_i^\vee) \to H^1_*(\mathcal{F} \otimes \Sigma_i^\vee)$ gives $V' \cong V$. □

Corollary 2.5. *Let $\mathcal{F}$ be a vector bundle on $\mathcal{Q}_n$. Then any graded vector subspace V of Σ_i-socle elements in $H^1_*(\mathcal{F} \otimes \Sigma_i^\vee)_{\text{soc}}$ is an A-submodule of $H^1_*(\mathcal{F} \otimes \Sigma_i^\vee)_{\text{soc}}$.*

Proof. In the proof above $H^1_*(\alpha \otimes 1_{\Sigma_i^\vee})$ is an A-module homomorphism, and by Lemma 2.2 the A-module $H^1_*((V' \otimes_k \Sigma_{\bar{i}}^\vee) \otimes \Sigma_i^\vee)$ has the trivial A-module structure, where multiplication by graded elements in A of positive degree is zero. □

For any vector bundle $\mathcal{E}$ on $\mathcal{Q}_n$, we will define invariants as follows:

Definition 2.6 (Horrocks invariants of $\mathscr{E}$). Let $\mathscr{E}$ be a vector bundle on $\mathscr{Q}_n$. It has a minimal associated Horrocks datum $(\mathscr{F}_{\min}, \beta)$. Let

$$V_i = \ker H^1(\beta \otimes \mathrm{Id}_{\Sigma_i^\vee}) : H^1_*(\mathscr{F}_{\min} \otimes \Sigma_i^\vee) \to H^1_*(\mathscr{E} \otimes \Sigma_i^\vee).$$

Then V_i is a graded subspace of $H^1_*(\mathscr{F}_{\min} \otimes \Sigma_i^\vee)_{\mathrm{soc}}$. The collection $(\mathscr{F}_{\min}, V_i)$ will be called Horrocks invariants for $\mathscr{E}$. (As usual, when n is even, this means $(\mathscr{F}_{\min}, V_1, V_2)$ and when n is odd, it means $(\mathscr{F}_{\min}, V_1)$.)

Remark 2.7. (1) $\mathscr{E}$ is ACM if and only if $\mathscr{F}_{\min}$ is the zero bundle. $V_i = 0$ as well.

(2) In general, $V_i = 0$ for all i if and only if $\mathscr{E}$ is a direct sum of a Horrocks data bundle and an ACM bundle.

(3) If $\mathscr{B}$ is an ACM bundle, then $\mathscr{E}$ and $\mathscr{E} \oplus \mathscr{B}$ will have the same Horrocks invariants.

(4) If $(\mathscr{F}_{\min}, \beta, V_i)$ is a collection of Horrocks invariants for $\mathscr{E}$ and ϕ is an automorphism of $\mathscr{F}_{\min}$, then ϕ can be used to change $\beta : \mathscr{F}_{\min} \to \mathscr{E}$ and hence also V_i to get a new collection of Horrocks invariants for $\mathscr{E}$.

(5) The definition could have used an arbitrary Horrocks data bundle $\mathscr{F}$ for $\mathscr{E}$ instead of the minimal one $\mathscr{F}_{\min}$, since $H^1_*(\Sigma_i^\vee) = 0$ and hence the description of V_i would not change.

A stronger existence theorem for quadrics can now be stated than was proved in Theorem 1.15. Below we have a statement that deals with arbitrary subspaces of socle elements:

Theorem 2.8 (existence). *Let $\mathscr{F}_{\min}$ be a minimal Horrocks data bundle on $\mathscr{Q}_n$ and let V_i be a graded vector subspace of $H^1_*(\mathscr{F}_{\min} \otimes \Sigma_i^\vee)_{\mathrm{soc}}$. Then there exists a vector bundle $\mathscr{E}$ with the Horrocks invariants $(\mathscr{F}_{\min}, V_1, V_2)$ (when n is even) and invariants $(\mathscr{F}_{\min}, V_1)$ (when n is odd).*

Proof. We follow the approach in Theorem 1.15. For notational convenience, assume n is even, so $i = 1, 2$. Let $\mathscr{B} = (V_1 \otimes_k \Sigma_1) \oplus (V_2 \otimes_k \Sigma_2)$. As in the earlier proof, we obtain a short exact sequence (defining $\mathscr{E}$):

$$0 \longrightarrow \mathscr{F}_{\min} \xrightarrow{\beta} \mathscr{E} \xrightarrow{\rho} (V_1 \otimes_k \Sigma_1) \oplus (V_2 \otimes_k \Sigma_2) \longrightarrow 0,$$

where $(\mathscr{F}_{\min}, \beta)$ is a Horrocks datum for the bundle $\mathscr{E}$ so obtained. Our goal is now to show that the image of $H^0_*(\mathscr{B} \otimes \Sigma_i^\vee) \to H^1_*(\mathscr{F}_{\min} \otimes \Sigma_i^\vee)$ is V_i, whereas in the earlier proof we showed that it contained V_i. Let $\Sigma_j(a)$ be any summand in $\mathscr{B}$, and pick a nonzero section $s \in H^0(\Sigma_j(a) \otimes \Sigma_i^\vee(b))$, or a map $s : \Sigma_i(-b) \to \Sigma_j(a)$. Then $a + b \geq 0$. The section $s \in H^0(\mathscr{B} \otimes \Sigma_i^\vee(b))$ maps to zero in $H^1_*(\mathscr{F}_{\min} \otimes \Sigma_i^\vee)$ if and only if the pullback of the short exact sequence by the map $s : \Sigma_i(-b) \to \mathscr{B}$ is a split sequence. If $a + b > 0$, by Lemma 2.2 the map $s : \Sigma_i(-b) \to \Sigma_j(a)$

factors through $\mathcal{O}^{\oplus 2^l}(a)$. The pullback of the short exact sequence by the map $\mathcal{O}^{\oplus 2^l}(a) \to \Sigma_j(a) \subseteq \mathcal{B}$ splits since the extension is defined by socle elements. Hence so does the pullback by the map $\Sigma_i(-b) \to \Sigma_j(a) \subseteq \mathcal{B}$.

It follows that the only nonzero contribution from this summand $\Sigma_j(a)$ to the image of $H^0(\mathcal{B} \otimes \Sigma_i^\vee(b))$ occurs when $a+b=0$. If $i \neq j$, $\mathrm{Hom}(\Sigma_i, \Sigma_j) = 0$ and so no section s can be found. If $i = j$, $\mathrm{End}(\Sigma_i) = k$ and it follows that the image of s lies in V_i. Thus the image of $H^0_*(\mathcal{B} \otimes \Sigma_i^\vee)$ is exactly V_i. □

As pointed out after Theorem 1.15, if $\mathcal{F}_{\min}$ has a Ψ-sequence $0 \to \mathcal{F}_{\min} \to \mathcal{P}^0 \to \mathcal{G}_{\min} \to 0$, then the $\mathcal{E}$ constructed in the above theorem has γ-sequence given as

$$0 \longrightarrow \mathcal{E} \longrightarrow \bigoplus_i (V_i \otimes_k \Sigma_i) \oplus \mathcal{P}^0 \longrightarrow \mathcal{G}_{\min} \longrightarrow 0.$$

It is also easy to see that since $\mathcal{F}_{\min}$ has no summands of type Σ_i, neither does $\mathcal{E}$. Conversely, suppose $\mathcal{E}$ is a vector bundle on $\mathcal{Q}_n$ with Horrocks invariants $(\mathcal{F}_{\min}, V_i)$ and with no summands of type Σ_i. It will follow from the next theorems that $\mathcal{E}$ has a γ-sequence with $\mathcal{A}_{\mathcal{E}} = \bigoplus_i (V_i \otimes_k \Sigma_i) \oplus \mathcal{P}'$, where $\mathcal{P}'$ is free.

The following two uniqueness results follow easily from the general theorems of Section 1.

Theorem 2.9 (uniqueness). *Given two bundles $\mathcal{E}, \mathcal{E}'$ on $\mathcal{Q}_n$ without ACM summands and with Horrocks invariants $(\mathcal{F}_{\min}, V_i)$, $(\mathcal{F}'_{\min}, V'_i)$, suppose that there exists $\phi : \mathcal{F}_{\min} \xrightarrow{\sim} \mathcal{F}'_{\min}$ such that the induced isomorphisms $H^1_*(\mathcal{F}_{\min} \otimes \Sigma_i^\vee) \cong H^1_*(\mathcal{F}'_{\min} \otimes \Sigma_i^\vee)$ carry V_i to V'_i for each i. Then $\mathcal{E}$ and $\mathcal{E}'$ are isomorphic.*

Proof. We may assume that $\mathcal{E}$ and $\mathcal{E}'$ have the same minimal Horrocks data bundle $\mathcal{F}_{\min}$. If $\mathcal{F}_{\min}$ is zero, $\mathcal{E}, \mathcal{E}'$ are ACM and the theorem does not apply. So we will assume that $\mathcal{F}_{\min}$ is a nonfree minimal Horrocks data bundle. If V_i is 0 for $i = 1, 2$, then $\mathcal{E}$ is stably equivalent to $\mathcal{F}_{\min}$, and, being without ACM summands, it must be isomorphic to $\mathcal{F}_{\min}$. Since V'_i will also be zero, the same is true for $\mathcal{E}'$ and we conclude that $\mathcal{E} \cong \mathcal{E}'$. So assume V_i is nonzero for some i. If there is an automorphism ϕ of $\mathcal{F}_{\min}$ which carries V_i to V'_i, in the diagram of Theorem 1.8 for $\mathcal{E}'$, we may replace $\beta' : \mathcal{F}_{\min} \to \mathcal{E}'$ by $\beta' \circ \phi^{-1}$ and so on, and assume that β and β' give the same kernel V_i in $H^1_*(F_{\min} \otimes \Sigma_i^\vee)$.

We can now apply Theorem 1.10 to conclude the result. □

Theorem 2.10. *Let $\mathcal{E}, \mathcal{E}'$ be vector bundles on $\mathcal{Q}_n$ with no ACM summands. Suppose $\sigma : \mathcal{E} \to \mathcal{E}'$ is a homomorphism such that σ induces $H^j_*(\mathcal{E}) \cong H^j_*(\mathcal{E}')$ for $1 \leq j \leq n-1$ and also isomorphisms $H^1_*(\mathcal{E} \otimes \Sigma_i^\vee) \cong H^1_*(\mathcal{E}' \otimes \Sigma_i^\vee)$ for all i. Then σ is an isomorphism.*

Proof. This is just Theorem 1.11 with the additional condition that $\mathcal{E}$ has no ACM summands. □

3. The Veronese surface

The Veronese surface $\mathcal{V} \subset \mathbb{P}^5$ is an arithmetically Cohen–Macaulay embedding which is not arithmetically Gorenstein. The study of vector bundles on $\mathcal{V}$ is trivial if we view $\mathcal{V}$ as $\mathbb{P}^2$. Below we discuss how the techniques of Section 1 apply to the embedded variety $\mathcal{V}$. With its polarization from the embedding, $\mathcal{V}$ has two irreducible, nonfree ACM bundles (up to twists). Hence, as in the case of quadric hypersurfaces of even dimension, we can define Horrocks invariants $(\mathcal{F}_{\min}, V, W)$ for any vector bundle $\mathcal{E}$ on $\mathcal{V}$. But unlike in the case of the quadric, where V, W were independent of each other, here there is a dependency between them.

In the following discussion, we will write $\mathbb{O}_{\mathcal{V}}(1)$ for $\mathbb{O}_{\mathbb{P}^5}(1)|_{\mathcal{V}}$ and $\mathbb{O}_{\mathcal{V}}(n)$ for $\mathbb{O}_{\mathcal{V}}(1)^{\otimes n}$. We will write $\mathcal{L}$ for $\mathbb{O}_{\mathbb{P}^2}(1)$ and $\mathcal{U}$ for $\Omega^1_{\mathcal{V}} \otimes \mathcal{L}$. Then the only irreducible ACM bundles on $\mathcal{V}$ (with respect to the polarization $\mathbb{O}_{\mathcal{V}}(1)$) are $\mathbb{O}_{\mathcal{V}}(n)$, $\mathcal{L}(n)$ and $\mathcal{U}(n)$. In the diagram of a bundle $\mathcal{E}$ on $\mathcal{V}$ in Theorem 1.8, the terms $\mathcal{A}_{\mathcal{E}}$ and $\mathcal{K}_{\mathcal{E}}$ are built out of these three types of irreducible ACM bundles. The vector bundle $\mathcal{G}$ is a free bundle and the Ψ-sequence is the sheafification of a free presentation of the A-module $H^1_*(\mathcal{E})$. The connection between $\mathcal{A}_{\mathcal{E}}$ and $\mathcal{K}_{\mathcal{E}}$, given by the Δ-sequence in the diagram of $\mathcal{E}$, is controlled by the canonical sequences

$$0 \longrightarrow \mathcal{U} \xrightarrow{u} 3\mathbb{O}_{\mathcal{V}} \xrightarrow{v} \mathcal{L} \longrightarrow 0 \tag{7}$$

and

$$0 \longrightarrow 3\mathcal{U}(-1) \oplus \mathbb{O}_{\mathcal{V}}(-1) \longrightarrow 9\mathbb{O}_{\mathcal{V}}(-1) \longrightarrow \mathcal{U} \longrightarrow 0, \tag{8}$$

where the second can be simplified noncanonically to

$$0 \longrightarrow 3\mathcal{U}(-1) \xrightarrow{u'} 8\mathbb{O}_{\mathcal{V}}(-1) \xrightarrow{v'} \mathcal{U} \longrightarrow 0. \tag{9}$$

In addition, there is the canonical sequence

$$0 \longrightarrow \mathbb{O}_{\mathcal{V}}(-1) \longrightarrow 3\mathcal{L}(-1) \longrightarrow \mathcal{U} \longrightarrow 0. \tag{10}$$

The two uniqueness theorems of Section 1 apply in this setting, where given a bundle $\mathcal{E}$ on $\mathcal{V}$ we can construct Horrocks invariants for $\mathcal{E}$ as $(\mathcal{F}_{\min}, V, W)$, where $(\mathcal{F}_{\min}, \beta)$ is a Horrocks datum for $\mathcal{E}$, $V = \ker(H^1_*(\mathcal{F}_{\min} \otimes \mathcal{L}^\vee) \to H^1_*(\mathcal{E} \otimes \mathcal{L}^\vee))$ and $W = \ker(H^1_*(\mathcal{F}_{\min} \otimes \mathcal{U}^\vee) \to H^1_*(\mathcal{E} \otimes \mathcal{U}^\vee))$. Thus to complete the classification of bundles on $\mathcal{V}$ by this method it remains to get a description of any constraints on $V \subseteq H^1_*(\mathcal{F} \otimes \mathcal{L}^\vee)$ and $W \subseteq H^1_*(\mathcal{F}_{\min} \otimes \mathcal{U}^\vee)$, and to finally show that given $(\mathcal{F}_{\min}, V, W)$ with these constraints, there exists a bundle $\mathcal{E}$ with those invariants.

