Algebra & Number Theory

Volume 14
2020
No. 6

msp

Algebra & Number Theory

msp.org/ant

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

The subscription price for 2020 is US $415/year for the electronic version, and $620/year (+$60, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber 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/

msp

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1331

Unobstructedness of Galois deformation rings associated to regular algebraic conjugate self-dual cuspidal automorphic representations

David-Alexandre Guiraud

Let F be a CM field and let $(\bar{r}_{\pi,\lambda})_\lambda$ be the compatible system of residual $\mathcal{G}_n$-valued representations of Gal_F attached to a regular algebraic conjugate self-dual cuspidal (RACSDC) automorphic representation π of $\mathrm{GL}_n(\mathbb{A})$, as studied by Clozel, Harris and Taylor (2008) and others. Under mild assumptions, we prove that the fixed-determinant universal deformation rings attached to $\bar{r}_{\pi,\lambda}$ are unobstructed for all places λ in a subset of Dirichlet density 1, continuing the investigations of Mazur, Weston and Gamzon. During the proof, we develop a general framework for proving unobstructedness (with future applications in mind) and an $R = T$-theorem, relating the universal crystalline deformation ring of $\bar{r}_{\pi,\lambda}$ and a certain unitary fixed-type Hecke algebra.

1. Introduction

This article studies unobstructedness of certain Galois deformation rings. For this introduction, let F be a number field, let k be a finite field of characteristic ℓ and fix an absolutely irreducible representation

$$\bar{\rho}\colon \mathrm{Gal}_{F,S} \to \mathrm{GL}_n(k),$$

where $S \subset \mathrm{Pl}_F$ is a finite set of places. Then assigning to a complete Noetherian local algebra A over the ring W of Witt vectors of k the set of all $\mathrm{GL}_n(A)$-valued deformations of $\bar{\rho}$ defines a functor, which is representable by a universal deformation ring $R_S(\bar{\rho})$, studied first by Mazur [1989].

If the cohomology group $H^2(\mathrm{Gal}_{F,S}, \mathrm{ad}\,\bar{\rho})$ vanishes, then $R_S(\bar{\rho})$ is easily seen to be formally smooth, i.e., isomorphic to a power series ring over W. In this sense, the group $H^2(\mathrm{Gal}_{F,S}, \mathrm{ad}\,\bar{\rho})$ can be interpreted as the obstruction to the smoothness of $R_S(\bar{\rho})$, and we say that $R_S(\bar{\rho})$ is *unobstructed* if $H^2(\mathrm{Gal}_{F,S}, \mathrm{ad}\,\bar{\rho}) = 0$.

We point out the following connection with a conjecture of Jannsen: Assume that $\bar{\rho}$ is the reduction of the ℓ-adic representation $\rho_{f,\ell}$ attached to a cuspidal modular eigenform f (see [Deligne 1973; Shimura 1971; Deligne and Serre 1974]). Then the Frobenius eigenvalues of $\rho_{f,\ell}$ are Weil-numbers of some fixed weight w, i.e., $\rho_{f,\ell}$ is pure of weight w. A conjecture of Jannsen [1989, Conjecture 1] (see also [Bellaïche 2009, Conjecture 5.1]) predicts the vanishing of $H^2(\mathrm{Gal}_{F,S}, \mathrm{ad}\,\rho)$. This implies that $H^2(\mathrm{Gal}_{F,S}, \Lambda)$

MSC2010: primary 11F80; secondary 11F70.
Keywords: Galois deformation, automorphic representation.

is finite and torsion, where $\Lambda \subset \operatorname{ad}\rho$ denotes an integral $\mathrm{Gal}_{F,S}$-stable lattice. Now our residual H^2-vanishing implies the vanishing of $H^2(\mathrm{Gal}_{F,S}, \Lambda)$ by Nakayama's lemma. This, in turn, implies the vanishing of $H^2(\mathrm{Gal}_{F,S}, \operatorname{ad}\rho)$, as predicted by Jannsen. Besides this application, numerous uses of Galois deformation-theoretic methods in number theory indicate that the structure of universal deformation rings is of independent interest.

Unobstructedness for Galois representations attached to automorphic objects rarely can be expected to hold for all choices of ℓ. The best we can hope for is that unobstructedness holds for almost all primes (or for all primes in a subset of Dirichlet density 1), and this question has been studied (under different technical assumptions) in the following cases:

(a) For $\bar\rho$ the reduction of the representation $\rho_{E,\ell}$ attached to an elliptic curve E over $F = \mathbb{Q}$; see [Mazur 1989].

(b) For $\bar\rho$ the reduction of the representation $\rho_{f,\ell}$ attached to a newform f of weight $k \geq 3$ over $F = \mathbb{Q}$; see [Weston 2004] (but see also [Yamagami 2004; Hatley 2015]).

(c) For $\bar\rho$ the reduction of the representation $\rho_{f,\ell}$ attached to a Hilbert eigenform f over a totally real field F; see [Gamzon 2016].

Note that $n = 2$ in all these cases.

For an example of $n = 3$, in [Chenevier 2011, Appendix] unobstructedness is shown under GRH for $\bar\rho = \mathrm{Sym}^2 A[\ell](-1)|_{G_{E,S}}$ where A is an elliptic curve over $\mathbb{Q}$ of one of the ten isogeny classes listed in [loc. cit., Proposition 6.15], $\ell = 5$, $E = \mathbb{Q}(i)$ and S is the set of places containing ∞, ℓ and the primes dividing $\mathrm{disc}(E).\,\mathrm{cond}(A)$; see [loc. cit., Appendix].

In this article, we develop a general framework for proving unobstructedness. To this end, we adjust the arguments in the existing literature to deal with framings, build on [Shotton 2018] to understand minimal lifts at $\ell \neq p$, and use results of [Allen 2016] together with ideas of Khare–Wintenberger to bootstrap the $R[1/p]^{\mathtt{red}} = T[1/p]$ theorems of [Barnet-Lamb et al. 2014] to obtain an $R^{\mathtt{min}} = T^{\mathtt{min}}$ theorem. We apply this framework to the reduction of the Galois representation attached to a regular algebraic conjugate self-dual cuspidal (RACSDC) automorphic representation π of $\mathrm{GL}_n(\mathbb{A}_F)$ with ramification set S, where F is a CM field.[1] To give a more precise statement, we have to recall that π gives rise, in first instance, not to GL_n-valued representations, but to morphisms $r_{\pi,\lambda}\colon \mathrm{Gal}_{F^+} \to \mathcal{G}_n(\overline{\mathbb{Q}}_{\ell(\lambda)})$, where λ runs through the places of the coefficient field of π, where $\mathcal{G}_n$ denotes the group scheme from [Clozel et al. 2008, Section 2.1] and where $\ell(\lambda)$ denotes the rational prime below λ.

We make the following assumption:

Assumption 1.1. The set of the λ for which the GL_n-valued representation $\bar r_{\pi,\lambda} \mid \mathrm{Gal}_F$ is absolutely irreducible has Dirichlet density 1.

[1] We remark that, in light of the results of [Barnet-Lamb et al. 2014], it should be possible to weaken the conjugate self-duality assumption to an essentially self-duality assumption, thus treating RACESDC automorphic representations.

We remark that this assumption is fulfilled, e.g., if $n \leq 5$ or if π is extremely regular, or would follow from absolute irreducibility of the ℓ-adic system $(r_{\pi,\lambda} \mid \mathrm{Gal}_F)$; see Remark 8.2. For the following, we fix for each λ a lift χ of the character $\mathtt{m} \circ \bar{r}_{\pi,\lambda}$ of Gal_F, where $\mathtt{m}$ is the multiplier character of the group $\mathcal{G}_n$; see Section 6A. By $R^{\chi}_{S_\ell}(\bar{r}_{\pi,\lambda})$ we denote the universal ring parametrizing deformations r of $\bar{r}_{\pi,\lambda}$ that are unramified outside the places that are in S or divide $\infty.\ell(\lambda)$ and that fulfill $\mathtt{m} \circ r = \chi$. The correct unobstructedness requirement is then the vanishing of $H^2(\mathrm{Gal}_{F,S_\ell}, \mathfrak{g}_n^{\mathtt{der}})$, where $\mathfrak{g}_n^{\mathtt{der}}$ denotes the Lie algebra of the derived subgroup of $\mathcal{G}_n$. Our main result is:

Theorem 1.2. *Assume that all Hodge–Tate weights of $r_{\pi,\lambda}$ (which are independent of λ, as the $r_{\pi,\lambda}$ form a compatible system) are nonconsecutive: if $a, b \in \mathbb{Z}$ show up as Hodge–Tate weights, then $|a-b| \neq 1$. Then, for all λ in a set of places of Dirichlet density 1 the universal deformation ring $R^{\chi}_{S_\ell}(\bar{r}_{\pi,\lambda})$ is unobstructed.*

Remark that we do not require a particular splitting behavior at the places in S. We also want to stress that the developed framework is flexible and in principle applicable to Galois representations with values in other groups and can be used to establish unobstructedness of universal deformation rings with imposed deformation conditions, which are more sophisticated than the fixed-determinant condition $\mathtt{m} \circ r = \chi$. Therefore, we hope that the framework will be useful for other applications, as better modularity lifting results become available in the future. We also remark that presently the condition on the Hodge–Tate weights is necessary for using a local unobstructedness property at the places above $\ell(\lambda)$; a technical inconvenience we expect to weaken in future work.

We give a short outline of the article: After some remarks about notation, we start in Section 3 with a collection of the general deformation theoretic methods we will use. Moreover, we will define a suitably flexible notion of unobstructedness for conditioned deformation functors (Definition 3.28). In Section 4, we state and prove the core framework (Theorem 4.2), which uses a list of six assumptions as input and provides unobstructedness as output. This framework is presented for local deformation conditions `crys, min, sm`, which have a purely formal meaning throughout Section 4. The main input is the formal smoothness of the deformation ring with respect to the conditions `min` and `crys`, which is the natural output of a suitable $R = T$-theorem, and the desired unobstructedness is then deduced by commutative algebra arguments and comparing dimensions. Section 5 introduces and studies useful local conditions that will go into the framework theorem later. After a reminder on the association of Galois deformations to automorphic forms, the additional results are provided in Section 7: We consider the deformation ring $R^{\mathtt{min},\mathtt{crys}} := R^{\chi,\mathtt{min},\mathtt{crys}}_{S_\ell}(\bar{r}_{\pi,\lambda})$ parametrizing those lifts that are minimally ramified (in the sense of Section 5D) at all places in S and crystalline (in the Fontaine–Laffaille range) at all places dividing ℓ. Moreover, we consider a corresponding Hecke algebra $\mathbb{T}^{\mathtt{min}}$ that is defined as the localization of a certain endomorphism algebra of automorphic forms of the same weight and level as π, and with a certain fixed type-requirement at the places in S. Then, using the modularity lifting results of [Barnet-Lamb et al. 2014], we show:

Theorem 1.3. *$R^{\mathtt{min},\mathtt{crys}} \cong \mathbb{T}^{\mathtt{min}}$ and, for almost all λ, $\mathbb{T}^{\mathtt{min}} \cong W$.*

This result is crucial to prove in Section 8 that, for almost all λ, there exists a suitable finite solvable extension F' of F such that the deformation ring $R^{\chi,\mathtt{min}}_{S_\ell}(\bar{r}_{\pi,\lambda} \mid \mathrm{Gal}_{F',S})$, parametrizing deformations of the base change of $\bar{r}_{\pi,\lambda}$ to F' that are minimally ramified at all places above S, is unobstructed. We go on to show that the minimally ramified condition can be waived for almost all λ (Theorem 7.10). It is important to keep track of the different field extensions necessary when running through all λ, so that we are left with a set of Dirichlet density 1 to which we can apply a result on potential unobstructedness (Lemma 4.8) and finally deduce Theorem 1.2.

2. Notation

Before we start with the main body of this article, let us make some remarks on the notation used: If F denotes a number field, we denote by Pl_F the set of places of F and by $\mathrm{Pl}_F^{\mathrm{fin}}$ the set of finite places of F. Moreover, we set $\Omega_\infty^F = \mathrm{Pl}_F \setminus \mathrm{Pl}_F^{\mathrm{fin}}$ and, for a rational prime ℓ, we denote by Ω_ℓ^F the set of places of F dividing ℓ. If F is understood, we will simply write Ω_∞ and Ω_ℓ. For a place $\lambda \in \mathrm{Pl}_F^{\mathrm{fin}}$ we define $\ell(\lambda)$ (or ℓ, if λ is understood) as the rational prime below λ. If $S \subset \mathrm{Pl}_F^{\mathrm{fin}}$ and ℓ is some rational prime, we set $S_\ell := S \cup \Omega_\infty \cup \Omega_\ell$.

We denote by $\bar{F}$ the Galois closure of F. When dealing with a quadratic extension $F \mid F^+$, we will denote by c the nontrivial element of the Galois group $\mathrm{Gal}(F \mid F^+)$. Moreover, for a rational prime ℓ, we denote by $\epsilon_\ell \colon \mathrm{Gal}_F \to \overline{\mathbb{Z}}_\ell^\times$ the ℓ-adic cyclotomic character and by $\bar{\epsilon}_\ell$ its mod-ℓ reduction.

If $L \mid F$ is a finite extension and S is a fixed set of places of F, then we will denote as well by S the set $\{\nu' \in \mathrm{Pl}_L : \nu' \text{ divides some } \nu \in S\}$. In a completely analogous way, if S is a subset of Pl_L, then we will denote as well by S the set $\{\nu' \in \mathrm{Pl}_F : \nu' \text{ is divided by some } \nu \in S\}$. If ρ is a representation of Gal_F and ν a place of F, we will use the symbol ρ_ν for the restriction of ρ to a decomposition subgroup at ν.

For a topological group Γ and a topological ring R, we denote by $\mathrm{Rep}_R(\Gamma)$ the category of finitely generated R-modules with a continuous Γ-action. If A is a Γ-module, we denote by A^* the Pontryagin dual and by $A^\vee$ the Tate dual of A.

We will often make statements concerning variations of deformation rings and we will shorten this using brackets; e.g., we will use the notation $R^{(\chi),[\mathtt{min}]}(\bar{\rho}) = 0$ as a shortcut for the four statements $R(\bar{\rho}) = 0$, $R^\chi(\bar{\rho}) = 0$, $R^{\mathtt{min}}(\bar{\rho}) = 0$ and $R^{\chi,\mathtt{min}}(\bar{\rho}) = 0$. For cohomology groups, we abbreviate $h^i(*,*)$ for $\dim H^i(*,*)$.

Let k be a finite field of characteristic ℓ. For the valuation ring Λ of a finite extension of $\mathbb{Q}_\ell$ with residue field $k_\Lambda = k$, we will consider the category $\mathcal{C}_\Lambda$ of complete Noetherian local Λ-algebras A fulfilling $k_A = k$.

3. Liftings and deformations

In this section, which contains nothing original, we recall the main results on deformation theory. For general background literature, we refer the reader to [Tilouine 1996; Mauger 2000; Levin 2013; Balaji 2012; Bleher and Chinburg 2003]. Let us first fix a finite field k and denote $\ell = \mathrm{char}(k)$. We will denote

the ring of Witt vectors over k by $W(k)$, or, if k is understood, by W. Moreover, let us fix a profinite group Γ which fulfills the ℓ-finiteness condition (Φ_ℓ) of [Mazur 1989]: For any open subgroup $H \subset \Gamma$, the maximal pro-ℓ quotient of H is topologically finitely generated.

Let G be a smooth linear algebraic group over W and fix a continuous group homomorphism $\bar{\rho}\colon \Gamma \to G(k)$, where $G(k)$ carries the discrete topology.

Basic facts on coefficient rings. Let us first state some basic facts on the category $\mathcal{C}_\Lambda$, whose proofs we leave to the reader: The pushout in $\mathcal{C}_\Lambda$ is realized by the completed tensor product $\widehat{\otimes}$; see [Mazur 1997, Section 12]. Consequently, if $C \leftarrow A \to B$ is a diagram in $\mathcal{C}_\Lambda$, then $\mathrm{Hom}_{\mathcal{C}_\Lambda}(B\widehat{\otimes}_A C, \_)$ is the pullback of the diagram of functors $\mathrm{Hom}_{\mathcal{C}_\Lambda}(C,\_) \to \mathrm{Hom}_{\mathcal{C}_\Lambda}(A,\_) \leftarrow \mathrm{Hom}_{\mathcal{C}_\Lambda}(B,\_)$. Consider a pushout diagram in $\mathcal{C}_\Lambda$ where one arrow (say, f) is surjective. This implies that the parallel arrow (say, g) is surjective as well, so taking $I = \ker(f)$ and $J = \ker(g)$ we can extend the orthogonal arrow (say, π) to a map of short exact sequences of Λ-modules:

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ {\scriptstyle\pi}\downarrow & & \vdots \\ C & \xrightarrow[g]{} & P \end{array} \quad \rightsquigarrow \quad \begin{array}{ccccccccc} 0 & \longrightarrow & I & \longrightarrow & A & \xrightarrow{f} & B & \longrightarrow & 0 \\ & & {\scriptstyle\pi\,|\,I}\downarrow & & {\scriptstyle\pi}\downarrow & & \downarrow & & \\ 0 & \longrightarrow & J & \longrightarrow & C & \xrightarrow[g]{} & P & \longrightarrow & 0 \end{array}$$

If $\mathfrak{I}$ is an ideal of some $D \in \mathcal{C}_\Lambda$ we denote cardinality of a minimal set of generators of $\mathfrak{I}$ by $\mathrm{gen}_D(\mathfrak{I}) := \dim_k \mathfrak{I}/\mathfrak{m}_D\mathfrak{I}$. Then, we easily see that the following holds for the above diagram:

Proposition 3.1. *In the above diagram,* $\mathrm{gen}_C(J) \le \mathrm{gen}_A(I)$.

Proof. This follows from the above extended diagram, using that both the map $I \to C\hat{\otimes}_A I$ induced by base change from A to C and the surjective module homomorphism $C\hat{\otimes}_A I \twoheadrightarrow J$ send systems of generators to systems of generators. □

Recall the following elementary facts about regular systems of parameters:

Proposition 3.2 [Serre 2000, Proposition 22 and the subsequent corollary]. (a) *Let $x_1, \dots, x_l$ be l elements of the maximal ideal $\mathfrak{m}_A$ of a regular local ring A. Then the following are equivalent*:

(i) *$x_1, \dots, x_l$ is a subset of a regular system of parameters of A.*

(ii) *The images of $x_1, \dots, x_l$ in $\mathfrak{m}_A/\mathfrak{m}_A^2$ are linearly independent over k.*

(iii) *The local ring $A/(x_1, \dots, x_l)$ is regular and has dimension* $\dim A - l$. (*In particular, $(x_1, \dots, x_l)$ is a prime ideal.*)

(b) *If $\mathfrak{I}$ is an ideal of a regular local ring A, the following properties are equivalent*:

(i) *$A/\mathfrak{I}$ is a regular local ring.*

(ii) *$\mathfrak{I}$ is generated by a subset of a regular system of parameters of A.*

Moreover, we have the following results, which follow easily from standard facts about regular systems of parameters (see [Serre 2000, Proposition 22] and its use in Section 2 of [Guiraud 2016]):

Lemma 3.3. *Suppose $A = \Lambda[\![x_1, \dots, x_a]\!]$, $B = \Lambda[\![x_1, \dots, x_b]\!] \in \mathcal{C}_\Lambda$ and let $J \subset A$ be an ideal of the form $J = (f_1, \dots, f_u)$ with $f_i \in A$ and $u \leq a$. Suppose moreover that there exists a surjective morphism $f : A/J \twoheadrightarrow B$ and denote its kernel by I. Then the following are equivalent*:

- $A/J \cong \Lambda[\![x_1, \dots, x_{a-u}]\!]$.
- $\mathrm{gen}_{A/J}(I) = a - u - b$.
- $\mathrm{gen}_{A/J}(I) \leq a - u - b$.

Proof. It is clear that there cannot be a negative number of generators of I. By Proposition 3.2(b), the ideal I can be generated by a subset (of, say, cardinality r) of a regular system of parameters of A. By part (a) of said proposition, the quotient A/I has dimension $\dim A - r = a + 1 - r$. We get $r = a - b$, which is thus an upper bound on $\mathrm{gen}(I)$. In order to derive a lower bound, consider the canonical surjection

$$\pi : A/\mathfrak{m}_A.I \twoheadrightarrow A/\mathfrak{m}_A^2.$$

The image of $I/\mathfrak{m}_A.I$ under π is $(I + \mathfrak{m}_A^2)/\mathfrak{m}_A^2 \cong I/(I \cap \mathfrak{m}_A^2)$. This implies $\mathrm{gen}(I) = \dim_k I/\mathfrak{m}_A I \geq \dim_k I/I \cap \mathfrak{m}_A^2 = r$, where the last equality is taken from the proof of [Serre 2000, Proposition 22]. □

Proposition 3.4. *Let $m \in \mathbb{N}$. Then $A \in \mathcal{C}_\Lambda$ is regular if and only if $A[\![x_1, \dots, x_m]\!]$ is regular.*

Proof. It is clearly sufficient to consider the case $m = 1$. The "only if" part is [Matsumura 1970, Proposition 24D]. For the other direction, assume that $A[\![x]\!]$ is regular. It is clear that x is not contained in $\mathfrak{m}_{A[\![x]\!]}^2 = (\mathfrak{m}_A, x)^2$, so implication (ii) ⇒ (iii) of Proposition 3.2(a) yields regularity of $A[\![x]\!]/(x) \cong A$. □

Proposition 3.5. *Let $f : A \to B$ be a morphism in $\mathcal{C}_\Lambda$. Then f is formally smooth* (see [EGA IV$_1$ 1964, Section 19]) *if and only if B is isomorphic to a formal power series ring over A.*

Proof. This is the equivalence (i)⇔(ii) of [Sernesi 2006, Proposition C.6]. □

Lemma 3.6. *Let $A \in \mathcal{C}_\Lambda$, $m \in \mathbb{N}$ such that $\Lambda[\![x_1, \dots, x_m]\!] \cong A \widehat{\otimes}_\Lambda \Lambda[\![x]\!]$. Then $A \cong \Lambda[\![x_1, \dots, x_{m-1}]\!]$.*

Proof. Let ϖ be a uniformizing element of Λ. Clearly, the unknown

$$x \in (R/\varpi.R)[\![x]\!] \cong k[\![x_1, \dots, x_m]\!]$$

is contained in a regular system of parameters, so

$$R/\varpi.R \cong k[\![x_1, \dots, x_{m-1}]\!]. \tag{1}$$

Now consider the diagram

$$\begin{array}{ccc} \Lambda & \longrightarrow & \Lambda[\![x_1, \dots, x_{m-1}]\!] \\ \downarrow & \overset{\tilde{h}}{\nearrow} & \downarrow h \\ R & \xrightarrow[g]{} & R/\varpi.R \end{array}$$

where h and g are the projection maps modulo ϖ. As $\Lambda[\![x_1, \dots, x_{m-1}]\!]$ is formally smooth over Λ, there exists a dotted map $\tilde{h}$. Because of the isomorphism (1), R modulo the maximal ideal of $\Lambda[\![x_1, \dots, x_{m-1}]\!]$ is k and hence, by Nakayama's lemma, the map $\tilde{h}$ is surjective.

Now we see that $\tilde{h}$ must be an isomorphism: Assume, this is not the case. Then $\dim R < m$, which is in conflict with the isomorphism $\Lambda[\![x_1, \dots, x_m]\!] \cong R\hat{\otimes}_\Lambda \Lambda[\![x]\!] \cong R[\![x]\!]$. □

Lemma 3.7. *Let $\Delta \in \mathcal{C}_\Lambda$ such that the structure morphism $\Lambda \to \Delta$ is flat and*

$$R = \Lambda[\![x_1, \dots, x_d]\!]/(f_1, \dots, f_b) \tag{2}$$

for some $f_1, \dots, f_b$ in $\Lambda[\![x_1, \dots, x_d]\!]$. Then R is formally smooth of relative dimension $d \in \mathbb{N}$ over Λ if and only if $\Delta\widehat{\otimes}_\Lambda R$ is formally smooth of relative dimension d over Δ.

Proof. Let $I := (f_1, \dots, f_b)$. Let $\mathfrak{m}$ denote the maximal ideal of R and $b = \dim I/\mathfrak{m}.I$, and let $\mathfrak{m}'$ denote the maximal ideal of $R' := \Delta\widehat{\otimes}_\Lambda R$ and $b' = \dim I/\mathfrak{m}'.I$.

To see that b equals $b' := \dim_{\Delta/\mathfrak{m}'} \Delta \otimes_\Lambda I/\mathfrak{m}'.\Delta \otimes_\Lambda I$, we use the isomorphism

$$\Delta \otimes_\Lambda I/\mathfrak{m}'.\Delta \otimes_\Lambda I \cong I/\mathfrak{m}.I \otimes_{\Lambda/\mathfrak{m}} \Delta/\mathfrak{m}'$$

and the fact that $\Lambda/\mathfrak{m} \to \Delta/\mathfrak{m}'$ is a monomorphism of fields:

$$b = \dim_{\Lambda/\mathfrak{m}} I/\mathfrak{m}.I = \dim_{\Delta/\mathfrak{m}'} \Delta \otimes_\Lambda I/\mathfrak{m}'.\Delta \otimes_\Lambda I = b'.$$ □

Liftings and deformations of G-valued representations.

Definition 3.8. (1) A lifting of $\bar{\rho}$ to an object $A \in \mathcal{C}_\Lambda$ is a continuous group homomorphism $\rho\colon \Gamma \to G(A)$ fulfilling $\mathrm{mod}_{\mathfrak{m}_A} \circ \rho = \bar{\rho}$, where $\mathrm{mod}_{\mathfrak{m}_A}\colon G(A) \to G(A/\mathfrak{m}_A) = G(k)$ is the canonical reduction.

(2) Denote by $D^\square_\Lambda(\bar{\rho})\colon \mathcal{C}_\Lambda \to$ Sets the functor which assigns to an object $A \in \mathcal{C}_\Lambda$ the set of all liftings of $\bar{\rho}$ to A.

By [Balaji 2012, Theorem 1.2.2], $D^\square_\Lambda(\bar{\rho})$ is representable by an object $R^\square_\Lambda(\bar{\rho}) \in \mathcal{C}_\Lambda$. As an examination of its proof easily yields, we get (with respect to the ring of integers Λ' of some finite extension of $\mathrm{Quot}(\Lambda)$ with residue field $k_{\Lambda'} = k$) an isomorphism

$$R^\square_{\Lambda'}(\bar{\rho}) \cong \Lambda'\widehat{\otimes}_\Lambda R^\square_\Lambda(\bar{\rho}). \tag{3}$$

Definition 3.9. A lifting condition is a family $\mathcal{D} = (S(A))_{A\in\mathcal{C}_\Lambda}$ of subsets $S(A) \subset D^\square_\Lambda(\bar{\rho})(A)$ such that:

(1) $\bar{\rho} \in S(k)$.

(2) If $f\colon A \to B$ is a morphism in $\mathcal{C}_\Lambda$ and $\rho \in S(A)$, then $G(f) \circ \rho \in S(B)$.

(3) Let $f_1\colon A_1 \to A$, $f_2\colon A_2 \to A$ be morphisms in $\mathcal{C}_\Lambda$ and let ρ_3 be a lifting of $\bar{\rho}$ to $A_3 := A_1 \times_A A_2$. For $i = 1, 2$ denote by $\pi_i : A_3 \to A_i$ the canonical map and by ρ_i the lifting $G(\pi_i) \circ \rho_3$ of $\bar{\rho}$ to A_i. Then, $\rho_3 \in S(A_3)$ if and only if $\rho_1 \in S(A_1)$ and $\rho_2 \in S(A_2)$.

Condition (2) guarantees that $\mathcal{D}$ defines a subfunctor $D^{\square,\mathcal{D}}_\Lambda(\bar{\rho}) \subset D^\square_\Lambda(\bar{\rho})$. Condition (3) is a variation of the Mayer–Vietoris property, so a standard argument yields:

Proposition 3.10. *$D_\Lambda^{\square,\mathcal{D}}(\bar\rho)$ is a relatively representable subfunctor (in the sense of* [Mazur 1997, Section 19]*) of $D_\Lambda^{\square}(\bar\rho)$, i.e., representable by some $R_\Lambda^{\square,\mathcal{D}}(\bar\rho) \in \mathcal{C}_\Lambda$. On the other hand, any representable subfunctor $F \subset D_\Lambda^{\square}(\bar\rho)$ yields a lifting condition $\mathcal{D} = (S(A))_{A\in\mathcal{C}_\Lambda}$ via $S(A) := F(A)$.*

We have the following conditioned version of (3):

$$R_{\Lambda'}^{\square,\mathcal{D}'}(\bar\rho) \cong \Lambda'\widehat{\otimes}_\Lambda R_\Lambda^{\square,\mathcal{D}}(\bar\rho), \tag{4}$$

where the condition $\mathcal{D}'$ on the left is a truncated version of $\mathcal{D}$, i.e., denotes the family of those $S(A)$ as in the definition of $\mathcal{D}$ for which $A \in \mathcal{C}_{\Lambda'}$. We will often omit this distinction and write $\mathcal{D}$ in place of $\mathcal{D}'$.