By Remark 1.13, V is an A-submodule of $\mathcal{L}$-socle elements in $H^1_*(\mathcal{F}_{\min} \otimes \mathcal{L}^\vee)_{\mathrm{soc}}$ and W is an A-submodule of $\mathcal{U}$-socle elements in $H^1_*(\mathcal{F}_{\min} \otimes \mathcal{U}^\vee)_{\mathrm{soc}}$. By the next lemma, there is no distinction between the concepts of graded A-submodules and graded vector subspaces of socle elements:

Lemma 3.1. *For any vector bundle $\mathcal{F}$ on $\mathcal{V}$, in the A-module structures of both $H^1_*(\mathcal{F}\otimes\mathcal{L}^\vee)_{\text{soc}}$ and $H^1_*(\mathcal{F}\otimes\mathcal{U}^\vee)_{\text{soc}}$, multiplication by graded elements of positive degree in A is zero.*

Proof. Let $\eta\in H^1(\mathcal{F}(d)\otimes\mathcal{L}^\vee)_{\text{soc}}$, giving a short exact sequence $0\to\mathcal{F}(d)\to\mathcal{A}\to\mathcal{L}\to 0$. Consider multiplication by $x\in A$ of degree one, $\cdot x:\mathcal{L}(-1)\to\mathcal{L}$. The pullback by this map of the short exact sequence (7) is split since $H^1(\mathcal{U}\otimes\mathcal{L}^\vee(1))=0$. So $\cdot x:\mathcal{L}(-1)\to\mathcal{L}$ factors through $3\mathcal{O}_{\mathcal{V}}$. By the definition of $\mathcal{L}$-socle elements, the pullback of η by $3\mathcal{O}_{\mathcal{V}}\to\mathcal{L}$ splits, hence also the pullback of η by $\cdot x:\mathcal{L}(-1)\to\mathcal{L}$. Thus $x\cdot\eta=0$.

A similar proof works for an element $\eta\in H^1(\mathcal{F}(d)\otimes\mathcal{U}^\vee)_{\text{soc}}$. One notices that the pullback by $\cdot x:\mathcal{U}(-1)\to\mathcal{U}$ of the short exact sequence (9) is split because $H^1_*(\mathcal{U}\otimes\mathcal{U}^\vee)=3k$ supported in $H^1(\mathcal{U}\otimes\mathcal{U}^\vee(-1))$. $\square$

In the definition of $\mathcal{U}$-socle elements for $\mathcal{F}$, the noncanonical inclusion $\mathcal{U}^\vee\hookrightarrow 8\mathcal{O}_{\mathcal{V}}(1)$ can be replaced by a canonical composite inclusion $\mathcal{U}^\vee\hookrightarrow 3\mathcal{L}^\vee(1)\hookrightarrow 9\mathcal{O}_{\mathcal{V}}(1)$. For any bundle $\mathcal{F}$, this gives a canonical map

$$\phi_{\mathcal{F}}:H^1_*(\mathcal{F}\otimes\mathcal{U}^\vee)_{\text{soc}}\to 3H^1_*(\mathcal{F}(1)\otimes\mathcal{L}^\vee)_{\text{soc}}.$$

When $\mathcal{E}$ is a vector bundle with Horrocks invariants $(\mathcal{F}_{\min},V,W)$, it is immediate to see that V and W are related by $\phi_{\mathcal{F}_{\min}}(W)\subseteq 3V(1)$. This is a dependency between V and W. In fact, this is the only requirement on the pair (V,W) for proving an existence theorem on the Veronese surface:

Theorem 3.2. *Let $\mathcal{F}_{\min}$ be a minimal Horrocks data bundle on $\mathcal{V}$, and let V,W be graded vector subspaces of $H^1_*(\mathcal{F}_{\min}\otimes\mathcal{L}^\vee)_{\text{soc}}$, $H^1_*(\mathcal{F}_{\min}\otimes\mathcal{U}^\vee)_{\text{soc}}$ with the property that $\phi_{\mathcal{F}_{\min}}(W)\subseteq 3V(1)$. Then there is a vector bundle $\mathcal{E}$ on $\mathcal{V}$ with Horrocks invariants $(\mathcal{F}_{\min},V,W)$.*

Proof. Construct $\mathcal{E}$ as an extension of $\mathcal{F}_{\min}$ by $\mathcal{B}=(V\otimes_k\mathcal{L})\oplus(W\otimes_k\mathcal{U})$:

$$0\longrightarrow\mathcal{F}_{\min}\xrightarrow{\beta}\mathcal{E}\longrightarrow\mathcal{B}\longrightarrow 0. \tag{$*$}$$

Since V, W are subspaces of socle elements, $\mathcal{E}$ has $(\mathcal{F}_{\min},\beta)$ as its Horrocks datum. We wish to understand the images of $H^0_*(\mathcal{B}\otimes\mathcal{L}^\vee)\to H^1_*(\mathcal{F}_{\min}\otimes\mathcal{L}^\vee)$ and $H^0_*(\mathcal{B}\otimes\mathcal{U}^\vee)\to H^1_*(\mathcal{F}_{\min}\otimes\mathcal{U}^\vee)$. $\operatorname{End}(\mathcal{L})=\operatorname{End}(\mathcal{U})=k$ and the image of $V\cdot I_{\mathcal{L}}\subseteq H^0(V\otimes\mathcal{L}\otimes\mathcal{L}^\vee)$ and $W\cdot I_{\mathcal{U}}\subseteq H^0(W\otimes\mathcal{U}\otimes\mathcal{U}^\vee)$ give V and W in $H^1_*(\mathcal{F}_{\min}\otimes\mathcal{L}^\vee)_{\text{soc}}$ and $H^1_*(\mathcal{F}_{\min}\otimes\mathcal{U}^\vee)_{\text{soc}}$. It remains to analyze any other contributions to the two images inside $H^1_*(\mathcal{F}_{\min}\otimes\mathcal{L}^\vee)_{\text{soc}}$ and $H^1_*(\mathcal{F}_{\min}\otimes\mathcal{U}^\vee)_{\text{soc}}$, and prove that the images are just V and W respectively.

Let $\mathcal{L}(b),\mathcal{U}(b)$ be any summands in $(V\otimes_k\mathcal{L})\oplus(W\otimes_k\mathcal{U})$. Consider maps $\sigma_1:\mathcal{L}(a)\to\mathcal{L}(b)$, $\sigma_2:\mathcal{L}(a)\to\mathcal{U}(b)$, $\sigma_3:\mathcal{U}(a)\to\mathcal{U}(b)$, $\sigma_4:\mathcal{U}(a)\to\mathcal{L}(b)$. For σ_1, assume $a<b$ since we wish to omit endomorphisms of $\mathcal{L}$. Likewise for σ_3. In

the sequence (7) tensored by $\mathscr{L}^\vee(b-a)$ we have $H^1(\mathscr{U}\otimes\mathscr{L}^\vee(b-a))=0$, and in the sequence (9) tensored by $\mathscr{U}^\vee(b-a)$ we have $H^1(3\mathscr{U}(-1)\otimes\mathscr{U}^\vee(b-a))=0$. Hence σ_1 factors through $3\mathscr{O}_{\mathscr{V}}(b)$ and σ_3 factors through $8\mathscr{O}_{\mathscr{V}}(b-1)$. By the socle nature of the extension $(*)$, pullbacks of $(*)$ by σ_1 and σ_3 split; hence the element $\sigma_1\in H^0(\mathscr{L}(b)\otimes\mathscr{L}^\vee(-a))$ maps to zero in $H^1_*(\mathscr{F}_{\min}\otimes\mathscr{L}^\vee)$, and likewise σ_3 maps to zero in $H^1_*(\mathscr{F}_{\min}\otimes\mathscr{U}^\vee)$.

For σ_4 to be nonzero, we require that $a<b+1$. We know $H^1(\mathscr{U}\otimes\mathscr{U}^\vee(b-a))=0$. Hence the same argument applies to show that σ_4 factors through $3\mathscr{O}_{\mathscr{V}}(b)$, and we are done. The arguments for σ_3,σ_4 show that the image of $H^0_*(\mathscr{B}\otimes\mathscr{U}^\vee)\to H^1_*(\mathscr{F}_{\min}\otimes\mathscr{U}^\vee)$ equals W.

For σ_2 to be nonzero we require that $a<b$, and we know that

$$H^1(3\mathscr{U}(-1)\otimes\mathscr{L}^\vee(b-a))=0$$

except when $b-a=1$. Hence the only situation of difficulty is when we have $\sigma_2:\mathscr{L}(b-1)\to\mathscr{U}(b)$. Suppose the pullback of our short exact sequence $(*)$ by $\mathscr{L}(b-1)\xrightarrow{\sigma_2}\mathscr{U}(b)\hookrightarrow\mathscr{B}$ is nonsplit. The pullback of $(*)$ by $\mathscr{U}(b)\hookrightarrow\mathscr{B}$ gives a nonzero element w of degree $-b$ in $W\subseteq H^1_*(\mathscr{F}_{\min}\otimes\mathscr{U}^\vee)_{\mathrm{soc}}$. The nonsplit pullback by $\mathscr{L}(b-1)\to\mathscr{B}$ gives a nonzero element v in $H^1(\mathscr{F}_{\min}\otimes\mathscr{L}^\vee(-b+1))_{\mathrm{soc}}$ which is the image of w under $\sigma_2^\vee$. Since $\sigma_2^\vee$ is one component in $\mathscr{U}^\vee(-b)\hookrightarrow 3\mathscr{L}^\vee(-b+1)$, the assumption that $\phi_{\mathscr{F}_{\min}}(W)\subseteq 3V(1)$ tells us that $v\in V$. Thus, the image of $H^0_*(\mathscr{B}\otimes\mathscr{L}^\vee)\to H^1_*(\mathscr{F}_{\min}\otimes\mathscr{L}^\vee)$ equals V. $\square$

We conclude with an example:

Example 3.3. The simplest non-ACM bundle on $\mathscr{V}$ is $\mathscr{E}=\Omega^1_{\mathscr{V}}=\mathscr{U}\otimes\mathscr{L}^\vee$, with $H^1_*(\mathscr{E})=k$ and γ-sequence $0\to\mathscr{E}\to 3\mathscr{L}^\vee\to\mathscr{O}_{\mathscr{V}}\to 0$, while its minimal Horrocks data bundle is $\mathscr{F}=\mathscr{F}_{\min}=\Omega^1_{\mathbb{P}^5}|_{\mathscr{V}}$, with Ψ sequence $0\to\mathscr{F}\to 6\mathscr{O}_{\mathscr{V}}(-1)\to\mathscr{O}_{\mathscr{V}}\to 0$. The map $\beta:\mathscr{F}\to\mathscr{E}$ is the standard map $\Omega^1_{\mathbb{P}^5}|_{\mathscr{V}}\to\Omega^1_{\mathscr{V}}$, which is a surjective map of vector bundles but not surjective on the module of global sections. The Horrocks invariants $(\mathscr{F},V,W)$ of $\mathscr{E}$ are easy to work out and are described below.

$H^1_*(\mathscr{F}\otimes\mathscr{L}^\vee)=H^1(\mathscr{F}(1)\otimes\mathscr{L}^\vee)=3k$, and $H^1_*(\mathscr{E}\otimes\mathscr{L}^\vee)=0$, hence $V=3k=H^1(\mathscr{F}(1)\otimes\mathscr{L}^\vee)$, where all elements in $H^1_*(\mathscr{F}\otimes\mathscr{L}^\vee)$ are $\mathscr{L}$-socle.

There is a commutative diagram that shows the only nonzero parts of $H^1_*(\mathscr{F}\otimes\mathscr{U}^\vee)$ and $H^1_*(\mathscr{E}\otimes\mathscr{U}^\vee)$

$$\begin{tikzcd}
H^0(\mathscr{U}^\vee) \arrow[r, hook] \arrow[d, equal] & H^1(\mathscr{F}\otimes\mathscr{U}^\vee) \arrow[r] \arrow[d, "\beta\otimes I_{\mathscr{U}^\vee}"] & H^1(6\mathscr{U}^\vee(-1)) \arrow[r] \arrow[d] & 0 \\
H^0(\mathscr{U}^\vee) \arrow[r, phantom, "\cong"] & H^1(\mathscr{E}\otimes\mathscr{U}^\vee) \arrow[r] & 0 &
\end{tikzcd}$$

Hence $H^1_*(\mathscr{F}\otimes\mathscr{U}^\vee)=H^1(\mathscr{F}\otimes\mathscr{U}^\vee)$ is nine-dimensional, and the kernel W of $H^1_*(\beta\otimes I_{\mathscr{U}^\vee})$ is a six-dimensional subspace (of $\mathscr{U}$-socle elements) that maps isomorphically to $H^1(6\mathscr{U}^\vee(-1))$.

When we apply the construction of the existence theorems (Theorems 1.15, 3.2) to the data $(\mathcal{F}, V, W)$, we obtain a vector bundle $\widetilde{\mathscr{E}}$ and a pushout diagram (refer to the discussion after Theorem 1.15)

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & & & \\
 & & \downarrow & & \downarrow & & & & \\
0 & \longrightarrow & \mathcal{F} & \longrightarrow & 6\mathcal{O}_{\mathcal{V}}(-1) & \longrightarrow & \mathcal{O}_{\mathcal{V}} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle\tilde{\beta}} & & \downarrow & & \| & & \\
0 & \longrightarrow & \widetilde{\mathscr{E}} & \longrightarrow & \mathscr{A}_{\widetilde{\mathscr{E}}} & \longrightarrow & \mathcal{O}_{\mathcal{V}} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & & & \\
 & & \mathscr{B} & \cong & \mathscr{B} & & & & \\
 & & \downarrow & & \downarrow & & & & \\
 & & 0 & & 0 & & & &
\end{array}$$

where $\mathscr{B} = (V \otimes_k \mathscr{L}) \oplus (W \otimes_k \mathscr{U})$.

According to the uniqueness theorems, $\mathscr{E}$ is a rank-two summand of the rank-20 bundle $\widetilde{\mathscr{E}}$, with the remaining summand of $\widetilde{\mathscr{E}}$ consisting of ACM bundles. In this example, even $\mathscr{A}_{\widetilde{\mathscr{E}}}$ is not obvious because the middle short exact sequence is not split. Indeed, the middle sequence is the pushout of the left sequence, hence it is split if and only if, under $\mathcal{F} \to 6\mathcal{O}_{\mathcal{V}}(-1)$, the image of the element $\tau \in H^1(\mathcal{F} \otimes \mathscr{B}^\vee)$ is zero in $H^1(6\mathcal{O}_{\mathcal{V}}(-1) \otimes \mathscr{B}^\vee)$. However, the components of τ in each of the $\mathscr{U}$-summands of $\mathscr{B}$ generate the vector space $W \subset H^1(\mathcal{F} \otimes \mathscr{U}^\vee)$, and W maps isomorphically to $H^1(6\mathscr{U}^\vee(-1))$. Hence the image of τ is nonzero.

To understand $\widetilde{\mathscr{E}}$ and $\mathscr{A}_{\widetilde{\mathscr{E}}}$, a little more work is needed. The fact that W maps isomorphically to $H^1(6\mathscr{U}^\vee(-1))$ tells us that the middle short exact sequence contains six copies of the canonical sequence (10). Hence $\mathscr{A}_{\widetilde{\mathscr{E}}} = 21\mathscr{L}^\vee$. The map $\mathscr{A}_{\widetilde{\mathscr{E}}} \to \mathcal{O}_{\mathcal{V}}$ is now easy to understand and shows that $\widetilde{\mathscr{E}} = \mathscr{E} \oplus 18\mathscr{L}^\vee$.

References

[Ancona and Ottaviani 1991] V. Ancona and G. Ottaviani, "Some applications of Beilinson's theorem to projective spaces and quadrics", *Forum Math.* **3**:2 (1991), 157–176. MR 92e:14039 Zbl 0725.14009

[Arrondo and Graña 1999] E. Arrondo and B. Graña, "Vector bundles on $G(1, 4)$ without intermediate cohomology", *J. Algebra* **214**:1 (1999), 128–142. MR 2000e:14069 Zbl 0963.14027

[Arrondo and Malaspina 2010] E. Arrondo and F. Malaspina, "Cohomological characterization of vector bundles on Grassmannians of lines", *J. Algebra* **323**:4 (2010), 1098–1106. MR 2010m:14054 Zbl 1200.14081

[Auslander and Bridger 1969] M. Auslander and M. Bridger, *Stable module theory*, Memoirs of the American Mathematical Society **94**, American Mathematical Society, Providence, R.I., 1969. MR 42 #4580 Zbl 0204.36402

[Ballico and Malaspina 2009] E. Ballico and F. Malaspina, "Qregularity and an extension of the Evans–Griffiths criterion to vector bundles on quadrics", *J. Pure Appl. Algebra* **213**:2 (2009), 194–202. MR 2009j:14055 Zbl 1153.14014

[Ballico and Malaspina 2011] E. Ballico and F. Malaspina, "Regularity and cohomological splitting conditions for vector bundles on multiprojective spaces", *J. Algebra* **345** (2011), 137–149. MR 2842058 Zbl 1246.14056

[Beĭlinson 1978] A. A. Beĭlinson, "Coherent sheaves on $\mathbb{P}^n$ and problems in linear algebra", *Funktsional. Anal. i Prilozhen.* **12**:3 (1978), 68–69. In Russian; translated in *Funct. Anal. Appl.* **12**:3 (1978), 214–216. MR 80c:14010b

[Buchweitz 1986] R.-O. Buchweitz, "Maximal Cohen–Macaulay modules and Tate-Cohomology over Gorenstein rings", preprint, 1986, http://hdl.handle.net/1807/16682.

[Buchweitz et al. 1987] R.-O. Buchweitz, G.-M. Greuel, and F.-O. Schreyer, "Cohen–Macaulay modules on hypersurface singularities, II", *Invent. Math.* **88**:1 (1987), 165–182. MR 88d:14005 Zbl 0617.14034

[Coandă 2010] I. Coandă, "The Horrocks correspondence for coherent sheaves on projective spaces", *Homology, Homotopy Appl.* **12**:1 (2010), 327–353. MR 2012a:14036 Zbl 1197.14012

[Evans and Griffith 1981] E. G. Evans and P. Griffith, "The syzygy problem", *Ann. of Math.* (2) **114**:2 (1981), 323–333. MR 83i:13006 Zbl 0497.13013

[Horrocks 1964] G. Horrocks, "Vector bundles on the punctured spectrum of a local ring", *Proc. London Math. Soc.* (3) **14** (1964), 689–713. MR 30 #120 Zbl 0126.16801

[Kapranov 1988] M. M. Kapranov, "On the derived categories of coherent sheaves on some homogeneous spaces", *Invent. Math.* **92**:3 (1988), 479–508. MR 89g:18018 Zbl 0651.18008

[Karrer 1973] G. Karrer, "Darstellung von Cliffordbündeln", *Ann. Acad. Sci. Fenn. Ser. A I* 521 (1973), 34. MR 47 #7641 Zbl 0262.53025

[Knörrer 1987] H. Knörrer, "Cohen–Macaulay modules on hypersurface singularities, I", *Invent. Math.* **88**:1 (1987), 153–164. MR 88d:14004 Zbl 0617.14033

[Malaspina and Rao 2014] F. Malaspina and A. P. Rao, "Horrocks correspondence on a quadric surface", *Geom. Dedicata* **169** (2014), 15–31. MR 3175233 Zbl 06309484

[Ottaviani 1987] G. Ottaviani, "Critères de scindage pour les fibrés vectoriels sur les Grassmanniennes et les quadriques", *C. R. Acad. Sci. Paris Sér. I Math.* **305**:6 (1987), 257–260. MR 88j:14024 Zbl 0629.14012

[Ottaviani 1988] G. Ottaviani, "Spinor bundles on quadrics", *Trans. Amer. Math. Soc.* **307**:1 (1988), 301–316. MR 89h:14012 Zbl 0657.14006

[Walter 1996] C. H. Walter, "Pfaffian subschemes", *J. Algebraic Geom.* **5**:4 (1996), 671–704. MR 99f:14064 Zbl 0864.14032

Communicated by David Eisenbud
Received 2014-10-06 Revised 2015-02-23 Accepted 2015-04-07

francesco.malaspina@polito.it — *Dipartimento di Scienze Matematiche, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy*

raoa@umsl.edu — *Department of Mathematics, University of Missouri – St. Louis, Saint Louis, MO 63121, United States*

ALGEBRA AND NUMBER THEORY 9:4 (2015)
dx.doi.org/10.2140/ant.2015.9.1005

The Elliott–Halberstam conjecture implies the Vinogradov least quadratic nonresidue conjecture

Terence Tao

For each prime p, let $n(p)$ denote the least quadratic nonresidue modulo p. Vinogradov conjectured that $n(p) = O(p^\varepsilon)$ for every fixed $\varepsilon > 0$. This conjecture follows from the generalized Riemann hypothesis and is known to hold for almost all primes p but remains open in general. In this paper, we show that Vinogradov's conjecture also follows from the Elliott–Halberstam conjecture on the distribution of primes in arithmetic progressions, thus providing a potential "nonmultiplicative" route to the Vinogradov conjecture. We also give a variant of this argument that obtains bounds on short centered character sums from "Type II" estimates of the type introduced recently by Zhang and improved upon by the Polymath project or from bounds on the level of distribution on variants of the higher-order divisor function. In particular, an improvement over the Burgess bound would be obtained if one had Type II estimates with level of distribution above $\frac{2}{3}$ (when the conductor is not cube-free) or $\frac{3}{4}$ (if the conductor is cube-free); morally, one would also obtain such a gain if one had distributional estimates on the third or fourth divisor functions τ_3 or τ_4 at level above $\frac{2}{3}$ or $\frac{3}{4}$, respectively. Some applications to the least primitive root are also given.

1. Introduction

For each prime p, let $n(p)$ denote the least natural number that is not a quadratic residue modulo p. Vinogradov [1985] established the asymptotic bound

$$n(p) \ll p^{1/2\sqrt{e}} \log^2 p \tag{1-1}$$

for all primes p and made the following conjecture:

Conjecture 1.1 (Vinogradov's conjecture). *For any fixed $\varepsilon > 0$, we have $n(p) \ll p^\varepsilon$.*

MSC2010: primary 11L40; secondary 11L20.
Keywords: quadratic nonresidue, Elliott–Halberstam conjecture, character sums, Burgess bound.

(See the end of the section for our conventions on asymptotic notation.) Linnik [1942] showed that this conjecture follows[1] from the generalized Riemann hypothesis; Ankeny [1952] improved the bound further to

$$n(p) \ll \log^2 p$$

on this hypothesis. However, Conjecture 1.1 remains open unconditionally; the best bound available (up to logarithmic factors) for general primes p is

$$n(p) \ll p^{1/4\sqrt{e}+\varepsilon} \tag{1-2}$$

for any fixed $\varepsilon > 0$, a well-known result of Burgess [1957]. It was also shown by Linnik [1942] unconditionally that, for any fixed $\varepsilon > 0$, the number of $p \le x$ with $n(p) > x^\varepsilon$ is bounded uniformly in x, and hence, the number of exceptions to the inequality $n(p) > p^\varepsilon$ with $p \le x$ is bounded by $O(\log\log x)$.

In this paper, we connect Vinogradov's conjecture to a standard conjecture in sieve theory, the *Elliott–Halberstam conjecture* [1970], as well as to a restricted fragment of this conjecture recently introduced by Zhang [2014]. The basic phenomenon being exploited here is that distribution estimates such as those given by the Elliott–Halberstam conjecture allow one to control correlations of the form[2]

$$\sum_n (\alpha * \beta)(n)(\gamma * \delta)(n+h) \tag{1-3}$$

for various arithmetic sequences α, β, γ, and δ and nontrivial shifts h, as long as all of the sequences α, β, γ, and δ vanish for very small values of n and provided that at least one of the sequences α, β, γ, or δ is "smooth" (e.g., if one of these sequences is an indicator function such as $1_{[N,2N]}$). On the other hand, by combining the multiplicativity and periodicity properties of Dirichlet characters with a hypothesis that the least quadratic residue is large (or that a character sum is large), we will be able to construct sums of the form (1-3) that deviate substantially from its expected value, giving the required contradiction. It is the periodicity of Dirichlet characters χ that allow us to introduce the shift h, thus transferring the problem from a multiplicative number theory problem (in which hypotheses

[1] In fact, the conjecture follows from even very weak fragments of this hypothesis; see, e.g., [Bateman and Diamond 2004, Theorem 10.6]. (Thanks to Kevin Ford for this reference.) The strongest result in this direction comes from a very recent work of Granville and Soundararajan [2015] (see also [Banks and Makarov 2014]), who showed (roughly speaking) that the only way this conjecture can fail is if a positive proportion of low-lying zeros of an L-function lies extremely close to the line $\operatorname{Re} s = 1$.

[2] If only the original Elliott–Halberstam conjecture is available, rather than its variants, then one of the convolutions $\alpha * \beta$ or $\gamma * \delta$ needs to be replaced by the von Mangoldt function Λ. Also, for technical reasons, it is convenient to ensure that one of the factors α, β, γ, or δ is supported on numbers coprime to the shift h.

such as the generalized Riemann hypothesis are useful) to a sieve theory problem (in which hypotheses such as the Elliott–Halberstam conjecture are useful). The arguments share some similarities with that of Burgess [1957] (which also relies heavily on the multiplicativity and periodicity properties of Dirichlet characters) but is ultimately powered by a somewhat different source of cancellation, namely the equidistribution assumptions of Elliott–Halberstam type rather[3] than the Weil exponential sum estimates.

To describe the results more precisely, we need some notation. For any function $\alpha : \mathbb{N} \to \mathbb{C}$ with finite support (that is, α is nonzero only on a finite set) and any primitive residue class $a\ (r)$, we define the (signed) *discrepancy* $\Delta(\alpha; a\ (r))$ to be the quantity

$$\Delta(\alpha; a\ (r)) := \sum_{n=a\ (r)} \alpha(n) - \frac{1}{\varphi(r)} \sum_{(n,r)=1} \alpha(n), \tag{1-4}$$

where φ is the Euler totient function.

Conjecture 1.2 (Elliott–Halberstam conjecture). *Let $0 < \vartheta < 1$ be fixed. Then*

$$\sum_{r<x^\vartheta} \sup_{a\in(\mathbb{Z}/r\mathbb{Z})^\times} |\Delta(\Lambda 1_{[1,x]}; a\ (r))| \ll x \log^{-A} x \tag{1-5}$$

for any fixed $A > 1$, where Λ is the von Mangoldt function. Equivalently, from the prime number theorem, one has

$$\sum_{r<x^\vartheta} \sup_{a\in(\mathbb{Z}/r\mathbb{Z})^\times} \left| \sum_{n\le x: n=a\ (r)} \Lambda(n) - \frac{x}{\varphi(r)} \right| \ll x \log^{-A} x$$

for any fixed $A > 1$.

The case $\vartheta < \frac{1}{2}$ of this conjecture is of course (a slightly weakened form of) the Bombieri–Vinogradov theorem [Bombieri 1965; Vinogradov 1965].

Our first theorem is then:

Theorem 1.3 (Elliott–Halberstam implies Vinogradov). *Conjecture 1.2 implies Conjecture 1.1.*

We prove this theorem in Section 2. The basic idea is to observe (from the general theory of mean values of multiplicative functions) that, if $n(q) > q^\varepsilon$ for some large prime q, then the character sum $\sum_{n\le x} \chi(n)\Lambda(n)$ will be anomalously large for

[3]It is worth noting however that much of the recent partial progress on the Elliott–Halberstam conjecture has proceeded by using Weil exponential sum estimates, although the precise estimates used there are different from those used in the Burgess argument. In Section 5, though, we sketch a version of the argument that allows for an improvement over the original bound (1-1) of Vinogradov using only the elementary bound on Kloosterman sums [1927] and does not require the full strength of the Weil conjectures.

some large $x = O(q^{O(1)})$, where χ is the quadratic character modulo q. As χ is periodic modulo q, this forces $\sum_{n\le x} \chi(n)\Lambda(n+q)$ to be large also. But one can use the Elliott–Halberstam conjecture (and an expansion of χ into divisor sums, using once again the largeness of $n(q)$) to obtain good bounds for $\sum_{n\le x} \chi(n)\Lambda(n+q)$ and obtain a contradiction.

With some additional combinatorial argument, we can obtain a similar implication[4] concerning the least primitive root modulo p, provided that $p-1$ has only boundedly many factors:

Theorem 1.4 (Elliott–Halberstam bounds least primitive roots). *Assume Conjecture 1.2. Then for any fixed $d \ge 1$ and fixed $\varepsilon > 0$ and any prime p for which $p-1$ is the product of at most d primes (counting multiplicity), the least primitive residue modulo p is $O(p^\varepsilon)$.*

We prove this theorem in Section 3.