Remark 3.11. Let Λ be as above and let ${}^*\mathcal{C}_\Lambda$ denote the category of complete Noetherian local Λ-algebras A such that $[k_A : k]$ is finite. Then one can extend $D_\Lambda^{\square}(\bar\rho)$ to a functor on ${}^*\mathcal{C}_\Lambda$ by considering A-valued liftings of $\bar\rho$ as continuous group homomorphisms $\rho : \Gamma \to G(A)$ which fulfill $\mathrm{mod}_{\mathfrak{m}_A} \circ \rho = \iota_{k\subset k_A} \circ \bar\rho$, where $\iota_{k\subset k_A} : G(k) \to G(k_A)$ is the map induced by the structure map $\Lambda \to A$. It is easy to check that this extended functor is representable by the same universal object $R_\Lambda^{\square}(\bar\rho)$ as the functor from Definition 3.8. Moreover, if Λ' is the ring of integers of some finite extension of $\mathrm{Quot}(\Lambda)$ such that $[k_{\Lambda'} : k] < \infty$, we have the following version of (3):

$$R_{\Lambda'}^{\square}(\iota_{k\subset k_A} \circ \bar\rho) \cong \Lambda'\widehat{\otimes}_\Lambda R_\Lambda^{\square}(\bar\rho).$$

Moreover, if $\mathcal{D}$ is an extended lifting condition, i.e., a family $(S(A))_{A\in{}^*\mathcal{C}_\Lambda}$ fulfilling the analogue conditions of Definition 3.9 (with $A, A_i, B \in {}^*\mathcal{C}_\Lambda$), we have the following conditioned version of (4):

$$R_{\Lambda'}^{\square,\mathcal{D}}(\iota_{k\subset k_A} \circ \bar\rho) \cong \Lambda'\widehat{\otimes}_\Lambda R_\Lambda^{\square,\mathcal{D}}(\bar\rho),$$

where $\mathcal{D}$ on the left hand side is to be understood as the Λ'-truncated version of the condition $\mathcal{D}$, i.e., a family indexed by ${}^*\mathcal{C}_{\Lambda'}$ instead of ${}^*\mathcal{C}_\Lambda$. Moreover, the statement of Lemma 3.7 holds if Λ' is in ${}^*\mathcal{C}_\Lambda$ instead of $\mathcal{C}_\Lambda$. (The content of this remark is strongly inspired by the treatment in [Conrad et al. 1999, Appendix A] and [Mazur 1997].)

Definition 3.12. (1) A deformation of $\bar\rho$ to $A \in \mathcal{C}_\Lambda$ is an equivalence class of liftings to A, where two lifts are taken to be equivalent if they are conjugate by some element of $\hat G(A) := \ker(\mathrm{mod}_{\mathfrak{m}_A})$.

(2) Denote by $D_\Lambda(\bar\rho) : \mathcal{C}_\Lambda \to$ Sets the functor which assigns to an object $A \in \mathcal{C}_\Lambda$ the set of all deformations of $\bar\rho$ to A.

For the following, denote by Z_G the center of G and by $\mathfrak{g}$ (resp. by $\mathfrak{z}$) the Lie algebra of the special fiber of G (resp. of Z_G). We assume from now on that Z_G is formally smooth over Λ.

Theorem 3.13 [Tilouine 1996, Theorem 3.3]. *If $H^0(\Gamma, \mathfrak{g}) = \mathfrak{z}$ then $D_\Lambda(\bar\rho)$ is representable by an object $R_\Lambda(\bar\rho) \in \mathcal{C}_\Lambda$.*

Observe that in the case $G = \mathrm{GL}_n$, the condition of Theorem 3.13 becomes the usual centralizer condition $\mathrm{End}_{k[\Gamma]}(\bar\rho) = k$. In practice, this is often deduced from absolute irreducibility of $\bar\rho$ by Schur's lemma. This reasoning can be adopted to more general groups G as follows:

Definition 3.14 (Absolute irreducibility, see [Serre 1998]). We say that $\bar{\rho}$ is absolutely irreducible if there does not exist a proper parabolic subgroup $P \subsetneq G$ over $\bar{k}$ such that $\bar{\rho}(\Gamma) \subset P$.

Then the following can be deduced from [Bate et al. 2005, Proposition 2.13]:

Lemma 3.15 (Schur's lemma). *Assume that ℓ is very good for G (see [Bate et al. 2010, Section 2]) or that there exists an embedding $G \hookrightarrow \mathrm{GL}(V)$ such that $(\mathrm{GL}(V), G)$ is a reductive pair (in the sense of [Bate et al. 2005, Definition 3.32]). Then $H^0(\Gamma, \mathfrak{g}) = \mathfrak{z}$ if $\bar{\rho}$ is absolutely irreducible.*

We now give an appropriate version of Definition 3.9:

Definition 3.16. A deformation condition is a lifting condition in the sense of Definition 3.9 which fulfills additionally:

(4) If $\rho \in S(A)$ and $g \in \hat{G}(A)$, then $g\rho g^{-1} \in S(A)$.

This defines a relatively representable subfunctor $D_\Lambda^{\mathcal{D}}(\bar{\rho})$ of $D_\Lambda(\bar{\rho})$: If $D_\Lambda(\bar{\rho})$ is representable, then so is $D_\Lambda^{\mathcal{D}}(\bar{\rho})$ and the representing object $R_\Lambda^{\mathcal{D}}(\bar{\rho})$ is a quotient of $R_\Lambda(\bar{\rho})$. In addition to the conditions appearing in Section 5 below, we will be interested in the following conditions:

(1) If $\Delta \subset \Gamma$ is a profinite subgroup and $\bar{\rho}(\Delta) = \{1\}$, then the assignment $S(A) := \{\rho \mid \rho(\Delta) = \{1\}\}$ defines a deformation condition. In the case $\Gamma = \mathrm{Gal}_K$ for a local field K and $\Delta = I_K$, we call this the unramified lifting condition and write $D_\Lambda^{(\square),\mathtt{nr}}(\bar{\rho})$ for the corresponding subfunctor.

(2) Fix a representation $\chi : \Gamma \to G^{\mathtt{ab}}(\Lambda)$ such that $d(k) \circ \bar{\rho} = \bar{\chi}$, where $d\colon G \to G^{\mathtt{ab}}$ is the canonical projection modulo the derived subgroup $G^{\mathtt{der}}$ and where $\bar{\chi}$ denotes the reduction of χ. In accordance with the case $G = \mathrm{GL}_n$, we call this the fixed determinant condition and write $D_\Lambda^{(\square),\chi}(\bar{\rho})$ for the corresponding subfunctor.

(3) Let F be a number field and let $\Sigma \subset S \subset \mathrm{Pl}_F$ be a finite set of finite places. Let $\Gamma = \mathrm{Gal}_{F,S}$ be the Galois group of the maximal unramified outside S extension F_S of F, and fix for each $v \in \Sigma$ a local condition D_v of the functor $D_\Lambda^{(\square)}(\bar{\rho}_v)$, where $\bar{\rho}_v$ denotes the restriction of $\bar{\rho}$ to a decomposition group at v. Then the assignment $S(A) = \{\rho \mid \rho_v \in D_\Lambda^{(\square),D_v}(\bar{\rho}_v) \forall v \in \Sigma\}$ defines a global deformation condition, denoted by $\mathcal{D} = (D_v)_{v\in\Sigma}$. The afforded subfunctor of $D_\Lambda^{(\square)}(\bar{\rho})$ is denoted by $D_\Lambda^{(\square),\mathcal{D}}(\bar{\rho})$.

(4) If Γ, F, Σ are as above and if $\bar{\rho}$ is unramified outside Σ, then requiring that a lift ρ is unramified outside Σ defines a global deformation condition, and we denote the corresponding subfunctor by $D_{\Sigma,\Lambda}^{(\square)}(\bar{\rho})$. It is easily seen that studying these lifts is equivalent to studying unconditioned lifts of $\bar{\rho}$, understood as a representation of the Galois group $\mathrm{Gal}_{F,\Sigma}$ of the maximal, unramified outside Σ, extension F_Σ of F.

It is easily seen that decreeing multiple conditions defines another condition, i.e., it makes sense to write for example $D_\Lambda^{\square,\chi,\mathtt{nr}}(\bar{\rho})$.

Multiply framed deformations. Fix finite subsets $\Sigma \subset S \subset \mathrm{Pl}_F$ such that $\bar\rho$ is unramified outside S and continue to denote $\Gamma = \mathrm{Gal}_{F,S}$.

Definition 3.17. Following [Khare and Wintenberger 2009b, Section 4.1.1], we define the functor $D_\Lambda^{\square_\Sigma}(\bar\rho) : \mathcal{C}_\Lambda \to \mathrm{Sets}$ by mapping A to

$$\{(\rho, (\rho_v, \beta_v)_{v\in\Sigma}) \mid \rho \in D_\Lambda^{\square}(\bar\rho)(A),\ \rho_v \in D_\Lambda^{\square}(\bar\rho_v)(A),\ \beta_v \in \hat G(A) \text{ such that } \rho \mid \mathrm{Gal}(F_v) = \beta_v \rho_v \beta_v^{-1}\} \Big/ \sim$$

where $(\rho, (\rho_v, \beta_v)_{v\in\Sigma})$ and $(\rho', (\rho'_v, \beta'_v)_{v\in\Sigma})$ are taken to be equivalent if $\rho_v = \rho'_v$ for all v and if there is a $\gamma \in \hat G(A)$ such that $\rho' = \gamma\rho\gamma^{-1}$ and $\beta'_v = \gamma\beta_v$ for all v.

Note that specifying the ρ_v is not strictly necessary, as they can be obtained from ρ and β_v. We can impose a deformation condition $\mathcal{D} = (S(A))_{A\in\mathcal{C}_\Lambda}$ on multiply framed deformations in the same way we did for liftings and deformations, i.e., we allow only those triples $(\rho, (\rho_v, \beta_v)_{v\in\Sigma})$ for which $\rho \in S(A)$. The following assertions are immediate; see [Khare and Wintenberger 2009b, Proposition 4.1] or [Guiraud 2016, Proposition 2.62]:

Proposition 3.18. (1) *$D_{[S],\Lambda}^{\square_\Sigma,(\chi),\mathcal{D}}$ is representable and we denote the afforded deformation ring by $R_{[S],\Lambda}^{\square_\Sigma,(\chi),\mathcal{D}}$ (if $\Sigma = \varnothing$, we have to assume $H^0(\Gamma, \mathfrak{g}) = \mathfrak{z}$).*

(2) *If $\#\Sigma = 1$, then the functors $D_{[S],\Lambda}^{\square_\Sigma,(\chi),\mathcal{D}}$ and $D_{[S],\Lambda}^{\square,(\chi),\mathcal{D}}$ are naturally isomorphic.*

(3) *If $\Sigma \neq \varnothing$, then*

$$R_{[S],\Lambda}^{\square_\Sigma,(\chi),\mathcal{D}} \cong R_{[S],\Lambda}^{\square,(\chi),\mathcal{D}}[\![x_1, \dots, x_t]\!]$$

and, if $H^0(\Gamma, \mathfrak{g}) = \mathfrak{z}$, then also

$$R_{[S],\Lambda}^{\square,(\chi),\mathcal{D}} \cong R_{[S],\Lambda}^{(\chi),\mathcal{D}}[\![x_1, \dots, x_u]\!]$$

with $t = \dim(\mathfrak{g}).(\#\Sigma - 1)$, $u = \dim(\mathfrak{g}) - \dim(\mathfrak{z}) = \dim(\mathfrak{g}^{\mathtt{der}})$.

From now on, let us suppose

Assumption 3.19. $H^0(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der}}) = 0$.

With respect to a deformation condition $\mathcal{D} = (D_v)_{v\in\Sigma}$ as in Example (3) above, we set

$$R_\Lambda^{\mathtt{loc}_\Sigma,(\chi),\mathcal{D}}(\bar\rho) := \widehat{\bigotimes}_{v\in\Sigma} R_\Lambda^{\square,(\chi_v),D_v}(\bar\rho_v).$$

The following is essentially a special case of [Balaji 2012, Proposition 4.2.5] (which goes back to [Kisin 2007, Proposition 4.1.5]):

Proposition 3.20. *If Σ contains all infinite places, $H^0(\Gamma, \mathfrak{g}^{\mathtt{der},\wedge}) = 0$, and $D_\Lambda^{(\chi)}(\bar\rho)$ is representable, then*

$$R_{S,\Lambda}^{\square_\Sigma,(\chi),\mathcal{D}}(\bar\rho) \cong R_\Lambda^{\mathtt{loc}_\Sigma,(\chi),\mathcal{D}}(\bar\rho)[\![x_1, \dots, x_{a+b}]\!]/(f_1, \dots, f_a)$$

for suitable $a \in \mathbb{N}$, $f_i \in R_\Lambda^{\mathtt{loc}_\Sigma,(\chi),\mathcal{D}}(\bar\rho)[\![x_1, \dots, x_{a+b}]\!]$ and with $b = 0$ if the determinant is not fixed (resp. $b = (\#\Sigma - 1).\dim(\mathfrak{g}^{\mathtt{ab}})$ if the determinant is fixed).

Corollary 3.21. *Assume that each $R^{\square,(\chi_v),D_v}(\bar\rho_v)$ is a complete intersection ring of relative dimension d_v over Λ. Assume moreover that $D^{\mathcal{D}}_{[S],\Lambda}(\bar\rho)$ is representable and that $d := \sum_{v\in\Sigma} d_v > \dim(\mathfrak{g}).\#\Sigma - \dim(\mathfrak{z}) - b$ (with b as in Proposition 3.20). Then there exists a presentation*

$$R^{\mathcal{D}}_{[S],\Lambda}(\bar\rho) \cong \Lambda[\![x_1,\dots,x_m]\!]/(f_1,\dots,f_m)$$

for a suitable $m \in \mathbb{N}$.