Our proof of Theorem 1.3 does not easily allow one to convert partial progress on the Elliott–Halberstam conjecture to partial progress on Vinogradov's conjecture. We now present a different argument that replaces the Elliott–Halberstam conjecture by a conjecture on "Type II sums" of the type introduced[5] by Zhang [2014] with the feature that partial progress on the Type II conjecture implies partial progress on Vinogradov's conjecture. In particular, the Type II estimates in [Polymath 2014a] can be used to improve slightly upon the Vinogradov bound (1-1) by a method different from the Burgess argument, although the numerical exponent obtained is inferior to that in [Burgess 1957].

Let us first state the Type II conjecture, in a formulation suited for the current application.

Conjecture 1.5 (Type II conjecture). *Let $0 < \varpi < \frac{1}{4}$, and let $\delta > 0$ be a sufficiently small fixed quantity depending on ϑ. Let x be an asymptotic parameter going to infinity. Let P be any number that is the product of some subset of the primes in $[1, x^\delta]$; equivalently, let P be a square-free number all of whose prime factors are at most x^δ. Let N and M be quantities such that*

$$x^{1/2-2\varpi} \ll N \ll M \ll x^{1/2+2\varpi}$$

with $NM \asymp x$, and let $\alpha, \beta : \mathbb{N} \to \mathbb{R}$ be sequences supported on $[M, 2M]$ and $[N, 2N]$, respectively, such that one has the pointwise bounds

$$|\alpha(n)| \ll 1 \tag{1-6}$$

[4]We are indebted to Felipe Voloch for suggesting this variant.

[5]Zhang also considered "Type I" and "Type III" sums, which will not be of direct relevance in this paper, although the τ_3 distribution estimates mentioned in Section 5 are related to the Type III sums. Similar sums had also been previously considered by Bombieri, Fouvry, Friedlander, and Iwaniec [Bombieri et al. 1986; 1987; 1989; Fouvry 1984; 1985; Fouvry and Iwaniec 1980; 1983; 1992].

for all natural numbers n. We also assume that β is simply the indicator function

$$\beta = 1_{[N,2N]}.$$

Then one has

$$\sup_{1\le a\le x:(a,P)=1} \sum_{r\ll x^{1/2+2\varpi}:r|P} |\Delta(\alpha\star\beta; a\ (r))| \ll x \log^{-A} x \tag{1-7}$$

for any fixed $A > 0$.

This conjecture is implied by the generalized Elliott–Halberstam conjecture in [Polymath 2014b], which was in turn inspired by a similar conjecture in [Bombieri et al. 1986]. In [Motohashi 1976] (see also [Gallagher 1968]), a generalization of the Bombieri–Vinogradov theorem is obtained that roughly speaking implies (up to logarithmic factors) the $\varpi = 0$ endpoint of this conjecture. The arguments in [Zhang 2014] implicitly establish the above conjecture for $0 < \varpi < \frac{1}{1168}$, and more explicitly, the estimate in [Polymath 2014a, Theorem 5.1(iv)] establishes the conjecture for $0 < \varpi < \frac{1}{68}$. The estimates in those papers allow for more general values of a and r and more general sequences α and β than those considered here; however, the restricted version of Conjecture 1.5 stated above will suffice for our application. It is likely that the additional restrictions imposed here (particularly the requirement that β be the indicator function of an interval) allow for some improvement in the exponent $\frac{1}{68}$ obtained in [Polymath 2014a]; see also Section 5 below for a slightly different way to improve upon this exponent, from $\frac{1}{68}$ to $\frac{1}{28}$.

Our next main result is then:

Theorem 1.6 (Type II sums bound character sums). *Suppose that Conjecture 1.5 holds for a fixed choice of $0 < \varpi < \frac{1}{4}$. Then one has*

$$\Bigl|\sum_{n<q^{1/2-2\varpi+\varepsilon}} \chi(n)\Bigr| \ll q^{1/2-2\varpi+\varepsilon} \log^{-A} q \tag{1-8}$$

for any sufficiently small fixed $\varepsilon > 0$, any fixed $A > 0$, and any natural number q (not necessarily prime) whenever χ is a nonprincipal primitive Dirichlet character of conductor q.

By the usual argument of Vinogradov, this gives:

Corollary 1.7. *Suppose that Conjecture 1.5 holds for a fixed choice of $0 < \varpi < \frac{1}{4}$. Then one has*

$$n(q) \ll q^{(1/\sqrt{e})(1/2-2\varpi)+\varepsilon}$$

for any fixed $\varepsilon > 0$ and any prime q.

Proof. From the pointwise estimate

$$\chi(n) \geq 1 - 2 \sum_{p|n: p > n(q)} 1$$

for the quadratic character $\chi(n) := \left(\frac{n}{q}\right)$, we see that

$$\sum_{n<x} \chi(n) \geq x - 1 - 2 \sum_{n(q)<p\leq x} \left(\frac{x}{p} + 1\right)$$

for any $x > 1$. Setting $x := q^{1/2-2\varpi+\varepsilon}$ for some $\varepsilon > 0$ and using Theorem 1.6, we see that

$$x - 2x \sum_{n(q)<p\leq x} \frac{1}{p} \leq o(x)$$

as $q \to \infty$. From Mertens' theorem, this implies that

$$\log \frac{\log x}{\log n(q)} \geq \tfrac{1}{2} + o(1),$$

and the claim follows. $\square$

In particular, the Type II estimates in [Polymath 2014a] give the improvement

$$n(p) \ll p^{(1/\sqrt{e})(1/2-1/34)+\varepsilon}$$

to (1-1) for any fixed $\varepsilon > 0$. This is well short of the improvement in (1-2); however, it represents a slightly different way to break the "square root barrier" from the Burgess argument; for instance, the arguments can extend to more general moduli than primes p without much difficulty, whereas the Burgess argument encounters some additional technical issues when the modulus is not cube-free. One will be able to surpass the Burgess bound as soon as one can establish a Type II estimate for some $\varpi > \frac{1}{8}$ (or $\varpi > \frac{1}{12}$ in the non-cube-free case); thus, one needs to improve the Type II exponents in [Polymath 2014a] by a factor of roughly eight. Interestingly, it was noted in [Bombieri et al. 1986, Conjecture 3] that, if one assumed square root cancellation in certain exponential sums, one could obtain Type II estimates for all $\varpi < \frac{1}{8}$, thus falling barely short of being able to improve upon the Burgess bound.

Theorem 1.6, when combined with the Type II estimates in [Polymath 2014a], establishes the short character sum bounds

$$\sum_{n<q^{1/2-1/34+\varepsilon}} \chi(n) = q^{1/2-1/34+\varepsilon} \log^{-A} q \tag{1-9}$$

for any primitive character χ of conductor q. This bound is inferior to that of Burgess [1957; 1963; 1986], which establishes

$$\sum_{M\leq n\leq M+N} \chi(n) = N^{1-\delta(\varepsilon)}$$

for arbitrary M when $N \gg q^{1/3+\varepsilon}$ (if q is not cube-free) or $N \gg q^{1/4+\varepsilon}$ (if q is cube-free), and $\delta(\varepsilon) > 0$ depends only on ε. With our methods, one would need Type II estimates at level of distribution at least $\frac{2}{3}$ (thus $\varpi > \frac{1}{12}$) to improve upon the Burgess bound in the non-cube-free setting or at least $\frac{3}{4}$ (thus $\varpi > 1/8$) in the cube-free setting. Note also the Burgess bound has also been improved for certain types of modulus q, such as smooth numbers (see, e.g., [Graham and Ringrose 1990; Goldmakher 2010]) or prime powers (see, e.g., [Postnikov 1956]).

Remark 1.8. If one had the Type II estimates for all $0 < \varpi < \frac{1}{4}$, then (by combining Corollary 1.7 with the Burgess bound) we would have

$$\sum_{n \le x} \chi(n) \ll x \log^{-A} x$$

for all $x \ge q^{\varepsilon}$ and fixed $A, \varepsilon > 0$, and hence (by summation by parts), one would obtain a very slight improvement $L(1, \chi) = o(\log q)$ to the standard upper bound $L(1, \chi) = O(\log q)$ for the sum $L(1, \chi) = \sum_n \chi(n)/n$. Furthermore, one obtains the bound $L(s, \chi) = O(\log^2 q)$ (say) when $|s - 1| \le A \log\log q / \log q$ for any fixed A. Using this and standard arguments (see, e.g., [Iwaniec and Kowalski 2004, Chapter 8]), one can enlarge[6] the classical zero-free region of $L(s, \chi)$ to include the region $|s-1| \le A/\log q$ for any fixed $A > 0$, except possibly for a Siegel zero. This in turn can be used to improve the prime number theorem of Gallagher [1970] and hence also the constant in Linnik's theorem on primes in an arithmetic progression, assuming the Type II estimates and possibly excluding an exceptional modulus; we omit the details.

Remark 1.9. By standard arguments (see, e.g., [Montgomery and Vaughan 2007, Corollary 9.20]) starting from the observation that the sum

$$\sum_{d|Q} \frac{\varphi(Q/d)\mu(d)}{Q} \sum_{\substack{\chi(Q) \\ \operatorname{ord}\chi = d}} \sum_{n \le x} \chi(n)$$

counts the number of primitive roots modulo a prime p up to x, where Q is the product of all the primes dividing $p-1$, we see that Theorem 1.6 implies that, if one has Type II estimates for a given $0 < \varpi < \frac{1}{4}$, then the least primitive root of $\mathbb{Z}/p\mathbb{Z}$ is $O(p^{1/2-2\varpi+\varepsilon})$ for any fixed ε and any prime p, provided that $p-1$ has at most $O(\log\log p)$ prime factors; we leave the details to the interested reader. In particular, we can strengthen the conclusion of Theorem 1.4 slightly if we replace the Elliott–Halberstam conjecture by the Type II conjecture for ϖ arbitrarily close to $\frac{1}{4}$. It may be possible[7] to remove the requirement on the number of prime factors

[6]We thank James Maynard for this remark.

[7]We thank the anonymous referee for this suggestion.

of $p-1$ by using zero-density estimates (together with a result of Rodosskiĭ [1956] linking L-function zeros with character sums; see also the recent preprints [Banks and Makarov 2014; Granville and Soundararajan 2015]) to show that $\sum_{n \le x} \chi(n)$ is small for most characters χ; we will not pursue this in detail here.

Remark 1.10. Suppose Conjecture 1.5 holds for some fixed $0 < \varpi < \frac{1}{4}$, and suppose that q is a large prime such that the least prime quadratic *residue* is at least[8] $q^{1/2-2\varpi+\varepsilon}$. Then, letting χ be the quadratic character of conductor q, one has $\chi(n) = \lambda(n)$ for all $n \le q^{1/2-2\varpi+\varepsilon}$, where λ is the Liouville function. From the prime number theorem (for $n \le q^{1/2-2\varpi+\varepsilon}$) and Theorem 1.6, we conclude that $\sum_n \chi(n)/n \ll \log^{-A} q$ and $\sum_n \chi(n) \log n/n \gg 1$, so $|L'(1,\chi)/L(1,\chi)| \gg \log^A q$ for any fixed A. From standard arguments, this implies that one has a Siegel zero $L(\sigma,\chi) = 0$ with $1-\sigma \ll \log^{-A} q$ for any fixed A. Thus, if one could rule out Siegel zeros, one could use Type II estimates to bound the least prime quadratic residue. If one could improve the $\log^{-A} q$ gain in (1-8) to a power saving $q^{-\varepsilon}$, then Siegel's theorem could be used to remove the need to consider Siegel zeros; for instance, this argument recovers the standard bound of $q^{1/4+o(1)}$ for the least prime quadratic residue coming from the Burgess bound. However, our arguments would require a similar power saving in the Type II estimates to achieve this, which may be an overly ambitious hypothesis.

We prove Theorem 1.6 in Section 4. The idea here is to exploit the fact that, if $\sum_{n \in [N/2,N]} \chi(n)$ is large, then on an interval $[1,x]$ with $x = q^{1+O(\varepsilon)}$, $\chi(n)$ will exhibit large correlation with $\alpha * \beta(n+jq)$ for any $j = O(q^\varepsilon)$, where $\beta := 1_{[N/2,N]}$ and α is the restriction of χ to smooth square-free numbers of magnitude close to x/N and that are coprime to q. This is because of the multiplicativity and periodicity properties of χ. An application of Cauchy–Schwarz (i.e., the dispersion method) then shows that $\alpha * \beta(n+jq)$ and $\alpha * \beta(n+j'q)$ correlate with each other for some distinct j and j', but one can use Type II estimates to prevent this scenario from occurring.

Remark 1.11. The above argument shares many similarities with the argument of Burgess [1957]. Both arguments rely heavily on the periodicity and multiplicativity of the Dirichlet character χ, which allows one to start with a hypothesis that a single character sum $\sum_{n \le x} \chi(n)$ is large and deduce that χ is biased on many arithmetic progressions. In the current argument, one exploits the bias of χ on medium-length arithmetic progressions (of length about $q^{1/2-2\varpi}$) and varying modulus; in contrast, the argument of Burgess exploits the bias of χ on many (close to $q^{1/2}$) very short progressions (of length q^ε for some small ε) and fixed modulus. Unfortunately, the author was not able to combine the two methods together to obtain any improvement

[8] We thank John Friedlander for suggesting this problem.

on (1-2) without assuming a large portion of the Elliott–Halberstam or Type II conjectures.

Remark 1.12. The proof of Theorem 1.6 may possibly extend to cover the shifted character sums $\sum_{M \le n \le M+N} \chi(n)$ appearing in the work of Burgess; however, the way the argument is currently presented, this would require a shifted version of a Type II estimate in which the convolution $\alpha * \beta$ is replaced by a shifted convolution. As such, one can no longer directly quote the results from [Polymath 2014a] to obtain a result for such shifted sums; however, it is plausible that some modification of the *proof* of the Type II estimate in [Polymath 2014a] can still be adapted to this shifted setting. We do not pursue this matter here (as with the centered sums, we do not seem to directly improve upon the Burgess bounds at the current level of technology for equidistribution estimates).

A variant of the argument used to prove Theorem 1.6, which we discuss in Section 5 below, allows one to use distributional estimates for the higher divisor functions

$$\tau_k(n) := \sum_{n_1,\dots,n_k : n_1 \cdots n_k = n} 1 \tag{1-10}$$

(or more precisely, from dyadic components of such functions) in place of Type II estimates to obtain similar results. Roughly speaking, a distributional estimate on τ_k at level θ implies a bound of the form (1-8) with $\frac{1}{2} - 2\varpi$ replaced by $\max(1-\theta, 1/(k\theta+1))$; thus, for instance, the classical distribution estimate of τ_2 at $\theta = \frac{2}{3}$ gives (1-8) with $\varpi = \frac{1}{28}$, slightly improving upon (1-9) though still short of the Burgess bounds in both cube-free and non-cube-free cases. More recently, a level of distribution $\frac{4}{7}$ has been established (in a restricted averaged sense) for τ_3 in [Fouvry et al. 2014], which (morally at least) also recovers (1-8) with $\varpi = \frac{1}{28}$. To improve upon the Burgess bound, one would need τ_k at level of distribution above $\frac{2}{3}$ for some $k \ge 3$ (in the non-cube-free case) or above $\frac{3}{4}$ for some $k \ge 4$ (in the cube-free case). Both results seem unfortunately to be out of reach of current methods.