Proof. Using Proposition 3.20 and the assumption on $\mathcal{D}$, we can write

$$R^{\square_\Sigma,(\chi),\mathcal{D}}_{S,\Lambda}(\bar\rho) \cong R^{\mathtt{loc}_\Sigma,(\chi),\mathcal{D}}_{\Lambda}(\bar\rho)[\![x_1,\dots,x_{a+b}]\!]/(f_1,\dots,f_a) \cong \Lambda[\![x_1,\dots,x_{a+b+c+d}]\!]/(f_1,\dots,f_{a+c})$$

for a, b as above and for a suitably chosen $c \in \mathbb{N}_0$. On the other hand, by Cohen's structure theorem we can write $R^{(\chi),\mathcal{D}}_{S,\Lambda}(\bar\rho) \cong \Lambda[\![x_1,\dots,x_u]\!]/(f_1,\dots,f_v)$ for suitable $u, v \in \mathbb{N}_0$ (and we assume that this is a minimal presentation, i.e., that the quantity $u - v$ is maximal among all ways to write $R^{(\chi),\mathcal{D}}_{S,\Lambda}(\bar\rho)$ as a quotient of a power series ring), so by the third part of Proposition 3.18 we have

$$R^{\square_\Sigma,(\chi),\mathcal{D}}_{S,\Lambda}(\bar\rho) \cong R^{(\chi),\mathcal{D}}_{S,\Lambda}(\bar\rho)[\![x_1,\dots,x_r]\!] \cong \Lambda[\![x_1,\dots,x_{r+u}]\!]/(f_1,\dots,f_v)$$

with $r = \dim(\mathfrak{g}).\#\Sigma - \dim(\mathfrak{z})$. Comparing these two presentations, we get

$$u - v + \dim(\mathfrak{g}).\#\Sigma - \dim(\mathfrak{z}) \geq b + d \Rightarrow u - v \geq b + d - \dim(\mathfrak{g}).\#\Sigma + \dim(\mathfrak{z}).$$

Thus, the claim follows immediately from our assumption on d. □

Tangent spaces and systems of local conditions. With respect to a deformation condition $\mathcal{D}$ will consider the tangent space $t_{D^{(\square)}} = D^{(\square),\mathcal{D}}_{\Lambda}(k[\epsilon])$, which we consider as a (finite-dimensional) k-vector space (see [Gouvêa 2001, Lecture 2]). There are canonical isomorphisms

$$t_{D^{(\chi)}} \cong Z^1(\Gamma,\mathfrak{g}^{(\mathtt{der})}),\ t_D \cong H^1(\Gamma,\mathfrak{g}) \quad \text{and} \quad t_{D^\chi} \cong H^1(\Gamma,\mathfrak{g}^{\mathtt{der}})' := \mathrm{im}(H^1(\Gamma,\mathfrak{g}^{\mathtt{der}}) \to H^1(\Gamma,\mathfrak{g})),$$

so via the embedding $D^{(\chi),\mathcal{D}}_{\Lambda}(k[\epsilon]) \hookrightarrow D^{(\chi)}_{\Lambda}(k[\epsilon])$ we are provided with an assignment $\mathcal{D} \mapsto L(\mathcal{D})^{(\chi)} := D^{(\chi),\mathcal{D}}_{\Lambda}(k[\epsilon])$ from deformation conditions to subspaces of $H^1(\Gamma,\mathfrak{g})$ (resp. $H^1(\Gamma,\mathfrak{g}^{\mathtt{der}})'$). In the case $\Gamma = \mathrm{Gal}_F$ for a number field F and if $\mathcal{D} = (D_v)_{v\in\Sigma}$, we call the afforded family $\mathcal{L}^{(\chi)} = (L(D_v))_{v\in\mathrm{Pl}_F}$ of subspaces of $H^1(\mathrm{Gal}_{F_v},\mathfrak{g})$ (resp. of $H^1(\mathrm{Gal}_{F_v},\mathfrak{g}^{\mathtt{der}})'$) a *system of local conditions*. Also note that there is an exact sequence

$$0 \to \mathfrak{g}/\mathfrak{g}^\Gamma \to t_{D^{(\chi)}} \to t_{D^{(\chi)}}$$

where, in case $\ell \gg 0$ (such that $\mathfrak{g} = \mathfrak{g}^{\mathtt{der}} \oplus \mathfrak{g}^{\mathtt{ab}}$), the object $\mathfrak{g}/\mathfrak{g}^\Gamma$ can be replaced by $\mathfrak{g}^{\mathtt{der}}/(\mathfrak{g}^{\mathtt{der}})^\Gamma$.

Liftings at infinity.

Proposition 3.22. *Assume $\Gamma = \mathbb{Z}/2\mathbb{Z} = \{1, c\}$ and $\ell = \mathrm{char}(\mathbb{F}) \neq 2$. Then*

$$R^{\square}_{\Lambda}(\bar\rho) \cong \Lambda[\![x_1,\dots,x_m]\!] \quad \textit{with } m = \dim(\mathfrak{g}^{c=-1}).$$

If ψ is a lift of the determinant, then the same result holds for $R^{\square,\psi}_{\Lambda}(\bar\rho)$ after replacing $\mathfrak{g}$ by $\mathfrak{g}^{\mathtt{der}}$.

Proof. We use the general formula $H^2(\mathbb{Z}/n\mathbb{Z}, M) = M^{\mathbb{Z}/n\mathbb{Z}}/\operatorname{im}(\varphi)$ with

$$\varphi : M \to M \quad m \mapsto \sum_{j=0}^{n-1} j.m.$$

Because $\ell > 2$, we have $H^2(\mathbb{Z}/2\mathbb{Z}, \mathfrak{g}) = H^2(\{1, c\}, \mathfrak{g}) = 0$ and the lifting ring is unobstructed. To get the number of variables we have to evaluate

$$Z^1(\{1, c\}, \mathfrak{g}) = \{f : \{1, c\} \to \mathfrak{g} \mid f(xy) = f(x) + {}^x f(y)\}.$$

Looking at $x = y = c$, we see that f is uniquely determined by a vector $v = f(c)$. Looking at $x = 1$, $y = c$, we see that $f(1) = v + {}^c v = 0$, i.e., that $v \in \mathfrak{g}^{c=-1}$. On the other hand, any such v defines an $f \in Z^1$ via $1 \mapsto 0, c \mapsto v$.
The modifications of this argument for the fixed-determinant case are straight-forward. □

A simple criterion for the vanishing of cohomology groups. Now assume that $\Gamma = \mathrm{Gal}_K$ for a local field K. Recall that, by local Tate duality, the Pontryagin dual of $H^2(\Gamma, \mathfrak{g})$ can be identified with $H^0(\Gamma, \mathfrak{g}^\vee) = (\mathfrak{g}^\vee)^\Gamma$. Together with the identification of $(\operatorname{ad}\bar\rho^{(0)})^\vee$ and $(\operatorname{ad}\bar\rho^{(0)})(1)$ via the trace pairing, this implies the following criterion for the vanishing of $H^2(\Gamma, \mathfrak{g}^{\mathtt{der}})$ in the case $G = \mathrm{GL}_n$:

Lemma 3.23 (Local case). *Let Γ be the absolute Galois group of a nonarchimedean local field, k be a finite field of characteristic ℓ and*

$$\bar\rho : \Gamma \to \mathrm{GL}_n(k)$$

a representation.

(1) *If $\operatorname{Hom}_\Gamma(\bar\rho, \bar\rho(1))$ vanishes, then $H^2(\Gamma, \operatorname{ad}\bar\rho)$ vanishes.*

(2) *Assume that $\ell \nmid n$. Then, if $\operatorname{Hom}_\Gamma(\bar\rho, \bar\rho(1))$ vanishes, also $H^2(\Gamma, \operatorname{ad}\bar\rho^0)$ vanishes.*

In the global case, there is no such duality and we record the following:

Lemma 3.24 (Global case). *Let $\Gamma = \mathrm{Gal}_{F,S}$ for a number field F and a (possibly) finite set S of places of F. Let k, $\bar\rho$ be as in Lemma 3.23 above.*

(1) *If $\operatorname{Hom}_\Gamma(\bar\rho, \bar\rho(1))$ vanishes, then $H^0(\Gamma, (\operatorname{ad}\bar\rho)^\vee)$ vanishes.*

(2) *Assume that $\ell \nmid n$. Then, if $\operatorname{Hom}_\Gamma(\bar\rho, \bar\rho(1))$ vanishes, also $H^0(\Gamma, (\operatorname{ad}\bar\rho^0)^\vee)$ vanishes.*

We easily deduce the following result, which also implies the vanishing of the error term δ in [Böckle 2013] (see Remark 5.2.3(d) of that work) for large ℓ:

Corollary 3.25. *There exists a constant C, depending only on n and F, such that Assumption 3.19 holds if $\operatorname{char}(k) > C$, $G = \mathrm{GL}_n$ and $\bar\rho$ is irreducible.*

Unobstructedness.

Definition 3.26. The functor $D_\Lambda^{(\square),[\chi]}(\bar\rho)$ is called unobstructed if $h^2(\Gamma, \mathfrak{g}^{[\mathtt{der}]}) = 0$.

Definition 3.27. A relatively representable subfunctor of $D_\Lambda^{(\square),[\chi]}(\bar\rho)$ is called smooth (of dimension m) if its representing object is isomorphic to $\Lambda[\![x_1, \dots, x_m]\!]$.

The most apparent application of the unobstructedness-property is that it implies the smoothness of the lifting/deformation ring; see [Böckle 2007]: Assume that $D_\Lambda^{\square,(\chi)}(\bar\rho)$ is smooth and (in the fixed-determinant case) that $\ell \gg 0$ and (in the unframed case) that $D_\Lambda^{(\chi)}(\bar\rho)$ is representable. Then

$$D_\Lambda^{\square,(\chi)}(\bar\rho) \cong \Lambda[\![x_1, \dots, x_{a(+c)}]\!] \quad \text{and} \quad D_\Lambda^{(\chi)}(\bar\rho) \cong \Lambda[\![x_1, \dots, x_{b(+c)}]\!]$$

with $b = h^1(\Gamma, \mathfrak{g})$, $c = h^1(\Gamma, \mathfrak{g}^{\mathtt{der}})' - b$, $a = b + \dim(\mathfrak{g}^{(\mathtt{der})}) - h^0(\Gamma, \mathfrak{g}^{(\mathtt{der})})$. The converse direction (i.e., that smoothness implies unobstructedness) is known not to hold (for general profinite groups Γ); see [Sprang 2013].

In order to relax this notion to functors corresponding to deformation conditions, we restrict to the case $\Gamma = \mathrm{Gal}_{F,S}$. Let $\mathcal{D}^{(\chi)} = (D_v^{(\chi)})_{v\in \mathrm{Pl}_F}$ be a system of deformation conditions and $\mathcal{L}^{(\chi)} = (L_v^{(\chi)})_{v\in\mathrm{Pl}_F}$ the corresponding system of local conditions.

Denote by $\mathfrak{g}^{(\mathtt{der}),\vee}$ the Tate dual of $\mathfrak{g}^{\mathtt{der}}$ and by $L_v^{(\chi),\perp}$ the annihilator of $L_v^{(\chi)}$ under the Tate pairing

$$H^i(F_v, \mathfrak{g}^{\mathtt{der},\vee}) \times H^{2-i}(F_v, \mathfrak{g}^{\mathtt{der}}) \to H^2(F_v, k(1)) \cong \mathbb{Q}/\mathbb{Z}$$

for $i = 1$; see [Neukirch et al. 2008, (7.2.6) Theorem]. Then we denote the corresponding dual Selmer group by

$$H^1_{\mathcal{L}^{(\chi),\perp}}(F, \mathfrak{g}^{(\mathtt{der}),\vee}) := \ker\left(\bigoplus_{v\in\mathrm{Pl}} \mathrm{res}_v \colon H^1(F, \mathfrak{g}^{(\mathtt{der}),\vee}) \to \bigoplus_{v\in\mathrm{Pl}} H^1(F_v, \mathfrak{g}^{(\mathtt{der}),\vee})/L_v^{(\chi),\perp}\right).$$

From now on, let us assume that $D_v^{(\chi)}$ for $v \notin S$ parametrizes unramified deformations.

Definition 3.28. We say that $D_{S,\Lambda}^{\mathcal{D}^{(\chi)}}(\bar\rho)$ (or $D_{S,\Lambda}^{\square,\mathcal{D}^{(\chi)}}(\bar\rho)$, or $D_{S,\Lambda}^{\square_\Sigma,\mathcal{D}^{(\chi)}}(\bar\rho)$ for some set of places Σ) has vanishing dual Selmer group if $H^1_{\mathcal{L}^{(\chi),\perp}}(F, \mathfrak{g}^{(\mathtt{der}),\vee}) = 0$.

Definition 3.29. Let $\boldsymbol{m} = (m_v)_{v\in S} \in \mathbb{N}_0^S$. We say that $D_{S,\Lambda}^{\mathcal{D}^{(\chi)}}(\bar\rho)$ (or $D_{S,\Lambda}^{\square,\mathcal{D}^{(\chi)}}(\bar\rho)$, or $D_{S,\Lambda}^{\square_\Sigma,\mathcal{D}^{(\chi)}}(\bar\rho)$) is globally unobstructed (of local dimensions $\boldsymbol{m}$) if its dual Selmer group vanishes and if each $D_\Lambda^{\square,\mathcal{D}^{(\chi_v)}}(\bar\rho_v)$ for $v \in S$ is smooth (of dimension m_v).

We remark that if $D_{S,\Lambda}^{\mathcal{D}^{(\chi)}}(\bar\rho)$ is globally unobstructed and representable, then by [Böckle 2007, Theorem 5.2] the representing object $R_{S,\Lambda}^{\mathcal{D}^{(\chi)}}(\bar\rho)$ is isomorphic to a power series ring in $h^1_{\mathcal{L}^{(\chi)}}(F, \mathfrak{g}^{(\mathtt{der})})^{(\prime)}$ variables.

Since $Ш^2_S(\mathfrak{g}^{(\mathtt{der})}) := H^1_{\mathcal{L}^\perp}(F, \mathfrak{g}^{(\mathtt{der}),\vee})^*$ vanishes,[2] the following short exact sequence results directly from that of [Böckle 2007, page 7]

$$0 \to Ш^2_S(\mathfrak{g}^{(\mathtt{der})}) \to H^2(\mathrm{Gal}_{F,S}, \mathfrak{g}^{(\mathtt{der})}) \to \bigoplus_{v\in S} H^2(F_v, \mathfrak{g}^{(\mathtt{der})}) \to H^0(F, \mathfrak{g}^{(\mathtt{der}),\vee})^* \to 0,$$

where $H^0(F, \mathfrak{g}^{(\mathtt{der}),\vee})^*$ vanishes for $\ell \gg 0$.

Proposition 3.30. *Assume that $D^{(\chi_v)}_\Lambda(\bar\rho_v)$ is unobstructed (for all $v \in S$) and that $D^{(\chi)}_{S,\Lambda}(\bar\rho)$ is globally unobstructed (without making an assumption on the dimension). Then $D^{(\chi)}_{S,\Lambda}(\bar\rho)$ is unobstructed in the sense of Definition 3.26. For $\ell \gg 0$, also the converse is true.*

4. A general framework for unobstructedness

For this section, we take the following static point of view: Let k be a finite field with ring of Witt vectors $W = W(k)$, let S be a finite set of finite places of F. We assume $\ell := \mathrm{char}(k) \notin S \cup \{2\}$. Then we fix a continuous representation

$$\bar\rho\colon \mathrm{Gal}_{F,S} \to G(k)$$

together with a lift $\chi\colon \mathrm{Gal}_{F,S} \to G^{\mathtt{ab}}(W)$ of the determinant. Let us moreover fix a Borel subgroup $B \subset G$ and denote by $\mathfrak{g}^{\mathtt{der}}$ (resp. $\mathfrak{b}^{\mathtt{der}}$) the Lie algebra of the derived subgroup $G^{\mathtt{der}}$ (resp. the Lie algebra of $B \cap G^{\mathtt{der}}$).

With respect to some choice of local deformation conditions[3]

- `min` of the restriction $\bar\rho_v$ of $\bar\rho$ to a decomposition group at $v \in S$,
- `sm` and `crys` of the restriction $\bar\rho_v$ of $\bar\rho$ to a decomposition group at $v \mid \ell$,

consider the following list of assumptions, where we leave out the W in the subscript of the occurring deformation functors and rings:

(sm/*k*) For each $v \mid \ell$, the subfunctor $D^{\square,\chi_v,\mathtt{sm}}(\bar\rho_v)$ of $D^{\square,\chi_v}(\bar\rho_v)$ is representable by a formally smooth (over W) object $R_v^{\square,\chi_v,\mathtt{sm}}$ (and we denote the relative dimension by $d_v^{\square,\mathtt{sm}}$).

(crys) For each $v \mid \ell$, the subfunctor $D^{\square,\chi_v,\mathtt{crys}}(\bar\rho_v)$ of $D^{\square,\chi_v}(\bar\rho_v)$ is representable by a formally smooth (over W) object $R_v^{\square,\chi_v,\mathtt{crys}}$ of relative dimension

$$d_v^{\square,\mathtt{crys}} = \dim(\mathfrak{g}^{\mathtt{der}}) + (\dim(\mathfrak{g}^{\mathtt{der}}) - \dim(\mathfrak{b}^{\mathtt{der}}))[F_v : \mathbb{Q}_\ell].$$

[2] We remark that the vanishing of the "Tate–Shafarevich group" $Ш^2_S(\mathfrak{g}^{(\mathtt{der})})$ implies that all obstructions for $D^{(\chi)}_\Lambda(\bar\rho_v)$ come from local obstructions, see [Böckle 2007, Theorem 3.1].

[3] During the following applications of the presented material, we will consider for `min` the condition of Section 5D, for `crys` the condition of Section 5C and for `sm` the unconditioned deformation condition. We stress, however, that for the purpose of this section we treat `min`, `crys`, `sm` purely formally as deformation conditions satisfying the listed assumptions of Definition 3.9.

(min) For each $v \in S$, the subfunctor $D^{\square,\chi_v,\mathtt{min}}(\bar\rho_v)$ of $D^{\square,\chi_v}(\bar\rho_v)$ is representable by a formally smooth (over W) object $R_v^{\square,\chi_v,\mathtt{min}}$ of relative dimension

$$d_v^{\square,\mathtt{min}} = \dim(\mathfrak{g}^{\mathtt{der}}).$$

(∞) For each $v \mid \infty$, the functor $D^{\square,\chi_v}(\bar\rho_v)$ is representable by an object (over W) of relative dimension $d_v^{\square} = \dim(\mathfrak{b}^{\mathtt{der}})$. (As $\ell > 2 = \#\operatorname{Gal}_{F_v}$, the strict ℓ-cohomological dimension $\operatorname{scd}_\ell(\operatorname{Gal}_{F_v})$ is zero, i.e., the representing object is automatically formally smooth over W.)

(Presentability) There exists a presentation

$$R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{sm}} \cong R_{S_\ell}^{\mathtt{loc},\mathtt{min},\mathtt{sm}}[\![x_1,\dots,x_a]\!]/(f_1,\dots,f_b)$$

for integers a, b fulfilling $a - b = (\#S_\ell - 1).\dim(\mathfrak{g}^{\mathtt{ab}})$. In this equation, we take

$$R_{S_\ell}^{\mathtt{loc},\mathtt{min},\mathtt{sm}} = \widehat{\bigotimes_{v\in S_\ell}} \tilde R_v \text{ with } \tilde R_v = \begin{cases} R_v^{\square,\chi_v,\mathtt{min}} & \text{if } v \in S; \\ R_v^{\square,\chi_v,\mathtt{sm}} & \text{if } v \mid \ell; \\ R_v^{\square,\chi_v} & \text{if } v \mid \infty. \end{cases} \tag{5}$$

($\boldsymbol{R} = \boldsymbol{T}$) The ring $R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{crys}}$ is formally smooth of relative dimension

$$r_0 := \dim(\mathfrak{g}).\#S_\ell - \dim(\mathfrak{g}^{\mathtt{ab}}).$$

Remark 4.1 (Taylor–Wiles condition). Let $v \mid \infty$ so that $\operatorname{scd}_\ell(\operatorname{Gal}_{F_v}) = 0$, then it follows from condition (∞), $\operatorname{scd}_\ell(\operatorname{Gal}_{F_v}) = 0$ and the remark following Definition 3.26 that

$$\dim(\mathfrak{b}^{\mathtt{der}}) = \dim_W(R^{\square}) = h^1(\operatorname{Gal}_{F_v},\mathfrak{g}^{\mathtt{der}})' + \dim(\mathfrak{g}^{\mathtt{der}}) - h^0(\operatorname{Gal}_{F_v},\mathfrak{g}^{\mathtt{der}}) = \dim(\mathfrak{g}^{\mathtt{der}}) - h^0(\operatorname{Gal}_{F_v},\mathfrak{g}^{\mathtt{der}}).$$

This implies

$$\sum_{v\mid\infty} h^0(\operatorname{Gal}_{F_v},\mathfrak{g}^{\mathtt{der}}) = [F:\mathbb{Q}].(\dim\mathfrak{g}^{\mathtt{der}} - \dim(\mathfrak{b}^{\mathtt{der}})). \tag{6}$$

We can now state the main result of this section.

Theorem 4.2. *Suppose the six conditions are met and, for $v \mid \ell$, write $d_v^{\square,\mathtt{sm}} = \dim(\mathfrak{g}^{\mathtt{der}}).([F_v:\mathbb{Q}_\ell]+1)$.*

(1) *The ring $R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{sm}}$ is formally smooth of relative dimension*

$$\#S_\ell.\dim(\mathfrak{g}) - \dim(\mathfrak{g}^{\mathtt{ab}}) + [F:\mathbb{Q}].\dim(\mathfrak{b}^{\mathtt{der}}).$$

If the unframed deformation functor $D_{S_\ell}^{\chi,\mathtt{min},\mathtt{sm}}$ is representable, then $R_{S_\ell}^{\chi,\mathtt{min},\mathtt{sm}}$ is formally smooth of relative dimension $[F:\mathbb{Q}].\dim(\mathfrak{b}^{\mathtt{der}})$.

(2) *Let $\mathcal{L} := (L_v^{\chi})_v$ be the system of local conditions corresponding to the deformation functor $D_{S_\ell}^{\chi,\mathtt{min},\mathtt{sm}}(\bar\rho)$. Assume*:
   (a) $\mathfrak{g} = \mathfrak{g}^{\mathtt{der}} \oplus \mathfrak{g}^{\mathtt{ab}}$ (*e.g., because* $\ell \gg 0$).
   (b) $H^0(\operatorname{Gal}_F,\mathfrak{g}^{\mathtt{der},\vee}) = 0$.
   (c) *For $v \in S$, we have* $\dim(L_v) = h^0(\operatorname{Gal}_{F_v},\mathfrak{g}^{\mathtt{der}})$.

   Then $H^1_{\mathcal{L}^\perp}(\operatorname{Gal}_{F,S},\mathfrak{g}^{\mathtt{der},\vee}) = H^0(\operatorname{Gal}_{F,S},\mathfrak{g}^{\mathtt{der}}) = 0$.

Remark 4.3. (1) As the deformation conditions $\mathtt{sm}$ and $\mathtt{crys}$ are relatively representable (see conditions (1) and (2)), $D_{S_\ell}^{\chi,\mathtt{min},\mathtt{sm}}$ is representable if $D_{S_\ell}^{\chi}$ is representable. For example, this is the case if $\bar\rho$ is absolutely irreducible (in the sense of Definition 3.14).

(2) For $v \notin S_\ell$, the equality $\dim(L_v) = h^0(\mathrm{Gal}_{F_v}, \mathfrak{g}^{\mathtt{der}})$ holds automatically if $\ell \gg 0$ (so that $\mathfrak{g} = \mathfrak{g}^{\mathtt{der}} \oplus \mathfrak{g}^{\mathtt{ab}}$).

Proof of Theorem 4.2. First remark that the second claim of part (1) follows by a straightforward lifting argument from Proposition 3.5, as $R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{sm}}$ is a power series ring over $R_{S_\ell}^{\chi,\mathtt{min},\mathtt{sm}}$, and from the formula $\dim \mathfrak{g} = \dim \mathfrak{g}^{\mathtt{der}} + \dim \mathfrak{g}^{\mathtt{ab}}$.

For the first sentence of (1), we use the shorthand notation $d_T^\star = \sum_{v\in T} d_v^\star$ for a subset T of Pl_F. Moreover, we write $d_\infty^\square$ for $d_{\Omega_\infty}^\square$ and $d_\ell^\star$ for $d_{\Omega_\ell}^\star$. Let us consider the commutative diagram:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & I & \longrightarrow & R_{S_\ell}^{\mathtt{loc},\mathtt{min},\mathtt{sm}} & \xrightarrow{f} & R_{S_\ell}^{\mathtt{loc},\mathtt{min},\mathtt{crys}} & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi'} & & \\
0 & \longrightarrow & J & \longrightarrow & R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{sm}} & \xrightarrow[g]{} & R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{crys}} & \longrightarrow & 0
\end{array}$$

In this diagram, the right square is a pushout square, $R_{S_\ell}^{\mathtt{loc},\mathtt{min},\mathtt{crys}}$ is defined as in (5) (but with $\tilde R_v = R_v^{\square,\chi_v,\mathtt{crys}}$ for $v \mid \ell$) and f, g are the canonical projections. Moreover, $\pi = \otimes_{v\in S_\ell}\pi_v$ is induced from the natural transformations

$$D_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{crys}} \to \tilde D_v,$$

where $\tilde D_v$ is the deformation functor corresponding to (i.e., represented by) the ring $\tilde R_v$ in (5) and, analogously, $\pi' = \otimes_{v\in S_\ell}\pi'_v$ is defined with $\mathtt{crys}$ in place of $\mathtt{sm}$.

Using the list of assumptions, we can rewrite the above diagram as:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & I & \longrightarrow & W[\![x_1,\dots,x_{d_\ell^{\square,\mathtt{sm}}+d_\infty^\square+d_S^{\square,\mathtt{min}}}]\!] & \xrightarrow{f} & W[\![x_1,\dots,x_{d_\ell^{\square,\mathtt{crys}}+d_\infty^\square+d_S^{\square,\mathtt{min}}}]\!] & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi'} & & \\
0 & \longrightarrow & J & \longrightarrow & W[\![x_1,\dots,x_m]\!]/(f_1,\dots,f_{m-\gamma}) & \xrightarrow[g]{} & W[\![x_1,\dots,x_{r_0}]\!] & \longrightarrow & 0
\end{array}$$

with $\gamma = (\#S_\ell - 1).\dim(\mathfrak{g}^{\mathtt{ab}}) + d_\ell^{\square,\mathtt{sm}} + d_\infty^\square + d_S^{\square,\mathtt{min}}$. By Lemma 3.3, $R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{sm}}$ is formally smooth if we can show $\mathrm{gen}(J) \le m - (m-\gamma) - r_0 = \gamma - r_0$. As f is a surjection of regular rings, it follows by the same token that $\mathrm{gen}(I) = d_\ell^{\square,\mathtt{sm}} - d_\ell^{\square,\mathtt{crys}}$. From the pushout property of the diagram, we can easily deduce that $\mathrm{gen}(J) \le \mathrm{gen}(I)$. Thus, we are left to show the inequality

$$\begin{aligned}
d_\ell^{\square,\mathtt{sm}} - d_\ell^{\square,\mathtt{crys}} \le \gamma - r_0 &= (\#S_\ell - 1).\dim(\mathfrak{g}^{\mathtt{ab}}) + d_\ell^{\square,\mathtt{sm}} + d_\infty^\square + d_S^{\square,\mathtt{min}} - \dim(\mathfrak{g}).\#S_\ell + \dim(\mathfrak{g}^{\mathtt{ab}}) \\
&= \#S_\ell.(\dim(\mathfrak{g}^{\mathtt{ab}}) - \dim(\mathfrak{g})) + d_\ell^{\square,\mathtt{sm}} + d_\infty^\square + d_S^{\square,\mathtt{min}}.
\end{aligned}$$

By assumptions **(min)** and (∞) and by the identity $\dim(\mathfrak{g}^{\mathtt{der}}) + \dim(\mathfrak{g}^{\mathtt{ab}}) = \dim(\mathfrak{g})$, this amounts to

$$d_\ell^{\square,\mathtt{crys}} \geq \dim(\mathfrak{g}^{\mathtt{der}}).(\#\Omega_\ell + [F:\mathbb{Q}]) - \dim(\mathfrak{b}^{\mathtt{der}})[F:\mathbb{Q}].$$

Assumption **(crys)** amounts precisely to the fact that this inequality is fulfilled (with equality), which implies the formal smoothness of $R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{sm}}$. Moreover, we easily check that the relative dimension of $R_{S_\ell}^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{sm}}$ is

$$\begin{aligned}\gamma &= (\#S_\ell - 1).\dim(\mathfrak{g}^{\mathtt{ab}}) + d_\ell^{\square,\mathtt{sm}} + d_\infty^{\square} + d_S^{\square,\mathtt{min}} \\ &= \#S_\ell.\dim\mathfrak{g}^{\mathtt{ab}} - \dim\mathfrak{g}^{\mathtt{ab}} + \dim\mathfrak{g}^{\mathtt{der}}.([F:\mathbb{Q}] + \#\Omega_\ell) + [F:\mathbb{Q}].\dim(\mathfrak{b}^{\mathtt{der}}) + \#S.\dim(\mathfrak{g}^{\mathtt{der}}) \\ &= \#S_\ell.\dim(\mathfrak{g}) + [F:\mathbb{Q}].\dim(\mathfrak{b}^{\mathtt{der}}) - \dim(\mathfrak{g}^{\mathtt{ab}}).\end{aligned}$$

Concerning part (2), note that (using condition (a)) we have an exact sequence

$$0 \to \mathfrak{g}/\mathfrak{g}^{\mathrm{Gal}_{F_v}} = \mathfrak{g}^{\mathtt{der}}/(\mathfrak{g}^{\mathtt{der}})^{\mathrm{Gal}_{F_v}} \to t_{D_W^{\square,\chi_v,\mathtt{sm}}(\bar\rho_v)} \to t_{D_W^{\chi_v,\mathtt{sm}}(\bar\rho_v)} \to 0$$

for $v \mid \ell$. Therefore, using condition (2), we have for $v \mid \ell$ the following:

$$\dim(L_v) = \dim t_{D_W^{\chi_v,\mathtt{sm}}(\bar\rho_v)} = h^0(\mathrm{Gal}_{F_v}, \mathfrak{g}^{\mathtt{der}}) + [F_v:\mathbb{Q}_\ell].\dim(\mathfrak{g}^{\mathtt{der}}).$$