A similar analysis, again discussed in Section 5 below, suggests that one should be able to improve the exponent $\frac{1}{2} - 2\varpi$ in (1-8) to $1/k - c$ for some $c > 0$ provided that one can obtain good asymptotics for sums such as

$$\sum_{n \le x} \tau_k(n)\tau_k(n+q)$$

with $q = o(x)$. In particular, controlling such sums for $k = 3$ would (morally, at least) improve upon the non-cube-free Burgess bound and for $k = 4$ would improve upon the cube-free Burgess bound. Unfortunately, rigorous asymptotics for these sums have only been established for $k = 2$.

Notation. We use the following asymptotic notation. We allow for an asymptotic parameter (e.g., x or q) to go to infinity; quantities in this paper may depend on this parameter unless they are explicitly labeled as *fixed*. We then write $X \ll Y$, $X = O(Y)$, or $Y \gg X$ if one has $|X| \leq CY$ for some fixed C (in particular, C can depend on other parameters as long as they are also fixed). We also write $X = o(Y)$ if we have $|X| \leq cY$ for some quantity c that goes to zero as the asymptotic parameter goes to infinity and write $X \asymp Y$ for $X \ll Y \ll X$.

Sums over p are understood to be over primes, and all other sums are over the natural numbers $\mathbb{N} = \{1, 2, 3, \dots\}$ unless otherwise indicated.

Given two functions $f, g : \mathbb{N} \to \mathbb{C}$, their Dirichlet convolution $f * g$ is defined by

$$f * g(n) := \sum_{d|n} f(d) g\left(\frac{n}{d}\right),$$

where $d \mid n$ denotes the assertion that d divides n.

Given two natural numbers a and b, we use (a, b) to denote the greatest common divisor of a and b and $a\ (b)$ to denote the residue class of integers equal to a modulo b. Given a natural number r, we use $(\mathbb{Z}/r\mathbb{Z})^\times = \{a\ (r) : (a, r) = 1\}$ to denote the primitive residue classes modulo r.

We use 1_E to denote the indicator function of E; thus, $1_E(n)$ equals 1 when $n \in E$ and equals 0 otherwise. Similarly, if S is a sentence, we write 1_S to equal 1 when S is true and 0 otherwise; thus, for instance, $1_E(n) = 1_{n \in E}$.

2. Vinogradov from Elliott–Halberstam

We now prove Theorem 1.3. We will in fact prove a slightly stronger implication, in which Conjecture 1.1 is replaced by:

Conjecture 2.1. *For any Dirichlet character χ, let n_χ be the first natural number with $\chi(n_\chi) \neq 1$. For any fixed $\varepsilon > 0$, we have $n_\chi \ll q^\varepsilon$ for any primitive Dirichlet character χ of prime conductor q.*

Clearly, Conjecture 1.1 is the special case of Conjecture 2.1 in which χ is a quadratic character.

Assume the Elliott–Halberstam conjecture. Suppose for sake of contradiction that Conjecture 1.1 failed; then we can find a fixed $\kappa > 0$ and a sequence q of primes going to infinity, as well as a character χ of modulus q, such that

$$n_\chi > q^\kappa.$$

Without loss of generality, we may take κ to be small, e.g., $\kappa < \frac{1}{2}$. We view q as an asymptotic parameter for the purposes of asymptotic notation and reserve the right to refine q to subsequences as necessary.

We will need some basic results from the theory of mean values of multiplicative functions in order to produce some anomalous distribution for $\chi(n)\Lambda(n)$ at large scales. This could be accomplished using the results of Granville and Soundararajan [2001] (or even the earlier work of Wirsing [1967]), but we do not need the full strength of their theory here since we will be satisfied with an analysis of logarithmic densities such as $(1/\log x)\sum_{n\le x}\chi(n)/n$ instead of natural densities such as $(1/x)\sum_{n\le x}\chi(n)$. As such, we give a self-contained treatment here.

It will be technically convenient to work in the asymptotic limit in which we extract the mean value after sending q to infinity (this is a luxury available in the logarithmic density setting that is not easily achievable for natural densities, at least if one is not willing to use the tools of nonstandard analysis). For any fixed $t\ge 0$, we consider the logarithmic densities

$$A_q(t):=\frac{1}{\log q}\sum_{n<q^t}\frac{\chi(n)}{n},$$

$$B_q(t):=\frac{1}{\log q}\sum_{n<q^t}\frac{\chi(n)\Lambda(n)}{n}.$$

From Mertens' theorem, we have the Lipschitz bounds

$$|A_q(t)-A_q(s)|,\ |B_q(t)-B_q(s)|\le |t-s|+o(1) \tag{2-1}$$

for all fixed $t,s\ge 0$; also we clearly have $A_q(0)=B_q(0)=0$. From the Arzelà–Ascoli theorem, and refining q to a subsequence as necessary, we may thus find *fixed* Lipschitz functions $A,B:[0,+\infty)\to\mathbb{C}$ such that

$$A_q(t)=A(t)+o(1),\qquad B_q(t)=B(t)+o(1) \tag{2-2}$$

for all fixed $t\ge 0$; that is to say that A_q and B_q converge locally uniformly to A and B, respectively. (The traditional form of the Arzelà–Ascoli theorem allows one to pass to a subsequence on which one has uniform convergence on $[0,n]$ for each natural number n, and then a further diagonalization gives locally uniform convergence on $[0,+\infty)$.) From (2-1), we have

$$|A(t)-A(s)|,\ |B(t)-B(s)|\le |t-s|$$

for all fixed $t,s\ge 0$. By the Rademacher differentiation theorem, we can thus find Lebesgue-measurable functions $a,b:[0,+\infty)\to\mathbb{C}$ bounded in magnitude by 1, defined up to almost-everywhere equivalence, such that

$$A(t)=\int_0^t a(u)\,du,\qquad B(t)=\int_0^t b(u)\,du$$

for all $t\in[0,+\infty)$.

We now establish some bounds on A and B. Since χ has mean zero on intervals of length q, it is easy to see that

$$A_q(t) = A_q(t') + o(1)$$

for all fixed $t, t' > 1$; in fact, one can extend this to $t, t' > \frac{1}{4}$ using the Burgess bound [1957], but we will not need to do so here. This implies that a is supported on [0, 1] (modulo null sets).

Next, since $\chi(n) = 1$ for $n \leq q^\kappa$, we have from Mertens' theorem that

$$A_q(t),\ B_q(t) = t + o(1)$$

for $t < \kappa$. Thus, $A(t) = B(t) = t$ for $t < \kappa$, and so $a(t) = b(t) = 1$ for $t < \kappa$ (again up to null sets).

Next, we claim that a and b obey the integral equation of Wirsing [1967]:

Lemma 2.2 (Wirsing equation). *We have*

$$ta(t) = \int_0^t a(u)b(t-u)\,du$$

for almost all $t > 0$.

This equation also holds for means other than logarithmic densities (replacing a and b by suitable substitutes, such as the functions $t \mapsto (1/q^t)\sum_{n\leq q^t}\chi(n)$ and $t \mapsto (1/q^t)\sum_{n\leq q^t}\chi(n)\Lambda(n)$, respectively), but the arguments are more complicated, and one has to work nonasymptotically and admit some $o(1)$ errors; see [Wirsing 1967; Granville and Soundararajan 2001].

Proof. We start with the Dirichlet convolution identity

$$\chi(n)\log n = (\chi\Lambda) * \chi(n)$$

and conclude for any fixed $t > 0$ that

$$\frac{1}{\log^2 q}\sum_{n\leq q^t}\frac{\chi(n)\log n}{n} = \frac{1}{\log q}\sum_{d\leq q^t}\frac{\chi(d)\Lambda(d)}{d}\frac{1}{\log q}\sum_{m\leq q^t/d}\frac{\chi(m)}{m}. \tag{2-3}$$

To estimate this expression, we use a Riemann sum argument. Let $J > 0$ be a large fixed natural number. If $q^{(j-1)t/J} \leq d < q^{jt/J}$ for some $1 \leq j \leq J$, then $(1/\log q)\sum_{m\leq q^t/d}\chi(m)/m = A(t - jt/J) + O(1/J) + o(1)$ (with implied constant uniform in J), and so the expression (2-3) may be written (after using Mertens' theorem to estimate error terms) as

$$\left(\sum_{j=1}^{J} A\left(t - \frac{jt}{J}\right)\frac{1}{\log q}\sum_{q^{(j-1)t/J}\leq d<q^{jt/J}}\frac{\chi(d)\Lambda(d)}{d}\right) + O\left(\frac{1}{J}\right) + o(1).$$

One has

$$\frac{1}{\log q}\sum_{q^{(j-1)t/J}\le d<q^{jt/J}}\frac{\chi(d)\Lambda(d)}{d}=B(jt/J)-B((j-1)t/J)+o(1)$$
$$=\int_{(j-1)t/J}^{jt/J}b(u)\,du+o(1),$$

and so (by the Lipschitz nature of A), the previous expression becomes

$$\int_0^1 A(t-u)b(u)\,du+O\left(\frac{1}{J}\right)+o(1).$$

As J can be arbitrarily large, we conclude that

$$\frac{1}{\log^2 q}\sum_{n\le q^t}\frac{\chi(n)\log n}{n}=\int_0^t A(t-u)b(u)\,du+o(1).$$

On the other hand, from the identity $\log n/\log q=t-\int_0^t 1_{n\le q^u}\,du$ and (2-2), we see (after a Riemann sum argument as before) that

$$\frac{1}{\log^2 q}\sum_{n\le q^t}\frac{\chi(n)\log n}{n}=tA(t)-\int_0^t A(u)\,du+o(1)$$

and hence

$$tA(t)-\int_0^t A(u)\,du=\int_0^t A(t-u)b(u)\,du$$

for all t. Differentiating using the Lebesgue differentiation theorem, we conclude that

$$ta(t)=\int_0^t a(t-u)b(u)\,du$$

almost everywhere, as desired. □

We will use this equation, together with some complex analysis and the previously established compact support of a, to derive the following consequence:

Corollary 2.3. *b is not compactly supported* (*up to null sets*).

Proof. Suppose for contradiction that b is compactly supported (modulo null sets). Now consider the Fourier–Laplace transforms

$$\mathscr{L}a(s):=\int_0^\infty a(t)e^{-ts}\,dt,$$
$$\mathscr{L}b(s):=\int_0^\infty b(t)e^{-ts}\,dt;$$

as a and b are both bounded and compactly supported, the functions $\mathscr{L}a$ and $\mathscr{L}b$ are entire and of at most exponential growth and are not identically zero since a

and b are not identically zero. On the other hand, from Lemma 2.2 and standard computations, we have

$$-\frac{d}{ds}\mathscr{L}a = \mathscr{L}a \times \mathscr{L}b. \tag{2-4}$$

As $\mathscr{L}b$ has no poles, $\mathscr{L}a$ cannot have any zeros; in particular, $\log \mathscr{L}a$ is entire and grows at most linearly and must therefore be a linear function so that $\mathscr{L}a$ is an exponential function, and hence, by (2-4), $\mathscr{L}b$ is a constant function. But this is absurd (it contradicts the Riemann–Lebesgue lemma). □

Remark 2.4. The above argument shows that a and b cannot both be compactly supported while still obeying Lemma 2.2, except in trivial cases. A stronger result in this regard, in which a and b are allowed to decay exponentially, can be found in [Granville and Soundararajan 2007]. Note that the argument used to establish this corollary would have been significantly messier if one had to contend with $o(1)$ errors in the Wirsing integral equation as one would need quantitative approximate versions of various basic qualitative facts about entire functions. This is the main reason why we took the asymptotic limit $q \to \infty$ previously. However, Andrew Granville (private communication) has informed me that such an approximate version of this observation was obtained in an unpublished work of Granville and Soundararajan. (See also the recent paper [Granville and Soundararajan 2015] for some related results.)

From the above corollary and the Lebesgue differentiation theorem, we can find fixed $1 < t_1 < t_2$ such that $|B(t_2) - B(t_1)| > 0$, and so

$$\left| \frac{1}{\log q} \sum_{q^{t_1} < n < q^{t_2}} \frac{\chi(n)\Lambda(n)}{n} \right| \gg 1$$

for q sufficiently large. By the pigeonhole principle, we may thus find $q^{t_1} \ll x \ll q^{t_2}$ such that

$$\left| \sum_{n \in [x/2, x]} \chi(n)\Lambda(n) \right| \gg x.$$

Of course, x will depend on q. Since $q = o(x)$, we may shift n by q, using the periodicity of χ, to conclude that

$$\left| \sum_{n \in [x/2, x]} \chi(n)\Lambda(n+q) \right| \gg x.$$

On the other hand, as χ has mean zero on intervals of length q, we have

$$\sum_{n \in [x/2, x]} \chi(n) = o(x).$$

Thus, if we let

$$X := \sum_{n\in[x/2,x]} \chi(n)(\Lambda(n+q)-1),$$

then we have

$$|X| \gg x \tag{2-5}$$

for sufficiently large q.

We now upper-bound X in order to contradict (2-5). The first step is to expand out χ in terms of Dirichlet convolutions. By Möbius inversion, we can express

$$\chi = 1 * f = 1 + 1 * \tilde{f},$$

where

$$\tilde{f}(n) := f(n) - 1_{n=1}$$

and

$$f = \chi * \mu;$$

in other words, f is the multiplicative function with

$$f(p^j) = \chi(p)^{j-1}(\chi(p)-1)$$

whenever p is a prime and $j \geq 1$, with the convention that $0^0 = 1$. In particular, we see that $f(n)$ is only nonzero when n is q^κ*-rough*, by which we mean that n has no prime factor less than or equal to q^κ; this implies furthermore that $\tilde{f}(n)$ vanishes unless $n > q^\kappa$ and that

$$|\tilde{f}(n)| \ll 1 \tag{2-6}$$

whenever $n = O(q^{O(1)})$.