Recall the Greenberg–Wiles formula [Neukirch et al. 2008, Theorem 8.7.9]:

$$\begin{aligned}&\dim H^1_{\mathcal{L}}(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der}}) - \dim H^1_{\mathcal{L}^\perp}(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der},\vee}) \\ &\qquad = h^0(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der}}) - h^0(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der},\vee}) + \sum_{v\in S_\ell}(\dim(L_v) - h^0(\mathrm{Gal}_{F_v}, \mathfrak{g}^{\mathtt{der}}))\end{aligned}$$

By [Böckle 2007, Section 5], we know that $H^1_{\mathcal{L}}(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der}})$ can be identified with the tangent space of the functor $D_{S_\ell}^{\chi,\mathtt{min},\mathtt{sm}}$ and hence (by part (2)) equals $[F:\mathbb{Q}].\dim(\mathfrak{b}^{\mathtt{der}})$. For $v \mid \infty$, we have $L_v \subset H^1(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der}}) = 0$. Thus, using the Taylor–Wiles formula (6) and assumption (b), the sum evaluates to

$$\sum_{v\in S_\ell}(\dim(L_v) - h^0(\mathrm{Gal}_{F_v}, \mathfrak{g}^{\mathtt{der}})) = [F:\mathbb{Q}].\dim(\mathfrak{g}^{\mathtt{der}}) - [F:\mathbb{Q}].(\dim(\mathfrak{g}^{\mathtt{der}}) - \dim(\mathfrak{b}^{\mathtt{der}})).$$

Therefore we get

$$-\dim H^1_{\mathcal{L}^\perp}(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der},\vee}) = h^0(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der}}).$$

As neither quantity can be negative, they must both vanish and the result follows. □

From the exact sequence

$$H^1_{\mathcal{L}^\perp}(\mathrm{Gal}_{F,S}, \mathfrak{g}^{\mathtt{der},\vee})^* \to Ш^2_{S_\ell}(\mathfrak{g}^{\mathtt{der}}) \to 0$$

(see, e.g., equation (9) on page 10 of [Böckle 2007]) we can deduce:

Corollary 4.4. *Under the assumptions of part* (2) *of Theorem 4.2,* $Ш^2_{S_\ell}(\mathfrak{g}^{\mathtt{der}})$ *vanishes. In particular, in view of Proposition 3.30 the unrestricted deformation functor* $D_{S_\ell}^{(\square_{S_\ell}),\chi}(\bar\rho)$ *is globally unobstructed precisely if the local deformation functors* $D^{(\square),\chi_v}(\bar\rho_v)$ *are relatively smooth for* $v \in S \cup \Omega_\ell$.

We remark that $D^{(\square),\chi_\nu}(\bar\rho_\nu)$ is relatively smooth for $\nu \in \Omega_\infty$ by Proposition 3.22, so Corollary 4.4 holds true with " ... for $\nu \in S_\ell$" in place of " ... for $\nu \in S \cup \Omega_\ell$".

Potential unobstructedness. We start with the following easy observation:

Proposition 4.5. *Let K be a local field and let K' be a finite extension of K such that ℓ does not divide the index $[K' : K]$. Let $\bar\rho$ be a G-valued residual representation of Gal_K and fix a lift χ of the determinant. Then unobstructedness of $D^{(\chi)}_\Lambda(\bar\rho \mid \mathrm{Gal}_{K'})$ implies unobstructedness of $D^{(\chi)}_\Lambda(\bar\rho)$.*

Proof. This follows immediately from the injectivity of

$$\mathrm{res}_{K' \mid K} \colon H^2(K, \mathfrak{g}^{(\mathtt{der})}) \to H^2(K', \mathfrak{g}^{(\mathtt{der})});$$

see [Neukirch et al. 2008, Corollary (1.5.7)]. □

This proof is not directly applicable to the global situation, as we have to keep track of the set of places at which we allow ramification. Therefore, we first describe a more flexible method which can also handle conditioned deformation functors:

Definition 4.6 (Dual-pre condition). Let $F' \mid F$ be a finite extension of number fields:

(1) Let $\nu' \in \mathrm{Pl}_{F'}$, $\nu \in \mathrm{Pl}_F$ such that $\nu' \mid \nu$. Moreover, let $L_\nu \subset H^1(F_\nu, \mathfrak{g}^{(\mathtt{der})})$, $L'_{\nu'} \subset H^1(F'_{\nu'}, \mathfrak{g}^{(\mathtt{der})})$ be local conditions. We say that L_ν is a dual-pre-$L'_{\nu'}$ condition if $\mathrm{res}^\vee_{\nu'}(L_\nu^\perp) \subset L_{\nu'}'^\perp$, where

$$\mathrm{res}^\vee_{\nu'} : H^1(F_\nu, \mathfrak{g}^{(\mathtt{der}),\vee}) \to H^1(F'_{\nu'}, \mathfrak{g}^{(\mathtt{der}),\vee})$$

denotes the usual restriction map.

(2) Let $\mathcal{L}' = (L'_{\nu'})_{\nu' \in \mathrm{Pl}_{F'}}$ be a system of local conditions for F'. We say that a system $\mathcal{L} = (L_\nu)_{\nu \in \mathrm{Pl}_F}$ of local conditions for F is dual-pre-$\mathcal{L}'$ if for each pair ν, ν' as above, L_ν is a dual-pre-$L'_{\nu'}$ condition.

Example 4.7. Let F, F' be as in Definition 4.6 and fix a finite set $S \subset \mathrm{Pl}_F$ such that $\bar\rho$ is unramified outside S. Take for $\mathcal{L}$ the local system parametrizing all deformations which are unramified outside S, i.e., $L_\nu = H^1(F_\nu, \mathfrak{g}^{(\mathtt{der})})$ if $\nu \in S$ and $L_\nu = H^1(\mathrm{Gal}_{F_\nu}/I_{F_\nu}, \mathfrak{g}^{(\mathtt{der})})$ otherwise. Analogously, let $\mathcal{L}'$ the local system parametrizing all deformations which are unramified outside S. Then any lift of $\bar\rho$ which is unramified outside S is, after restriction to $\mathrm{Gal}_{F'}$, a lift of $\bar\rho \mid \mathrm{Gal}_{F'}$ which is unramified outside S. But this implies easily that the restriction map $\mathrm{res}_{\nu'} : H^1(F_\nu, \mathfrak{g}^{(\mathtt{der})}) \to H^1(F'_{\nu'}, \mathfrak{g}^{(\mathtt{der})})$ maps L_ν into $L'_{\nu'}$ for any pair of places ν, ν' with $\nu' \mid \nu$. Using the fact that Tate duality is given by the cup product which sends unramified classes to zero, we see that $\mathcal{L}$ is dual-pre-$\mathcal{L}'$.

Lemma 4.8. *Let $\bar\rho$, F and F' be as above and assume $(\ell, [F' : F]) = 1$. Let $\mathcal{L} = (L_\nu)_{\nu \in \mathrm{Pl}_F}$, $\mathcal{L}' = (L'_{\nu'})_{\nu' \in \mathrm{Pl}_{F'}}$ be systems of local conditions (with associated deformation conditions $\mathcal{D}$ and $\mathcal{D}'$) such that $\mathcal{L}$ is dual-pre-$\mathcal{L}'$. Moreover, assume that $D^{[\square],(\chi),\mathcal{D}'}(\bar\rho \mid \mathrm{Gal}_{F'})$ has vanishing dual Selmer group. Then also $D^{[\square],(\chi),\mathcal{D}}(\bar\rho)$ has vanishing dual Selmer group.*

Proof. As above, the invertibility of $[F' : F]$ implies that the restriction map

$$H^1(\mathrm{Gal}_F, \mathfrak{g}^{(\mathtt{der}),\vee}) \to H^1(\mathrm{Gal}_{F'}, \mathfrak{g}^{(\mathtt{der}),\vee})$$

is injective. Consider the diagram with exact rows:

$$\begin{array}{ccccc} H^1_{\mathcal{L}^\perp}(F, \mathfrak{g}^{(\mathtt{der}),\vee}) & \hookrightarrow & H^1(F, \mathfrak{g}^{(\mathtt{der}),\vee}) & \longrightarrow & \bigoplus_{v\in \mathrm{Pl}_F} H^1(F_v, \mathfrak{g}^{(\mathtt{der}),\vee})/L_v^\perp \\ \downarrow{\scriptstyle\varphi} & & \downarrow & & \downarrow \\ H^1_{\mathcal{L}'^\perp}(F', \mathfrak{g}^{(\mathtt{der}),\vee}) & \hookrightarrow & H^1(F', \mathfrak{g}^{(\mathtt{der}),\vee}) & \longrightarrow & \bigoplus_{v'\in \mathrm{Pl}_{F'}} H^1(F'_{v'}, \mathfrak{g}^{(\mathtt{der}),\vee})/L'^{,\perp}_{v'} \end{array}$$

The vertical map on the right is defined because $\mathcal{L}$ is dual-pre-$\mathcal{L}'$, and this implies the well-definedness of φ. A simple diagram chase implies injectivity of φ, from which the claim follows. □

The following follows now directly from Example 4.7 and Lemma 4.8:

Corollary 4.9. *Let F be a number field and let F' be a finite extension of F such that ℓ does not divide the index $[F' : F]$. Let $\bar\rho$ be a G-valued residual representation of Gal_F which is unramified outside a finite set of places S and fix a lift χ of the determinant. Then unobstructedness of $D^{(\chi \mid \mathrm{Gal}_{F'})}_{\Lambda,S}(\bar\rho \mid \mathrm{Gal}_{F'})$ implies unobstructedness of $D^{(\chi)}_\Lambda(\bar\rho)$.*

5. Local deformation conditions for $G = \mathrm{GL}_n$

Let K be a finite extension of $\mathbb{Q}_p$ and let k be a finite field of characteristic ℓ. In the following, we consider deformation conditions for a continuous representation $\bar\rho : \mathrm{Gal}_K \to \mathrm{GL}_n(k)$.

5A. ***Unrestricted deformations ($p \neq \ell$).*** In the case $p \neq \ell$, we have the following result:

Theorem 5.1. $\mathrm{Spec}\, R^\square(\bar\rho)$ *is a reduced complete intersection, flat and equidimensional of relative dimension n^2 over* $\mathrm{Spec}\, W$.

Proof. This is Theorem 2.5 in [Shotton 2018]. □

5B. ***Unrestricted deformations ($p = \ell$).*** For the remainder of this subsection, we assume that $\bar\rho$ is the semisimplification of the reduction of a crystalline representation

$$\rho : \mathrm{Gal}_K \to \mathrm{GL}_n(L)$$

for a suitable finite extension L of $\mathbb{Q}_p$ with residue field k and for $p = \ell$. Denote the set of embeddings $\tau : K \hookrightarrow \overline{\mathbb{Q}}_p$ by $\mathbb{E}_K$ and for $\tau \in \mathbb{E}_K$ denote by $\mathrm{HT}_\tau(\rho)$ the multiset of Hodge–Tate weights of ρ with respect to τ.

Theorem 5.2. *Assume that $K \mid \mathbb{Q}_p$ is unramified and that for each $\tau \in \mathbb{E}_K$:*

(1) *There exists an $\alpha \in \mathbb{Z}$ such that all Hodge–Tate weights in $\mathrm{HT}_\tau(\rho)$ lie in the range $[\alpha, \alpha + \ell - 3]$.*

(2) *The Hodge–Tate weights of ρ are nonconsecutive, i.e., if two numbers $a, b \in \mathbb{Z}$ occur in $\mathrm{HT}_\tau(\rho)$, then $|a - b| \neq 1$.*

Then $R^\square(\bar\rho) \cong W[\![x_1, \ldots, x_m]\!]$ with $m = n^2.([K : \mathbb{Q}_\ell] + 1)$.

Before we come to the proof, recall the theory of Fontaine and Laffaille [1982], as normalized in [Clozel et al. 2008] (see also [Barnet-Lamb et al. 2014, Section 1.4]): We consider the category $\underline{\mathrm{FL}}_{\mathcal{O}_K,\mathcal{O}_L}$, consisting of $\mathcal{O}_K \otimes_{\mathbb{Z}_\ell} \mathcal{O}_L$-modules M, endowed with a decreasing filtration $(\mathrm{Fil}^i M)_{i\in\mathbb{Z}}$ with $\mathrm{Fil}^0 M = M$ and $\mathrm{Fil}^{\ell-1} M = 0$ and a family of $\mathrm{Frob}\otimes 1$-linear maps $\mathrm{Fil}^i M \to M$ such that $\varphi^i \mid \mathrm{Fil}^{i+1} = \ell.\varphi^{i+1}$ and $\sum_i \varphi^i(\mathrm{Fil}^i M) = M$. Let $\underline{\mathrm{FL}}_{\mathcal{O}_K,k}$ denote the full subcategory of finite length objects which are annihilated by the maximal ideal $\varpi_L.\mathcal{O}_L$. We need the following well-known facts:

- There exists an exact, fully faithful, covariant and $\mathcal{O}_L$-linear functor
$$\mathtt{G}_K : \underline{\mathrm{FL}}_{\mathcal{O}_K,\mathcal{O}_L} \to \mathrm{Rep}_{\mathcal{O}_L}(\mathrm{Gal}_K).$$
The essential image is closed under taking subobjects and quotients. Moreover, $\mathtt{G}_K$ restricts to a functor
$$\underline{\mathrm{FL}}_{\mathcal{O}_K,k} \to \mathrm{Rep}_k(\mathrm{Gal}_K).$$
- For $\boldsymbol{M} \in \underline{\mathrm{FL}}_{\mathcal{O}_K,\mathcal{O}_L}$ projective over $\mathcal{O}_L$, we have
$$\mathrm{HT}_\tau(\mathtt{G}_K(\boldsymbol{M} \otimes_{\mathbb{Z}_p} \mathbb{Q}_p)) = \mathrm{FL}_\tau(\boldsymbol{M} \otimes_{\mathcal{O}_L} k),$$
where for $\boldsymbol{N} \in \underline{\mathrm{FL}}_{\mathcal{O}_K,k}$ we denote by $\mathrm{FL}_\tau(\boldsymbol{N})$ the multiset of integers i, such that
$$\mathrm{gr}^i(\boldsymbol{N}^\tau) = \mathrm{Fil}^i N \otimes_{\mathcal{O}_K\otimes_{\mathbb{Z}_p}\mathcal{O}_L,\tau\otimes 1} \mathcal{O}_L / \mathrm{Fil}^{i+1} N \otimes_{\mathcal{O}_K\otimes_{\mathbb{Z}_p}\mathcal{O}_L,\tau\otimes 1} \mathcal{O}_L$$
does not vanish, where i is counted with multiplicity $\dim_k \mathrm{gr}^i(\boldsymbol{N}^\tau)$.
- Assuming condition (1) of Theorem 5.2, any Gal_K-stable $\mathcal{O}_L$-lattice of ρ is in the image of $\mathtt{G}_K$, and so is its reduction $\Lambda/\varpi_L.\Lambda$.
- Morphisms in $\underline{\mathrm{FL}}_{\mathcal{O}_K,k}$ are strict with filtrations. If $f : \boldsymbol{M} \to \boldsymbol{N}$ is such a morphism, then $f(\mathrm{Fil}^i M) = f(M) \cap \mathrm{Fil}^i N$ for all $i \in \mathbb{Z}$. In particular, if $\boldsymbol{M}, \boldsymbol{N} \in \underline{\mathrm{FL}}_{\mathcal{O}_K,k}$ fulfill
$$\mathrm{FL}_\tau(\boldsymbol{M}) \cap \mathrm{FL}_\tau(\boldsymbol{N}) \tag{7}$$
for all $\tau \in \mathbb{E}_K$, then $\mathrm{Hom}_{\underline{\mathrm{FL}}_{\mathcal{O}_K,k}}(\boldsymbol{M}, \boldsymbol{N}) = 0$.