Let $\nu > 0$ be a small fixed constant to be chosen later. We expand X using the identity

$$\chi 1_{[x/2,x]} = 1_{[x/2,x]} + (1_{[1,x^\nu)} * \tilde{f})1_{[x/2,x]} + (1_{[x^\nu,q^{-\kappa}x]} * \tilde{f})1_{[x/2,x]}, \tag{2-7}$$

where we have used the fact that $\tilde{f}(n)$ vanishes for $n < q^\kappa$. This gives the splitting

$$X = X_1 + X_2 + X_3$$

where

$$X_1 = \sum_{n\in[x/2,x]} (\Lambda(n+q)-1),$$

$$X_2 = \sum_{n\in[x/2,x]} (1_{[1,x^\nu)} * \tilde{f})(n)(\Lambda(n+q)-1),$$

$$X_3 = \sum_{n\in[x/2,x]} (1_{[x^\nu,q^{-\kappa}x]} * \tilde{f})(n)(\Lambda(n+q)-1).$$

From the prime number theorem, we have

$$X_1 = o(x).$$

For X_2, we use the triangle inequality to bound

$$|X_2| \leq \sum_{d<x^\nu} \sum_{x/2d \leq m \leq x/d} |\tilde{f}(m)|(\Lambda(dm+q)+1).$$

We claim that

$$\sum_{x/2d \leq m \leq x/d} |\tilde{f}(m)|\Lambda(dm+q) \ll \frac{x}{\varphi(d)\log x} \tag{2-8}$$

and

$$\sum_{x/2d \leq m \leq x/d} |\tilde{f}(m)| \ll \frac{x}{d \log x} \tag{2-9}$$

for all $d < x^\nu$, and hence,

$$X_2 \ll \nu x$$

with implied constant independent of ν.

We first prove (2-8). From (2-6), we have $|\tilde{f}(m)|\Lambda(dm+q) = O(\log x)$, and this expression vanishes unless m and $dm+q$ are both q^κ-rough, except for a small exceptional contribution (coming from when $dm+q$ is the power of a small prime) that can easily be seen to be negligible. Removing this exceptional contribution, we see that we are removing two residue classes modulus p from the interval of m for each prime $p < x^\kappa$ not dividing d. Using a standard upper-bound sieve (see, e.g., [Friedlander and Iwaniec 2010]), we conclude that the number of surviving summands m is $O(x/(\varphi(d)\log^2 x))$, and the claim follows. The bound (2-9) is established similarly, except now we bound $|\tilde{f}(m)| = O(1)$ and we remove just a single residue class for each prime p, rather than two.

Finally we turn to X_3. We expand

$$X_3 = \sum_{q^\kappa \ll r \ll x^{1-\nu}} \tilde{f}(r) \sum_{m \in [x/2r, x/r] \cap [x^\nu, q^{-\kappa} x]} (\Lambda(rm+q)-1).$$

The contribution when $r \asymp q^\kappa$ or $r \asymp x^{1-\nu}$ can be seen to be $O(x/\log x)$ using the Brun–Titchmarsh inequality (and upper-bound sieve bounds on q^κ-rough numbers, as in the estimation of X_2). The contribution when r is divisible by q can be treated similarly (in fact one has the better bound of $O(x/q)$ in this case). So we may write

$$X_3 = \sum_{2q^\kappa < r < x^{1-\nu}/2;\, (r,q)=1} \tilde{f}(r) \sum_{x/2r \leq m \leq x/r} (\Lambda(rm+q)-1) + o(x)$$

or equivalently (since q is significantly smaller than x)

$$X_3 = \sum_{2q^\kappa < r < x^{1-\nu}/2;\, (r,q)=1} \tilde{f}(r) \sum_{n \in [x/2, x]: n = q\ (r)} (\Lambda(n)-1) + o(x).$$

Invoking the Elliott–Halberstam conjecture and the prime number theorem, we then have

$$X_3 = \sum_{2q^\kappa < r < x^{1-\nu}/2;\,(r,q)=1} \tilde{f}(r)\left(\frac{1}{\varphi(r)}\frac{x}{2} - \frac{1}{r}\frac{x}{2}\right) + o(x).$$

If r contributes to the above sum, then it is the product of $O(1)$ primes of size at least q^κ, and so $1/\varphi(r) = 1/r + O(q^{-\kappa}/r)$. From this, we see that

$$X_3 = o(x).$$

Putting all this together, we conclude that

$$|X| \ll (\nu + o(1))x,$$

contradicting (2-5) for ν small enough. This completes the proof of Theorem 1.3.

Remark 2.5. Our arguments here do not easily give any effective quantitative bound on $n(p)$ due to our use of asymptotic limits; in particular, the fixed quantities t_1 and t_2 appearing above were obtained by what is essentially a compactness argument and thus not obviously effective. It is likely that a more carefully quantitative version of the above argument (perhaps using the estimates from [Granville and Soundararajan 2001]) can make this portion of the argument effective, thus allowing one to derive partial progress on the Vinogradov conjecture from sufficiently strong partial progress on the Elliott–Halberstam conjecture; however, the dependence of constants will be far worse than in Theorem 1.6. We will not pursue this question further here.

Remark 2.6. Suppose the Burgess bound (1-2) was sharp up to epsilon factors, in the sense that one could find a sequence of primes q going to infinity with $n(q) = q^{1/4\sqrt{e}+o(1)}$. Then by extracting a limit to obtain the functions a and b as above, we see that $a(t) = b(t) = 1$ for $t \le 1/4\sqrt{e}$ and (from the Burgess character sum bounds) $a(t) = 0$ for $t > \frac{1}{4}$. As was first observed by Heath-Brown (see, e.g., Appendix 2 of [Diamond et al. 2006]), this information allows one in this case to determine the functions a and b completely. Indeed, in the range $1/4\sqrt{e} \le t < 1/2\sqrt{e}$, one has from Lemma 2.2 that

$$ta(t) = \int_0^t a(u)\,du - \int_0^{t-1/4\sqrt{e}} (1 - b(t-u))\,du.$$

Bounding $1 - b(t-u)$ by 2, we thus have

$$ta(t) \ge \int_0^t a(u)\,du - 2(t - 1/4\sqrt{e})$$

and thus by Gronwall's inequality

$$a(t) \ge 1 - 2\log(4\sqrt{e}t).$$

(Indeed, one can verify that the difference $f(t) := a(t) - 1 + 2\log(4\sqrt{e}t)$ obeys the inequality $tf(t) \geq \int_{1/4\sqrt{e}}^{t} f(u)\,du$ for $1/4\sqrt{e} \leq t < 1/2\sqrt{e}$ with $f(1/4\sqrt{e}) = 0$.) Since equality is attained for $t = \frac{1}{4}$ (note from Lemma 2.2 that a is continuous), we must have $1 - b(t-u) = 2$ whenever $t \leq \frac{1}{4}$ and $0 \leq u < t - 1/4\sqrt{e}$; that is to say $b(t) = -1$ for $1/4\sqrt{e} < t \leq \frac{1}{4}$; also $a(t) = 1 - 2\log(4\sqrt{e}t)$ in this range. For $t > \frac{1}{4}$, Lemma 2.2 gives

$$0 = \int_0^t a(t-u)b(u)\,du,$$

which on differentiation gives the integral equation

$$b(t) = 2\int_{1/4\sqrt{e}}^{1/4} b(t-u)\,\frac{du}{u},$$

which can then be used to complete the description of b, for instance via Laplace transforms. For instance, we see that $b(t) = 1$ for $\frac{1}{4} < t \leq 1/2\sqrt{e}$. One can compute that b does not vanish near $t = 1$, in which case the argument above shows that some improvement upon (1-2) can be made provided one can establish the Elliott–Halberstam conjecture for some $\vartheta > 1 - 1/4\sqrt{e} \approx 0.8484$.

3. From Elliott–Halberstam to the least primitive root

We now prove Theorem 1.4. The key new tool is the following combinatorial statement. Given a subset A of an additive group $G = (G, +)$ and a natural number k, define the iterated sumset kA to be the set of all sums $a_1 + \cdots + a_k$, where $a_1, \ldots, a_k$ are elements in A (allowing repetition).

Proposition 3.1 (escape from cosets). *Let $d, m \geq 1$ be fixed integers. Then there exists a natural number k with the following property: whenever G is a finite additive group whose order is the product of at most d primes (counting multiplicity) and A is a subset of G containing 0 for which one has inclusions of the form*

$$kA \subset \bigcup_{i=1}^{m} x_i + H_i \subsetneq G$$

for some cosets $x_i + H_i$ of subgroups H_i of G, then A is contained in a proper subgroup of G.

In the contrapositive, Proposition 3.1 asserts that, if A generates G and contains 0, then the iterated sumsets kA for k large enough cannot be covered by a small number of cosets of subgroups of G, unless these cosets of subgroups already covered all of G. Thus, the sumsets kA "escape" all nontrivial unions of boundedly many cosets. This result can be viewed as a simple abelian variant of the nonabelian "escape from subvarieties" lemma that first appeared in [Eskin et al. 2005].

Let us assume this proposition for the moment and see how it implies Theorem 1.4. Assume the Elliott–Halberstam conjecture, and assume for sake of contradiction that the conclusion of Theorem 1.4 failed. Carefully negating the quantifiers, this means that we can find a sequence of primes p going off to infinity, with $p-1$ being the product of $O(1)$ primes, and a fixed $\kappa > 0$, with the property that the least primitive root of $\mathbb{Z}/p\mathbb{Z}$ is at least p^κ.

Using a discrete logarithm, we have an isomorphism $\log : (\mathbb{Z}/p\mathbb{Z})^\times \to G$ from the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ to the additive cyclic group $G := \mathbb{Z}/(p-1)\mathbb{Z}$. If n is a natural number less than p^κ, then by hypothesis n is not a primitive root of $(\mathbb{Z}/p\mathbb{Z})^\times$, which implies that

$$\log(n) \subset \bigcup_{r|p-1:r<p-1} \{x \in G : rx = 0\} \subsetneq G.$$

In particular, for any natural number k, if we set $A := \{\log(n) : 1 \le n < p^{\kappa/k}\}$, then

$$kA \subset \bigcup_{r|p-1:r<p-1} \{x \in G : rx = 0\} \subsetneq G.$$

Since $\log(1) = 0$, A contains 0. Applying Proposition 3.1 (and using the hypothesis that $p-1$ is the product of $O(1)$ primes), we conclude (for k large enough) that A is contained in a proper subgroup of G. Equivalently, A lies in the kernel of a primitive character χ of conductor p; thus, $\chi(n) = 1$ for all $n < p^{\kappa/k}$. But this contradicts Conjecture 2.1, which as we saw in the previous section was a consequence of the Elliott–Halberstam conjecture.

It remains to prove Proposition 3.1. To illustrate the proposition, let us first give a simple case when G is a direct product $H_1 \times H_2$ and we are given that $0 \in A$ and

$$2A \subset (H_1 \times \{0\}) \cup (\{0\} \times H_2).$$

We claim that this forces either $A \subset H_1 \times \{0\}$ or $A \subset \{0\} \times H_2$. Indeed, if neither of these statements were true, then either there would exist $a \in A$ that was outside both $H_1 \times \{0\}$ and $\{0\} \times H_2$ or else there would exist $a_1, a_2 \in A$ with $a_1 \in H_1 \times \{0\}$, $a_2 \in \{0\} \times H_2$, and $a_1, a_2 \neq 0$. In either case, we could find an element of $2A$ ($a + 0$ or $a_1 + a_2$, respectively) that was outside of $(H_1 \times \{0\}) \cup (\{0\} \times H_2)$, giving the desired contradiction. This simple special case is already sufficient to handle the case of Theorem 1.4 in which $p - 1$ is the product of just two primes (that is, $p - 1 = 2q$ for some prime q) although in this case it turns out that the least primitive root is also the least quadratic nonresidue (for p large enough, at least), so the claim in this case is already immediate from Theorem 1.3.

The general case can be obtained by a rather complicated induction on the "complexity" of the covering set $\bigcup_{i=1}^m x_i + H_i$, as follows. Fix a natural number d.

Define a *configuration* to be a tuple

$$(k, G, A, m, (x_i + H_i)_{i=1}^m), \tag{3-1}$$

where k and m are natural numbers, G is a finite additive group with $|G|$ the product of d primes, A is a subset of G containing 0 and not contained in any proper subgroup of G, and the $x_i + H_i$ are distinct cosets in G, such that

$$kA \subset \bigcup_{i=1}^{m} x_i + H_i \subsetneq G. \tag{3-2}$$

In particular, this implies that $H_i \neq G$ for each i. Our task is to show that, for any configuration (3-1), k is bounded by a quantity depending only on d and m.

Suppose for contradiction that this claim failed. Then we can find a sequence of configurations (3-1) in which m stays constant but k goes to infinity. (The other data G, A, x_i, and H_i in the sequence may vary arbitrarily.)

Now we define a measure of complexity of a configuration (3-1). Given a subgroup H of G, define the *dimension* $\dim H$ of H to be the quantity such that the order $|H|$ of H is the product of $\dim H$ primes (counting multiplicity). This is a natural number between 0 and d, and any proper subgroup of G has dimension at most $d-1$.

Given a configuration (3-1), define the *complexity* of the configuration to be the tuple $(m_0, \dots, m_{d-1})$, where, for each $j = 0, \dots, d-1$, m_j is the number of cosets $x_i + H_i$ in the configuration such that H_i has dimension j. Since all the H_i have dimensions between 0 and $d-1$, we see that the $m_0, \dots, m_{d-1}$ are natural numbers that sum to m. In particular, if m is constant, there are only finitely many possible complexities. Thus, by passing to a subsequence if necessary, we can find a sequence of configurations (3-1) whose complexity $(m_0, \dots, m_{d-1})$ stays constant, but k goes to infinity.

We give the space of tuples $(m_0, \dots, m_{d-1}) \in \mathbb{N}^d$ the lexicographical ordering: we write $(m_0, \dots, m_{d-1}) < (n_0, \dots, n_{d-1})$ if there exists $0 \leq i \leq d-1$ such that $m_i < n_i$ and $m_j = n_j$ for $i < j \leq d-1$. As is well-known, this makes $\mathbb{N}^d$ a well-ordered set.

Call a tuple $(m_0, \dots, m_{d-1})$ *good* if there exists a sequence of configurations (3-1) with constant complexity $(m_0, \dots, m_{d-1})$, for which k goes to infinity. We have seen that there is at least one good tuple; by the well-ordering of $\mathbb{N}^d$, we may thus find a minimal good tuple $(m_0, \dots, m_{d-1})$.

By rounding k down to an even number and then dividing by two, we may thus find a sequence of configurations

$$(2k, G, A, m, (x_i + H_i)_{i=1}^m) \tag{3-3}$$

of complexity $(m_0, \dots, m_{d-1})$ with k going to infinity.

Let d_* be the largest j for which m_j is nonzero; thus, $0 \le d_* \le d-1$ (note that at least one of the m_j must be nonzero; otherwise, the first inclusion in (3-2) could not hold). By relabeling, we may assume without loss of generality that H_1 has dimension d_* for any configuration (3-3) in the above sequence.

Consider a configuration (3-3) in the above sequence; then

$$2kA \subset \bigcup_{i=1}^{m} x_i + H_i.$$

In particular, for any $y \in kA$, we have

$$kA \subset 2kA \cap (2kA - y) \subset \bigcup_{i=1}^{m} \bigcup_{j=1}^{m} (x_i + H_i) \cap (x_j - y + H_j).$$

Note that the set $(x_i + H_i) \cap (x_j - y + H_j)$ is either empty or a coset of $H_i \cap H_j$, which has dimension at most d_*, with equality if and only if $H_i = H_j$ has dimension d_*. In particular, since all the cosets $x_j + H_j$ are assumed distinct, we see that, if H_i has dimension d_*, there is at most one set $(x_i+H_i)\cap(x_j-y+H_j)$ that is a coset of a d_*-dimensional subgroup. In particular, at most m_{d_*} of the $(x_i+H_i)\cap(x_j-y+H_j)$ arise as cosets of d_*-dimensional subgroups.

Now suppose that we can find $y \in kA$ such that

$$y \notin \bigcup_{1 \le j \le m : H_j = H_1} x_j - x_1 + H_1. \tag{3-4}$$

Then we see that $x_1 + H_1 \ne x_j - y + H_j$ for any $j = 1, \dots, m$. As such, now at most $m_{d_*} - 1$ of the $(x_i + H_i) \cap (x_j - y + H_j)$ arise as cosets of d_*-dimensional subgroups. Collecting all the cosets of the form $(x_i + H_i) \cap (x_j - y + H_j)$ and eliminating duplicates, we obtain a new configuration

$$(k, G, A, m', (x'_i + H'_i)_{i=1}^{m'}),$$

which has strictly lower complexity than $(m_0, \dots, m_{d-1})$. By the minimality of $(m_0, \dots, m_{d-1})$, this situation can only occur for finitely many of the sequence of configurations (3-3). Thus, after discarding finitely many terms, we may assume that the situation (3-4) does not occur for any $y \in kA$; that is to say, we have

$$kA \subset \bigcup_{1 \le j \le m : H_j = H_1} x_j - x_1 + H_1.$$

This gives rise to a configuration of strictly lower complexity than $(m_0, \dots, m_{d-1})$, unless $(m_0, \dots, m_{d-1}) = (0, \dots, 0, m, 0, \dots, 0)$ (with m in the d_* position), and all of the H_j are equal to H_1. Thus, after discarding finitely many terms in the sequence, we may assume that $H_j = H_1$ for all j, and so

$$kA \subset \bigcup_{j=1}^{m} x_j - x_1 + H_1.$$

Intersecting this with the inclusion $kA \subset \bigcup_{j=1}^{m} x_j + H_1$, we again obtain a configuration of lower complexity, unless the set of cosets $\{x_j + H_1 : 1 \leq j \leq m\}$ is invariant with respect to translation by x_1; so by discarding another finite number of terms in the sequence, we may assume that this is the case. By permuting indices, we can then assume that $\{x_j + H_1 : 1 \leq j \leq m\}$ is invariant under translation by x_i for any $1 \leq i \leq m$. In other words, $\{x_j + H_1 : 1 \leq j \leq m\}$ is a subgroup of the quotient group G/H_1, so $\bigcup_{j=1}^{m} x_j + H_1$ is a subgroup of G. But this has to be a proper subgroup by (3-2), and so A is in a proper subgroup of G, a contradiction.

4. Character sums from Type II sums

We now prove Theorem 1.6. Suppose that Conjecture 1.5 holds for a fixed choice of $0 < \varpi < \frac{1}{4}$. Let $\delta > 0$ be as in Conjecture 1.5; we may assume that δ is small, e.g., $\delta < \frac{1}{4}$. Let $\varepsilon > 0$ be a sufficiently small fixed quantity depending on δ. If the claim (1-8) failed, then we could find a sequence of nonprincipal primitive characters χ with conductor q going to infinity such that

$$\Bigl| \sum_{n < q^{1/2-2\varpi+\varepsilon}} \chi(n) \Bigr| \gg q^{1/2-2\varpi+\varepsilon} \log^{-A} q$$

for some fixed $A > 0$. From the pigeonhole principle, we have

$$\Bigl| \sum_{n \in [N/2, N]} \chi(n) \Bigr| \gg N \log^{-A} q \tag{4-1}$$

for some $N = q^{1/2-2\varpi+\varepsilon} \log^{-O(A)} q$ (of course, N will depend on q).

Set $x := N^{1/(1/2-2\varpi)}$ and $M := x/N$; thus,

$$N = x^{1/2-2\varpi}, \qquad M = x^{1/2+2\varpi}$$

and

$$x \geq q^{1+2\varepsilon}. \tag{4-2}$$

Let $\mathfrak{D}$ be the set of square-free natural numbers in $[(1-\log^{-10A-10} x)M, M]$ whose prime factors all lie in $[q^\varepsilon, x^\delta]$ not dividing q. Note that the number of primes dividing q may be crudely bounded by $O(\log q)$ and are thus a negligible proportion of the primes in $[q^\varepsilon, x^\delta]$. If ε is small enough, then the prime number theorem gives the cardinality bound

$$|\mathfrak{D}| \asymp M \log^{-10A-11} x. \tag{4-3}$$

(We allow implied constants to depend on the fixed quantities ε, δ, and A.)

We now set

$$\alpha(m) := 1_{\mathcal{D}}(m)\overline{\chi(m)}$$

and

$$\beta(n) := 1_{[N/2,N]}(n) \tag{4-4}$$

and consider the quantity

$$\sum_{j \le q^\varepsilon} \sum_{n \le x} \chi(n)\alpha * \beta(n + jq).$$

Shifting n by jq and using the periodicity of χ, we may write this as

$$\sum_{j \le q^\varepsilon} \sum_{jq < n \le x + jq} \chi(n)\alpha * \beta(n).$$

Since $\alpha * \beta$ is supported on $[MN/4, MN] = [x/4, x]$, this is equal (by (4-2)) to

$$\sum_{j \le q^\varepsilon} \sum_{n} \chi(n)\alpha * \beta(n),$$

which factorizes as

$$\sum_{j \le q^\varepsilon} \biggl(\sum_{m} \chi(m)\alpha(m)\biggr)\biggl(\sum_{n} \chi(n)\beta(n)\biggr),$$

and hence, by (4-1) and (4-3), we have

$$\biggl|\sum_{n \le x} \chi(n) \sum_{j \le q^\varepsilon} \alpha * \beta(n + jq)\biggr| \gg xq^\varepsilon \log^{-11A-11} x.$$

We now "disperse" the $\alpha * \beta$ factors and eliminate the χ factors by a Cauchy–Schwarz argument. Let γ denote the quantity

$$\gamma := \frac{1}{x/2} \sum_{n} \alpha * \beta(n), \tag{4-5}$$

which (since $\sum_n \beta(n) = (1 + o(1))N/2$) factorizes as

$$\gamma = \frac{1 + o(1)}{M} \sum_{m} \alpha(m). \tag{4-6}$$

In particular, from (4-3) we have

$$\gamma = O(\log^{-10A-11} x). \tag{4-7}$$

Since χ has mean 0 on intervals of length q, we have

$$\biggl|\sum_{n \le x} \chi(n) \sum_{j \le q^\varepsilon} \gamma 1_{[x/2,x]}(n + jq)\biggr| \ll \gamma q q^\varepsilon = o(xq^\varepsilon \log^{-11A-11} x)$$

and thus

$$\left|\sum_{n\le x}\chi(n)\sum_{j\le q^{\varepsilon}}(\alpha*\beta-\gamma 1_{[x/2,x]})(n+jq)\right|\gg xq^{\varepsilon}\log^{-11A-11}x.$$

Applying the Cauchy–Schwarz inequality, we conclude that

$$\sum_{n\le x}\left|\sum_{j\le q^{\varepsilon}}(\alpha*\beta-\gamma 1_{[x/2,x]})(n+jq)\right|^2\gg xq^{2\varepsilon}\log^{-22A-22}x,$$

which we rearrange (using the support of $\alpha*\beta-\gamma 1_{[x/2,x]}$ to remove the restriction $n\le x$) as

$$\left|\sum_{j,j'\le q^{\varepsilon}}\sum_n(\alpha*\beta-\gamma 1_{[x/2,x]})(n)(\alpha*\beta-\gamma 1_{[x/2,x]})(n+(j'-j)q)\right| \gg xq^{2\varepsilon}\log^{-22A-22}x. \quad (4\text{-}8)$$

From the divisor bound, we have $\alpha*\beta=x^{o(1)}$, and the inner sum

$$\sum_n(\alpha*\beta-\gamma 1_{[x/2,x]})(n)(\alpha*\beta-\gamma 1_{[x/2,x]})(n+(j'-j)q)$$

may then be crudely bounded as $x^{1+o(1)}$. From this, we may remove the diagonal contribution $j=j'$ from (4-8); by symmetry, we may then reduce to the case $j'<j$. By the pigeonhole principle, we thus have

$$\left|\sum_n(\alpha*\beta-\gamma 1_{[x/2,x]})(n)(\alpha*\beta-\gamma 1_{[x/2,x]})(n-jq)\right|\gg x\log^{-22A-22}x \quad (4\text{-}9)$$

for some $1\le j\le q^{\varepsilon}$.

Let j be as above. We have

$$\sum_n\gamma 1_{[x/2,x]}(n)\times\gamma 1_{[x/2,x]}(n-jq)=\gamma^2\frac{x}{2}+o(x\log^{-22A-22}x).$$

Also, the quantity $\alpha*\beta$ is supported in $[(1-\log^{-10A-10}x)x/2,x]$. Standard divisor sum calculations using (4-3) give

$$\sum_n|\alpha*\beta(n)|1_{[(1-O(\log^{-10A-10}x))x/2,x/2]}(n)=O(x\log^{-20A-21}x) \quad (4\text{-}10)$$

and similarly

$$\sum_n|\alpha*\beta(n)|1_{[x,x(1+O(\log^{-10A-10}x))]}(n)=O(x\log^{-20A-21}x) \quad (4\text{-}11)$$

while from (4-5) one has

$$\sum_n \alpha * \beta(n)\gamma = \gamma^2 \frac{x}{2}.$$

We conclude (using (4-7)) that

$$\sum_n \alpha * \beta(n) \times \gamma 1_{[x/2,x]}(n - jq) = \gamma^2 \frac{x}{2} + o(x \log^{-22A-22} x).$$

A similar argument gives

$$\sum_n \gamma 1_{[x/2,x]}(n) \times \alpha * \beta(n - jq) = \gamma^2 \frac{x}{2} + o(x \log^{-22A-22} x).$$

Inserting these bounds into (4-9), we conclude that, if X denotes the quantity

$$X := \sum_n \alpha * \beta(n) \alpha * \beta(n - jq), \tag{4-12}$$

then we have

$$\left| X - \gamma^2 \frac{x}{2} \right| \gg x \log^{-22A-22} x \tag{4-13}$$

for q large enough.

Now we estimate X using Type II estimates in order to contradict (4-13). Expanding out the convolution $\alpha * \beta(n)$, we have

$$X = \sum_r \alpha(r) \sum_{N/2 \le m \le N} \alpha * \beta(rm - jq)$$

or equivalently

$$X = \sum_r \alpha(r) \sum_{\substack{rN/2 - jq \le n \le rN - jq \\ n = jq\ (r)}} \alpha * \beta(n).$$

Note from the support of α that $rN/2 - jq = x/2 + O(x \log^{-10A-10} x)$ and $rN - jq = x + O(x \log^{-10A-10} x)$ if $\alpha(r)$ is nonzero. A modification of (4-10) and (4-11) then shows that

$$\sum_{\substack{rN/2 + jq \le n \le rN + jq \\ n = jq\ (r)}} \alpha * \beta(n) = \sum_{n : n = jq\ (r)} \alpha * \beta(n) + O\left(\frac{x}{r} \log^{-20A-21} x \right),$$

and thus (by (4-3)),

$$X = \sum_r \alpha(r) \sum_{n : n = jq\ (r)} \alpha * \beta(n) + o(x \log^{-22A-22} x).$$

From construction, we see that jq is coprime to every prime between x^ε and x^δ that does not divide q and is in particular coprime to r. From the Type II estimate hypothesis, we have

$$\sum_r |\alpha(r)| \left| \sum_{n:n=jq\ (r)} \alpha*\beta(n) - \frac{1}{\varphi(r)} \sum_{n:(n,r)=1} \alpha*\beta(n) \right| \ll x \log^{-A'} x$$

for any fixed $A' > 0$. We conclude that

$$X = \sum_r \frac{\alpha(r)}{\varphi(r)} \sum_{n:(n,r)=1} \alpha*\beta(n) + o(x \log^{-22A-22} x).$$

If $\alpha(r)$ is nonzero, then r is the product of $O(1)$ primes between q^ε and x^δ, and so $1/\varphi(r) = 1/r + O(q^{-\varepsilon}/r)$; the contribution of the error $O(q^{-\varepsilon}/r)$ is then $o(x \log^{-22A-22} x)$ by (4-7). Also, from standard divisor bound bounds, one has

$$\sum_{n:p|n} \alpha*\beta(n) \ll \frac{x}{p}$$

for any prime p between q^ε and x^δ, and so

$$\sum_{n:(n,r)\neq 1} \alpha*\beta(n) \ll q^{-\varepsilon} x.$$

We conclude that

$$X = \sum_r \frac{\alpha(r)}{r} \sum_n \alpha*\beta(n) + o(x \log^{-22A-22} x),$$

and hence, by (4-5), (4-6), (4-7), and the estimate $1/r = 1/M + O((\log^{-10A-10} x)/M)$ on the support of α, one has

$$X = \gamma^2 \frac{x}{2} + o(x \log^{-22A-22} x),$$

which contradicts (4-13) for x large enough. This concludes the proof of Theorem 1.6.

Remark 4.1. If we have $n(q) > x^\delta$, then the sequence α in the above argument is simply $\alpha = 1_{\mathfrak{D}}$. Thus, for the purposes of establishing Vinogradov's conjecture, it suffices to consider Type II sums when α is a sequence of the form $1_{\mathfrak{D}}$; there is also considerable flexibility in how to choose the set $\mathfrak{D}$, and other choices than the one given here are available. For similar reasons, one can relax (1-7) by moving the absolute values outside of the r summation. This leads to some further numerical improvements in the $\frac{1}{68}$ exponent in [Polymath 2014a] for the purposes of the applications to Vinogradov's conjecture; see Section 5 below.

5. A variant of the method

In this section, we sketch how to modify the arguments in Section 4 to be able to utilize distributional estimates for (components of) the divisor functions τ_k.