Proof of Theorem 5.2. As $h^2(K, \mathrm{ad}\,\bar\rho)$ is an upper bound on the number of generators of the kernel of a surjection $W[\![x_1, \dots, x_s]\!] \twoheadrightarrow R^\square(\bar\rho)$ with $s = \dim Z^1(K, \mathrm{ad}\,\bar\rho)$ (see [Allen 2016, Proposition 2.1.2]), we have to prove
$$H^2(K, \mathrm{ad}\,\bar\rho) = 0. \tag{8}$$
Moreover, using the exact sequence
$$0 \to \mathrm{ad}\,\bar\rho/(\mathrm{ad}\,\bar\rho)^{\mathrm{Gal}_K} \to t_{D^\square_W(\bar\rho)} \to t_{D_W(\bar\rho)} \to 0$$
and the local Euler–Poincaré formula, we can compute
$$s = h^1(K, \mathrm{ad}\,\bar\rho) + n^2 - h^0(K, \mathrm{ad}\,\bar\rho) = n^2 - \chi(K, \mathrm{ad}\,\bar\rho) = n^2([K : \mathbb{Q}_p] + 1).$$
Thus, (8) implies the claim.

As the trace pairing identifies $\operatorname{ad}\bar\rho^\vee$ and $\operatorname{ad}\bar\rho(1)$, we are finished if we can show that

$$H^2(K, \operatorname{ad}\bar\rho)^* \cong H^0(K, \operatorname{ad}\bar\rho^\vee) \cong H^0(K, \operatorname{ad}\bar\rho(1)) \cong \operatorname{Hom}_{\mathrm{Gal}_K}(\bar\rho, \bar\rho(1)) = 0.$$

Because $\operatorname{Hom}_{\mathrm{Gal}_K}(\bar\rho, \bar\rho(1)) \cong \operatorname{Hom}_{\mathrm{Gal}_K}(\bar\rho(1-\alpha), \bar\rho(2-\alpha))$, we can assume without loss of generality that $\alpha = 1$.

It is easy to see that we can choose a Gal_K-stable $\mathcal{O}_L$-lattice Λ of ρ such that its reduction is semisimple, i.e., $\Lambda/\varpi_L.\Lambda \cong \bar\rho$ (if necessary, after replacing ρ by a base change $\rho \otimes_L L'$ to a sufficiently ramified finite extension L' of L, which does not affect the validity of (8)). By our first assumption that all weights of ρ lie in the range $[1, \ell-2]$, it thus follows that $\bar\rho$ is of the form $\mathtt{G}_\mathtt{K}(\boldsymbol{M})$ for a suitable $\boldsymbol{M} \in \underline{\mathrm{FL}}_{\mathcal{O}_K,k}$. By the same argument, $\bar\rho(1) = \mathtt{G}_\mathtt{K}(\boldsymbol{N})$ for a suitable $\boldsymbol{N} \in \underline{\mathrm{FL}}_{\mathcal{O}_K,k}$. As the cyclotomic character shifts the weights by -1, the second condition translates precisely into the condition (7). Thus, using that $\mathtt{G}_\mathtt{K}$ is fully faithful, we get

$$0 = \operatorname{Hom}_{\underline{\mathrm{FL}}_{\mathcal{O}_K,k}}(\boldsymbol{M}, \boldsymbol{N}) \cong \operatorname{Hom}_{\mathrm{Gal}_K}(\bar\rho, \bar\rho(1)). \qquad \square$$

5C. ***Crystalline deformations ($\ell = p$).*** Let K be unramified. Consider again a representation $\rho\colon \mathrm{Gal}_K \to \mathrm{GL}_n(L)$ which fulfills the conditions of Theorem 5.2. We will also make the additional assumption that all occurring Hodge–Tate weights of $\bar\rho$ have multiplicity one. We will consider the deformation problem $\mathtt{crys}$ of $\bar\rho$ consisting of those lifts $\tilde\rho : \mathrm{Gal}_K \to \mathrm{GL}_n(A)$ of $\bar\rho$ for which $\tilde\rho \otimes_A A'$ lies in the essential image of $\mathtt{G}_K$ for all Artinian quotients A' of A (see [Clozel et al. 2008, Section 2.4.1]). We refer to those lifts as FL-*crystalline lifts* of $\bar\rho$.

That $\mathtt{crys}$ defines a deformation condition in the sense of Definition 3.16 was already remarked in [Clozel et al. 2008] and follows easily from the Ramakrishna framework [1993]: We remarked already in Section 5B that the essential image of $\mathtt{G}_K$ is closed under subobjects and quotients. That the essential image is closed under direct sums follows immediately from the exactness of $\mathtt{G}_K$, since then $\mathtt{G}_K$ preserves direct sums (see [Freyd 1964, Theorem $3.12^{(*)}$]). Thus we can record the following (where for part (2) we refer to the remark just below Proposition 3.10):

Lemma 5.3. *Let Λ be the ring of integers of a finite, totally ramified extension E of $\mathrm{Quot}(W(k))$ and let Λ' be the ring of integers of a finite, totally ramified extension of E (so that we have $k = k_\Lambda = k_{\Lambda'}$.) Then*:

(1) *The functor $D_\Lambda^{\square,\mathtt{crys}}(\bar\rho)$ is representable by a quotient $R_\Lambda^{\square,\mathtt{crys}}(\bar\rho)$ of $R_\Lambda^\square(\bar\rho)$.*

(2) *The functor $D_{\Lambda'}^{\square,\mathtt{crys}}(\bar\rho)$ is representable by*

$$R_{\Lambda'}^{\square,\mathtt{crys}}(\bar\rho) \cong \Lambda' \otimes_\Lambda R_\Lambda^{\square,\mathtt{crys}}(\bar\rho). \tag{9}$$

We remark that the condition $\mathtt{crys}$ fulfills the extended requirements as described in Remark 3.11, so that (9) holds even if $\infty > [k_{\Lambda'} : k_\Lambda] > 1$.

Lemma 5.4. *Under the above hypotheses*

$$R_\Lambda^{\square,\mathtt{crys}}(\bar\rho) \cong \Lambda[\![x_1, \dots, x_m]\!]$$

with $m = n^2 + [K : \mathbb{Q}_\ell].n.(n-1)/2$.

Proof. This is a part of the statement of [Clozel et al. 2008, Corollary 2.4.3]. □

Let us also note the following useful compatibility with base change:

Lemma 5.5. *Let K' be a finite unramified extension of K with associated inclusion map $\iota_{K'|K} : \mathrm{Gal}_{K'} \to \mathrm{Gal}_K$. Set $\bar\rho' = \bar\rho \circ \iota_{K'|K}$. Let $\tilde\rho$ be a crystalline lift of $\bar\rho$. Then $\tilde\rho' = \tilde\rho \circ \iota_{K'|K}$ is a crystalline lift of $\bar\rho'$.*

In particular, the restriction map $\mathrm{res} : H^1(K, \mathrm{ad}\,\bar\rho) \to H^1(K', \mathrm{ad}\,\bar\rho')$ *maps the tangent subspace associated to the crystalline deformation condition for $\bar\rho$ into the tangent subspace associated to the crystalline deformation condition for $\bar\rho'$.*

Proof. This is a direct consequence of the following compatibility of the Fontaine–Laffaille functor with base change: Let $\boldsymbol{M} \in \underline{\mathrm{MF}}_{\mathcal{O}_K,\mathcal{O}_L}$, then $\mathcal{O}_{K'} \otimes_{\mathcal{O}_K} \boldsymbol{M}$ defines an object of $\underline{\mathrm{MF}}_{\mathcal{O}_{K'},\mathcal{O}_L}$. It follows from the definition of the functors $\mathtt{G}_K, \mathtt{G}_{K'}$ and a calculation analogous to the one in Section 3.11 of [Fontaine and Laffaille 1982] that $\mathtt{G}_K(\boldsymbol{M})$ and $\mathtt{G}_{K'}(\mathcal{O}_{K'} \otimes_{\mathcal{O}_K} \boldsymbol{M})$ are isomorphic as $\mathcal{O}_L$-modules and that this isomorphism commutes with the action of $\mathrm{Gal}_{K'}$. In other words,

$$r^K_{K'}(\mathtt{G}_K(\boldsymbol{M})) \cong \mathtt{G}_{K'}(\mathcal{O}_{K'} \otimes_{\mathcal{O}_K} \boldsymbol{M})$$

where $r^K_{K'}$ denotes the restriction to $\mathrm{Gal}_{K'}$. □

5D. ***Minimally ramified deformations ($p \neq \ell$).*** For this subsection, recall from [Clozel et al. 2008, Section 2.4.4] the minimal ramification condition for a lift ρ of $\bar\rho$. Let P_K denote the kernel of one (hence, any) surjection $I_K \twoheadrightarrow \mathbb{Z}_\ell$. Moreover, let $\Delta_{\bar\rho}$ denote the set of equivalence classes of P_K-representations over k such that $\mathrm{Hom}_{P_K}(\tau, \bar\rho) \neq 0$. Then the following can easily be deduced from the material in [loc. cit., Section 2.4.4], in particular [loc. cit., Corollary 2.4.21]:

Proposition 5.6. *Assume that any $\tau \in \Delta_{\bar\rho}$ is absolutely irreducible. Then we have*:

(1) *The condition of being minimally ramified defines a lifting condition, denoted* `min`*. The representing universal object fulfills*

$$R^{\square,\mathtt{min}}_\Lambda(\bar\rho) \cong \Lambda[\![X_1, \dots, X_{n^2}]\!].$$

(2) *If Λ is the ring of integers of some finite extension of* $\mathrm{Quot}(\Lambda)$ *with residue field $k_{\Lambda'} = k$, we have*

$$R^{\square,\mathtt{min}}_{\Lambda'}(\bar\rho') \cong \Lambda' \otimes_\Lambda R^{\square,\mathtt{min}}_\Lambda(\bar\rho).$$

We will be particularly interested in the case where $\bar\rho$ has *unipotent ramification* i.e., where $\bar\rho(P_K) = \{1\}$.[4] In the unipotent case, we have a strong connection between minimally ramified liftings and liftings of prescribed type as considered in [Shotton 2015]. In order to make this precise, let E denote the quotient field of Λ and $\bar E$ its algebraic closure.

[4] This notion is explained by the observation that $\bar\rho$ is unipotently ramified if and only if $\bar\rho(I_K)$ lies in a conjugate of the standard unipotent subgroup consisting of upper-triangular matrices in $\mathrm{GL}_n(k)$ with diagonal entries all equal to 1.

Definition 5.7 [Shotton 2015, Definition 2.10]. Let $\tau : I_K \to \mathrm{GL}_n(\bar{E})$ be a representation which extends to a continuous representation of the Weil group W_K of K (considered with the ℓ-adic topology). Then the isomorphism class of τ is called an *inertial type*. (*Warning*: This differs from the usual definition of an inertial type as e.g., in [Gee and Kisin 2014].)

Let ρ be a lift of $\bar{\rho}$ which has values in $\bar{E}$, then we say that ρ "is of type τ" if $\rho \mid I_K$ is isomorphic to τ.

For the following we consider a τ which is defined over E. Then we say that a morphism $x : \mathrm{Spec}\, \bar{E} \to \mathrm{Spec}\, R^{\square}_{\Lambda}(\bar{\rho})$ is of type τ if the associated $\bar{E}$-valued representation ρ_x is of type τ. This notion depends only on the image of x (because τ is defined over E).

Definition 5.8 (Fixed type deformation ring [Shotton 2015, Definiton 2.14]). Let $R^{\square,\tau}_{\Lambda}(\bar{\rho})$ be the reduced quotient of $R^{\square}_{\Lambda}(\bar{\rho})$ which is characterized by the requirement that $\mathrm{Spec}\, R^{\square,\tau}_{\Lambda}(\bar{\rho})$ is the Zariski closure of the $\bar{E}$-points of type τ in $\mathrm{Spec}\, R^{\square}_{\Lambda}(\bar{\rho})$.

A general classification of inertial types is given in Section 2.2.1 of [Shotton 2015]. Under the unipotent ramification assumption, this becomes particularly simple: The set $\mathcal{I}^{\mathtt{uni}}$ of the isomorphism classes of inertial types which are trivial on P_K is in bijection with the set $\mathcal{Y}_n$ of Young diagrams of size n. The partition $(l_1, \ldots, l_k)$ (with $l_i \geq l_{i+1}$) corresponds (using the notation of [loc. cit.]) to the type given by the I_K-restriction of the W_K-representation

$$\bigoplus_{i=1}^{k} \mathrm{Sp}(\mathbf{1}, l_i),$$

where $\mathrm{Sp}(\bullet, \bullet)$ is defined as in [loc. cit., Section 3.1]. We can express this differently: Each member of $\mathcal{I}^{\mathtt{uni}}$ is uniquely characterized by (the conjugacy class of) its value on the generator $\zeta := \zeta_{\mathrm{triv}}$ of I_K/P_K, and a bijection $\nabla : \mathcal{Y}_n \to \mathcal{I}^{\mathtt{uni}}$ is given by

$$(l_1, \ldots, l_k) \overset{\nabla}{\longmapsto} \tau(\zeta) = \left[1 + \begin{pmatrix} \mathcal{B}_{l_1} & & & \\ & \mathcal{B}_{l_2} & & \\ & & \ddots & \\ & & & \mathcal{B}_{l_k} \end{pmatrix} \right] \quad \text{with} \quad \mathcal{B}_m = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix} \in \mathbb{M}_{m\times m}(E). \tag{10}$$

On the other hand, we can associate to a $\tau \in \mathcal{I}^{\mathtt{uni}}$ a partition of n by considering the kernel sequences:

$$\Theta : \mathcal{I}^{\mathtt{uni}} \to \mathcal{Y}_n, \quad \tau \mapsto (s_1, \ldots, s_r)$$

with

$$s_i := \dim \ker(\tau(\zeta) - \mathbf{1})^i - \dim \ker(\tau(\zeta) - \mathbf{1})^{i-1}$$

and

$$r := \min\{i \mid \dim \ker(\tau(\zeta) - \mathbf{1})^i = \dim \ker(\tau(\zeta) - \mathbf{1})^{i+1}\} = \min\{i \mid \ker(\tau(\zeta) - \mathbf{1})^i = V\}.$$

(Here, V is the vector space underlying τ and we use the convention that f^0 is the identity map for any linear map f.) It follows easily from the characterization of $\mathcal{I}^{\mathtt{uni}}$ in (10) that $s_i \geq s_{i+1}$, i.e., that Θ has values in $\mathcal{Y}_n$.