We start with a setup similar to that in Section 4; namely, (4-1) holds for some N (and some character χ of conductor q going off to infinity) and some fixed $A \geq 1$. We set $x := q^{1+2\varepsilon}$ for some small fixed $\varepsilon > 0$. Let $k \geq 2$ be a fixed natural number, and suppose first that $N \leq x^{1/k}$. Then the quantity $M := \lfloor x/N^k \rfloor$ is at least 1. If we set $\alpha(m) := \overline{\chi(m)} 1_{[(1-\log^{10A} x)M, M]}(m)$ and $\beta(n) := 1_{[N/2,N]}(n)$, a brief calculation similar to that in the previous section reveals that

$$\left| \sum_{j \leq q^{\varepsilon}} \sum_{n \leq x} \chi(n) \alpha * \beta^{*k}(n + jq) \right| \gg x q^{\varepsilon} \log^{-(10+k)A} x,$$

where β^{*k} denotes the Dirichlet convolution of k copies of β; one should think of β^{*k} here as a component of the divisor function $\tau_k = 1^{*k}$ defined on (1-10). We then approximate $\alpha * \beta^{*k}$ by $\gamma \psi(n/x)$, where

$$\psi(t) := \int_{t_1 \cdots t_k = t} 1_{[1/2,1]}(t_1) \cdots 1_{[1/2,1]}(t_k) \frac{dt_1 \cdots dt_{k-1}}{t_1 \cdots t_k}$$

is the multiplicative convolution of k copies of $1_{[1/2,1]}$ and

$$\gamma := \frac{1}{M(N/2)^k} \sum_n \alpha * \beta^{*k}(n).$$

A repetition of the arguments of the previous section (with $\alpha * \beta^{*(k-1)}$ playing the role of α) then shows that there is $1 \leq j \leq q^{\varepsilon}$ for which one has

$$\left| X - \gamma^2 x \int_{\mathbb{R}} \psi^2(t)\, dt \right| \gg x \log^{-(20+2k)A} x,$$

where

$$X := \sum_n \alpha * \beta^{*k}(n) \alpha * \beta^{*k}(n - jq).$$

However, a somewhat tedious calculation (similar to that in the preceding section) shows that, if one has an Elliott–Halberstam-type distributional estimate for β^{*k} on residue classes to moduli up to $MN^{k-1} \asymp q^{1+2\varepsilon}/N$, one can obtain an asymptotic of the form

$$X = \gamma^2 x \int_{\mathbb{R}} \psi^2(t)\, dt + o(x \log^{-(20+2k)A} x)$$

giving the desired contradiction. If τ_k has a level of distribution θ for some $0 < \theta < 1$, this suggests we can establish cancellation in sums such as $\sum_{n \leq N} \chi(n)$ whenever $N \leq q^{1/k}$ and $q^{1+2\varepsilon}/N \leq (N^k)^{\theta - \varepsilon}$, which suggests that N can be as low as $q^{1/(1+k\theta)+\varepsilon}$ if $\theta > 1 - 1/k$. For instance, using the well-known level of distribution $\theta = \frac{2}{3}$ for the divisor function τ_2 or for the variant $\beta * \beta$ (an old observation of Linnik and Selberg, arising from the Weil bound on Kloosterman sums), this argument gives (1-8) with $\varpi = \frac{1}{28}$ (in fact, one can replace $\log^{-A} q$ by a power savings because the Linnik–Selberg argument provides such a savings in the equidistribution estimate). Using only the elementary bound of Kloosterman [1927], one gets a level of distribution $\theta = \frac{4}{7}$, corresponding to the value $\varpi = \frac{1}{60}$, thus giving a slight improvement over the Pólya–Vinogradov bound (or even the currently best known consequence of Theorem 1.6) that requires no knowledge of the Weil conjectures.

If instead $N < q^{1/k}$, one can repeat the above analysis with the convolution $\alpha * \beta^{*k}$ replaced by $\beta_1 * \cdots * \beta_k$, where $\beta_i = 1_{[N_i/2, N_i]}$ and $N_1, \ldots, N_k \geq 1$ are quantities with $N = N_1 \geq N_2, \ldots, N_k$ and $N_1 \cdots N_k = x$. If (4-1) holds for all $N_1, \ldots, N_k$, then the above analysis again leads to a contradiction if $q^{1+2\varepsilon}/N \leq x^{\theta-\varepsilon}$, which suggests that N can be as low as $q^{1-\theta+\varepsilon}$ if $\theta \leq 1 - 1/k$. By a numerical coincidence, the best known distribution results (at $\theta = \frac{4}{7}$) on τ_3, due to Fouvry, Kowalski, and Michel, correspond to the same value of ϖ, namely $\frac{1}{28}$, as the Linnik–Selberg distribution result discussed above.

In the endpoint case $N = x^{1/k}$, α becomes trivial and the quantity X discussed above is analogous to the sum

$$\sum_{n \leq x} \tau_k(n) \tau_k(n + jq),$$

with jq being slightly smaller than x. Thus, if one were able to obtain good asymptotics for such sums (with error terms that were smaller than the main term by an arbitrary power of the logarithm), one would expect to be able to obtain bounds such as (1-8) with $q^{1/2 - 2\varpi + \varepsilon}$ replaced by a quantity slightly smaller than $q^{1/k}$. Unfortunately, asymptotics for such sums are currently only known for $k = 2$.

Acknowledgments

The author was supported by a Simons Investigator grant, the James and Carol Collins Chair, the Mathematical Analysis and Application Research Fund Endowment, and NSF grant DMS-1266164. He also thanks John Friedlander, Andrew Granville, James Maynard, Lillian Pierce, and Felipe Voloch for several useful discussions and the anonymous referee for many valuable comments and suggestions.

References

[Ankeny 1952] N. C. Ankeny, "The least quadratic non residue", *Ann. of Math.* (2) **55** (1952), 65–72. MR 13,538c Zbl 0046.04006

[Banks and Makarov 2014] W. D. Banks and K. A. Makarov, "Convolutions with probability distributions, zeros of L-functions, and the least quadratic nonresidue", preprint, 2014. arXiv 1411.2009v1

[Bateman and Diamond 2004] P. T. Bateman and H. G. Diamond, *Analytic number theory: an introductory course*, Monographs in Number Theory **1**, World Scientific, Hackensack, NJ, 2004. MR 2005h:11208 Zbl 1074.11001

[Bombieri 1965] E. Bombieri, "On the large sieve", *Mathematika* **12** (1965), 201–225. MR 33 #5590 Zbl 0136.33004

[Bombieri et al. 1986] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli", *Acta Math.* **156**:3–4 (1986), 203–251. MR 88b:11058 Zbl 0588.10042

[Bombieri et al. 1987] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli, II", *Math. Ann.* **277**:3 (1987), 361–393. MR 88f:11085 Zbl 0625.10036

[Bombieri et al. 1989] E. Bombieri, J. B. Friedlander, and H. Iwaniec, "Primes in arithmetic progressions to large moduli, III", *J. Amer. Math. Soc.* **2**:2 (1989), 215–224. MR 89m:11087 Zbl 0674.10036

[Burgess 1957] D. A. Burgess, "The distribution of quadratic residues and non-residues", *Mathematika* **4** (1957), 106–112. MR 20 #28 Zbl 0081.27101

[Burgess 1963] D. A. Burgess, "On character sums and L-series, II", *Proc. London Math. Soc.* (3) **13** (1963), 524–536. MR 26 #6133 Zbl 0123.04404

[Burgess 1986] D. A. Burgess, "The character sum estimate with $r = 3$", *J. London Math. Soc.* (2) **33**:2 (1986), 219–226. MR 87g:11098 Zbl 0593.10033

[Diamond et al. 2006] H. G. Diamond, H. L. Montgomery, and U. M. A. Vorhauer, "Beurling primes with large oscillation", *Math. Ann.* **334**:1 (2006), 1–36. MR 2006j:11131 Zbl 1207.11105

[Elliott and Halberstam 1970] P. D. T. A. Elliott and H. Halberstam, "A conjecture in prime number theory", pp. 59–72 in *Symposia mathematica IV* (Istituto Nazionale di Alta Matematica, Rome, 1968–1969), Academic Press, London, 1970. MR 43 #1943 Zbl 0238.10030

[Eskin et al. 2005] A. Eskin, S. Mozes, and H. Oh, "On uniform exponential growth for linear groups", *Invent. Math.* **160**:1 (2005), 1–30. MR 2006a:20081 Zbl 1137.20024

[Fouvry 1984] É. Fouvry, "Autour du théorème de Bombieri–Vinogradov", *Acta Math.* **152**:3–4 (1984), 219–244. MR 85m:11052 Zbl 0552.10024

[Fouvry 1985] É. Fouvry, "Sur le problème des diviseurs de Titchmarsh", *J. Reine Angew. Math.* **357** (1985), 51–76. MR 87b:11090 Zbl 0547.10039

[Fouvry and Iwaniec 1980] E. Fouvry and H. Iwaniec, "On a theorem of Bombieri–Vinogradov type", *Mathematika* **27**:2 (1980), 135–152. MR 82h:10057 Zbl 0469.10027

[Fouvry and Iwaniec 1983] E. Fouvry and H. Iwaniec, "Primes in arithmetic progressions", *Acta Arith.* **42**:2 (1983), 197–218. MR 84k:10035 Zbl 0517.10045

[Fouvry and Iwaniec 1992] É. Fouvry and H. Iwaniec, "The divisor function over arithmetic progressions", *Acta Arith.* **61**:3 (1992), 271–287. MR 93g:11089 Zbl 0764.11040

[Fouvry et al. 2014] E. Fouvry, E. Kowalski, and P. Michel, "On the exponent of distribution of the ternary divisor function", preprint, 2014. arXiv 1304.3199v2

[Friedlander and Iwaniec 2010] J. Friedlander and H. Iwaniec, *Opera de cribro*, Colloquium Publications **57**, American Mathematical Society, Providence, RI, 2010. MR 2011d:11227 Zbl 1226.11099

[Gallagher 1968] P. X. Gallagher, "Bombieri's mean value theorem", *Mathematika* **15** (1968), 1–6. MR 38 #5724 Zbl 0174.08103

[Gallagher 1970] P. X. Gallagher, "A large sieve density estimate near $\sigma = 1$", *Invent. Math.* **11** (1970), 329–339. MR 43 #4775 Zbl 0219.10048

[Goldmakher 2010] L. Goldmakher, "Character sums to smooth moduli are small", *Canad. J. Math.* **62**:5 (2010), 1099–1115. MR 2011k:11108 Zbl 1273.11125

[Graham and Ringrose 1990] S. W. Graham and C. J. Ringrose, "Lower bounds for least quadratic non-residues", pp. 269–309 in *Analytic number theory* (Allerton Park, IL, 1989), edited by B. C. Berndt et al., Progr. Math. **85**, Birkhäuser, Boston, 1990. MR 92d:11108 Zbl 0719.11006

[Granville and Soundararajan 2001] A. Granville and K. Soundararajan, "The spectrum of multiplicative functions", *Ann. of Math.* (2) **153**:2 (2001), 407–470. MR 2002g:11127 Zbl 1036.11042

[Granville and Soundararajan 2007] A. Granville and K. Soundararajan, "An uncertainty principle for arithmetic sequences", *Ann. of Math.* (2) **165**:2 (2007), 593–635. MR 2008g:11165 Zbl 1139.11040

[Granville and Soundararajan 2015] A. Granville and K. Soundararajan, "Large character sums: Burgess's theorem and zeros of L-functions", preprint, 2015. arXiv 1501.01804v1

[Iwaniec and Kowalski 2004] H. Iwaniec and E. Kowalski, *Analytic number theory*, Colloquium Publications **53**, American Mathematical Society, Providence, RI, 2004. MR 2005h:11005 Zbl 1059.11001

[Kloosterman 1927] H. D. Kloosterman, "On the representation of numbers in the form $ax^2+by^2+cz^2+dt^2$", *Acta Math.* **49**:3–4 (1927), 407–464. MR 1555249 Zbl 53.0155.01

[Linnik 1942] U. V. Linnik, "A remark on the least quadratic non-residue", *C. R.* (*Doklady*) *Acad. Sci. URSS* (*N.S.*) **36** (1942), 119–120. MR 4,189a Zbl 0063.03570

[Montgomery and Vaughan 2007] H. L. Montgomery and R. C. Vaughan, *Multiplicative number theory, I: Classical theory*, Cambridge Studies in Advanced Mathematics **97**, Cambridge University, 2007. MR 2009b:11001 Zbl 1142.11001

[Motohashi 1976] Y. Motohashi, "An induction principle for the generalization of Bombieri's prime number theorem", *Proc. Japan Acad.* **52**:6 (1976), 273–275. MR 54 #10171 Zbl 0355.10035

[Polymath 2014a] D. H. J. Polymath, "New equidistribution estimates of Zhang type", *Algebra Number Theory* **8**:9 (2014), 2067–2199. MR 3294387 Zbl 06387014

[Polymath 2014b] D. H. J. Polymath, "Variants of the Selberg sieve, and bounded intervals containing many primes", *Res. Math. Sci.* **1**:12 (2014).

[Postnikov 1956] A. G. Postnikov, "On Dirichlet L-series with the character modulus equal to the power of a prime number", *J. Indian Math. Soc.* (*N.S.*) **20** (1956), 217–226. MR 18,793a Zbl 0072.27304

[Rodosskiĭ 1956] K. A. Rodosskiĭ, "On non-residues and zeros of L-functions", *Izv. Akad. Nauk SSSR. Ser. Mat.* **20** (1956), 303–306. In Russian. MR 18,564b Zbl 0071.04501

[Vinogradov 1965] A. I. Vinogradov, "The density hypothesis for Dirichet L-series", *Izv. Akad. Nauk SSSR Ser. Mat.* **29** (1965), 903–934. In Russian. MR 33 #5579 Zbl 0128.04205

[Vinogradov 1985] I. M. Vinogradov, *Selected works*, edited by L. D. Faddeev et al., Springer, Berlin, 1985. MR 87a:01042 Zbl 0577.01049

[Wirsing 1967] E. Wirsing, "Das asymptotische Verhalten von Summen über multiplikative Funktionen, II", *Acta Math. Acad. Sci. Hungar.* **18** (1967), 411–467. MR 36 #6366 Zbl 0165.05901

[Zhang 2014] Y. Zhang, "Bounded gaps between primes", *Ann. of Math.* (2) **179**:3 (2014), 1121–1174. MR 3171761 Zbl 1290.11128

Communicated by Andrew Granville
Received 2014-10-26 Revised 2015-01-12 Accepted 2015-03-18

tao@math.ucla.edu
Department of Mathematics, University of California, Los Angeles, 405 Hilgard Avenue, Los Angeles, CA 90095-1555, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the ANT website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *ANT* are usually in English, but articles written in other languages are welcome.

Length There is no a priori limit on the length of an *ANT* article, but *ANT* considers long articles only if the significance-to-length ratio is appropriate. Very long manuscripts might be more suitable elsewhere as a memoir instead of a journal article.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use LaTeX but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Algebra & Number Theory

Volume 9 No. 4 2015