It is an easy combinatorial calculation to check that τ is uniquely characterized by its value under Θ and that each Young diagram occurs as a kernel sequence (i.e., that Θ is a bijection). More precisely, we have:

Lemma 5.9. *The map $\Theta \circ \nabla^{-1} : \mathcal{Y}_n \to \mathcal{Y}_n$ is given by the conjugation operation on Young diagrams (see [Fulton and Harris 1991, Section 4.1] or [Harris et al. 2008, Section 2.8]). In particular, for a given $\tau \in \mathcal{I}^{\mathtt{uni}}$, the block matrix structure of $\tau(\zeta)$ (up to reordering blocks) as in (10) determines its kernel sequence and vice versa.*

Proof. Retaining the notation used in (10), we first remark that for $i \in \mathbb{N}_0$ we have

$$\dim \ker \mathcal{B}_m^i = \min(i, m).$$

Thus, setting $\mathcal{B} = \operatorname{diag}(\mathcal{B}_{l_1}, \dots, \mathcal{B}_{l_k})$, we get

$$\dim \ker \mathcal{B}^i = \sum_{j=1}^{k} \min(i, l_j).$$

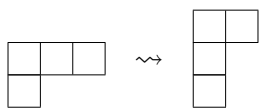
Consequently the kernel sequence $(s_1, \dots, s_r)$ associated to $(l_1, \dots, l_k)$ is given by

$$s_i = \sum_{j=1}^{k} \min(i, l_j) - \min(i-1, l_j) = \#\{j \mid l_j \geq i\} = \max\{j \mid l_j \geq i\}$$

and

$$r = \max\{l_j \mid j = 1, \dots, k\} = l_1.$$

Hence, the transition $(l_1, \dots, l_k) \rightsquigarrow (s_1, \dots, s_r)$ is precisely the conjugation operation of reflecting a Young diagram at the main diagonal (see [Harris et al. 2008, Section 2.8]), e.g.,



□

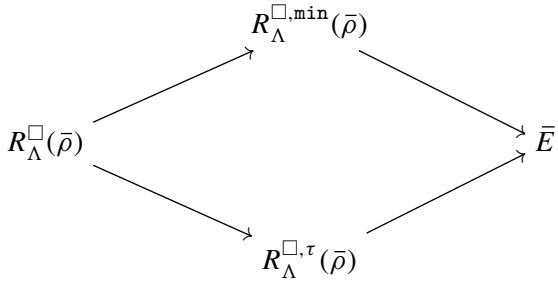
In order to state the desired comparison result, let us recall that we consider a residual representation $\bar{\rho} : \mathrm{Gal}_K \to \mathrm{GL}_n(k)$ with unipotent ramification. Let $\underline{\lambda} = (l_1, \dots, l_k) \in \mathcal{Y}_n$ such that $\bar{\rho}(\zeta) \sim \mathbf{1} + \operatorname{diag}(\mathcal{B}_{l_1}, \dots, \mathcal{B}_{l_k})$. Let $\tau = \nabla(\underline{\lambda}) \in \mathcal{I}^{\mathtt{uni}}$.

Theorem 5.10. *Assume $\bar{\rho}$ is unipotently ramified and τ as above. Then there is an isomorphism of the quotients*

$$R_\Lambda^{\square,\tau}(\bar{\rho}) \cong R_\Lambda^{\square,\mathtt{min}}(\bar{\rho}) \cong \Lambda[\![X_1, \dots, X_{n^2}]\!]$$

of $R_\Lambda^{\square}(\bar{\rho})$, i.e., a lifting of $\bar{\rho}$ is minimally ramified if and only if it is of type τ.

Proof. The diagram



allows us to consider the $\bar E$-points of $\operatorname{Spec} R_\Lambda^{\square,\mathtt{min}}(\bar\rho)$ and $\operatorname{Spec} R_\Lambda^{\square,\tau}(\bar\rho)$ as subsets of the $\bar E$-points of $\operatorname{Spec} R_\Lambda^{\square}(\bar\rho)$. We claim that they are equal: Translated into terms of $\bar E$-valued representations, we have to compare the sets

$$\Xi^{\mathtt{min}} = \left\{ \rho : \mathrm{Gal}_K \to \mathrm{GL}_n(\bar E) \;\middle|\; \begin{array}{l} \rho \text{ lifts } \bar\rho \text{ and has values in } \mathcal{O}_{\bar E}, \\ \dim \ker(\rho(\zeta)-\mathbf{1})^{i-1} - \dim \ker(\rho(\zeta)-\mathbf{1})^{i} = l_i \forall i \end{array} \right\}$$

and

$$\Xi^{\tau} = \{\rho : \mathrm{Gal}_K \to \mathrm{GL}_n(\bar E) \mid \rho \text{ lifts } \bar\rho \text{ and has values in } \mathcal{O}_{\bar E},\ \rho \mid I_K \cong \tau\}.$$

Lemma 5.9 implies that $\Xi^{\mathtt{min}} = \Xi^{\tau}$.

Now by definition of the ring $R_\Lambda^{\square,\tau}(\bar\rho)$ (as the schematic closure of the points in Ξ^{τ}) we have

$$\ker(R_\Lambda^{\square}(\bar\rho) \to R_\Lambda^{\square,\tau}(\bar\rho)) = \bigcap_{\rho\in\Xi^{\tau}} \ker(\rho).$$

Moreover, we clearly have

$$\ker(R_\Lambda^{\square}(\bar\rho) \to R_\Lambda^{\square,\mathtt{min}}(\bar\rho)) \subseteq \bigcap_{\rho\in\Xi^{\mathtt{min}}} \ker(\rho).$$

Hence, by $\Xi^{\tau} = \Xi^{\mathtt{min}}$ we get a factorization

$$R_\Lambda^{\square}(\bar\rho) \twoheadrightarrow R_\Lambda^{\square,\mathtt{min}}(\bar\rho) \overset{\varphi}{\twoheadrightarrow} R_\Lambda^{\square,\tau}(\bar\rho)$$

where the middle and the right ring have the same spectrum as topological spaces. Now we know by Proposition 5.6 that $R_\Lambda^{\square,\mathtt{min}}(\bar\rho)$ is formally smooth over Λ of relative dimension n^2 and that $\dim R_\Lambda^{\square,\tau}(\bar\rho) = n^2+1$ (combine Theorem 2.4 and Proposition 2.15 of [Shotton 2015]). Thus, φ is an isomorphism by Proposition 3.4 and the claim follows. □

5E. ***Taylors deformation condition*** **$(1,\dots,1)$ ($\ell \neq p$).** We continue to consider a unipotently ramified residual representation $\bar\rho : \mathrm{Gal}_K \to \mathrm{GL}_n(k)$. If $A \in \mathcal{C}_{\mathcal{O}}$ is a coefficient ring, we say that an A-valued lift ρ of $\bar\rho$ fulfills the condition $(1,\dots,1)$ if $\mathrm{charPoly}(\rho(\xi)) = (T-1)^n$ for all $\zeta \in I_K$. By our assumption that $\bar\rho$ is unipotently ramified, it is sufficient to check the case where ξ is a topological generator of the tame

inertia. This defines a deformation condition (and, in comparison to [Taylor 2008], we don't assume that $\bar\rho$ is trivial; see [Thorne 2012, Remark before Proposition 3.17]).

Proposition 5.11. *If a lift ρ is minimally ramified, it fulfills the Taylor condition. In particular, there is a canonical surjection*

$$R^{\square,(1,\dots,1)}(\bar\rho) \twoheadrightarrow R^{\square,\mathtt{min}}(\bar\rho),$$

and a morphism $R^{\square,(1,\dots,1)}(\bar\rho) \to A$ factors through this surjection if and only if the associated A-valued lift of $\bar\rho$ is minimally ramified.

Proof. By the unipotency assumption, we can assume that $\bar\rho \mid P_K$ is trivial and $\bar\rho(\zeta)$ is upper-triangular with each diagonal entry equal to 1 (where ζ is a topological generator of I_K/P_K). If a lift ρ is minimal, it follows that $\rho \mid P_K$ is trivial and that $\rho(\zeta)$ is unipotent; see [Clozel et al. 2008, Lemma 2.4.15, Assertion 3 $\Rightarrow$ 1]. It follows that $\rho(\sigma)$ is unipotent for any $\sigma \in I_K$. This proves the claim. □

Proposition 5.12. *Let L be a finite extension of K. Let*

$$\rho^{\square,(1,\dots,1)} : G_K \to \mathrm{GL}_n(R^{\square,(1,\dots,1)}(\bar\rho))$$

be the universal lifting of $\bar\rho$ with respect to the condition $(1,\dots,1)$ and let

$$\rho_L^{\square,(1,\dots,1)} : G_L \to \mathrm{GL}_n(R^{\square,(1,\dots,1)}(\bar\rho \mid G_L))$$

be the universal lifting of $\bar\rho \mid G_L$ with respect to the condition $(1,\dots,1)$. Then there exists a unique morphism of $\mathcal{C}_W$-algebras $\varphi : R^{\square,(1,\dots,1)}(\bar\rho \mid G_L)/(\ell) \to R^{\square,(1,\dots,1)}(\bar\rho)/(\ell)$ such that

$$\overline{\rho^{\square,(1,\dots,1)}} \mid G_L = \varphi \circ \overline{\rho_L^{\square,(1,\dots,1)}}.$$

Proof. The lifting $\rho^{\square,(1,\dots,1)}$ fulfills the condition $(1,\dots,1)$, i.e., $\mathrm{charPoly}(\rho^{\square,(1,\dots,1)}(\sigma)) = (T-1)^n$ for all $\sigma \in I_K$. As $I_L \subset I_K$, $\rho^{\square,(1,\dots,1)} \mid G_L$ is a deformation of $\bar\rho \mid G_L$ which fulfills the condition $(1,\dots,1)$, i.e., factors via $\mathrm{GL}_n(R^{\square,(1,\dots,1)}(\bar\rho \mid G_L))$. This implies the existence of a map $R^{\square,(1,\dots,1)}(\bar\rho \mid G_L) \to R^{\square,(1,\dots,1)}(\bar\rho)$ whose mod-ℓ reduction fulfills the required properties. □

Lemma 5.13. *Let $\tilde\rho$ be an A-valued lift, where we assume that A is reduced. Write $X = \tilde\rho(\zeta)$. Then $\chi_X := \mathrm{charPoly}(X)$ equals $(T-1)^n$ if $\ell \geq q^{n!}$.*

Proof. Assume first that A is an integral domain. By the condition $\varphi X \varphi^{-1} = X^q$ we see that raising to the q-th power permutes the eigenvalues of X (understood as a list of n elements). Thus, any eigenvalue of X must be a $(q^{\#S_n} - 1) = (q^{n!} - 1)$-th root of unity. Thus, if $Q(\mu)$ denotes the decomposition field of the polynomial $f(T) = T^{q^{n!}-1} - 1$ over the quotient field of A and $A(\mu)$ denotes the integral closure of A in $Q(\mu)$, then χ_X decomposes completely in $A(\mu)[T]$. On the other hand, each eigenvalue of X is sent to 1 by the canonical reduction map

$$\pi' : A(\mu) = A(\mu) \otimes_A A \to A(\mu) \otimes_A k.$$

As the kernel of π' is a pro-ℓ-subgroup and as $(\ell^m, q^{n!}-1)=1$ for any $m\in\mathbb{N}$, it follows that any eigenvalue of X is 1, i.e., that $\chi_X=(T-1)^n$. The result for a general (reduced) A follows easily from using the embedding

$$A\hookrightarrow\prod_{\mathfrak{q}}A/\mathfrak{q},$$

where $\mathfrak{q}$ runs through the minimal primes of A. □

Corollary 5.14. *If $\ell\geq q^{n!}$, then $R^{\square,(1,\dots,1)}(\bar\rho)=R^{\square}(\bar\rho)$. In particular, $R^{\square,(1,\dots,1)}(\bar\rho)$ is reduced (see Theorem 5.1).*

Proof. By Lemma 5.13 (together with Theorem 5.1), we see that the identity map on $R^{\square}(\bar\rho)$ factors through $R^{\square,(1,\dots,1)}(\bar\rho)$. On the other hand, $R^{\square,(1,\dots,1)}(\bar\rho)$ is by definition a quotient of $R^{\square}(\bar\rho)$. Thus, we have found a surjective endomorphism of $R^{\square,(1,\dots,1)}(\bar\rho)$ (which must then be an isomorphism, as the rings in question are noetherian) which factors via $R(\bar\rho)^{\square}$. This proves the claim. □

6. On automorphic forms on unitary groups

6A. ***The group $\mathcal{G}_n$.*** For $n\in\mathbb{N}$ recall from [Clozel et al. 2008, Section 2.1] the definition of the group scheme $\mathcal{G}_n$ over $\mathbb{Z}$ and the multiplier character $\mathtt{m}:\mathcal{G}_n\to\mathrm{GL}_1$. We write $\mathcal{G}_n^0$ for the connected component of the identity and $\mathfrak{g}_n$ for the Lie algebra of $\mathcal{G}_n$ (where we differ in notation from [loc. cit.]). We have $\mathcal{G}_n^{\mathtt{der}}\cong\mathrm{GL}_n$ and $\mathcal{G}_n^{\mathtt{ab}}\cong\mathrm{GL}_1\times\mathbb{Z}/2\mathbb{Z}$. If F is a CM-field with totally real subfield F^+, recall in particular the connection between GL_n-valued conjugate self-dual representations of Gal_F and $\mathcal{G}_n$-valued representations of Gal_{F^+}; see [loc. cit., Lemma 1.1.4] or [Gee 2011, Lemma 5.1.1].

We will be particularly interested in deformations of $\mathcal{G}_n$-valued residual representations. In the local split case, there is a substantial simplification possible: Let k be a finite field and let $\bar\rho$ be a GL_n-valued representation of Gal_F, let $\bar\chi$ a character such that $\bar\chi\bar\rho^\vee\cong\bar\rho^c$ and let $\bar r$ be the associated $\mathcal{G}_n(k)$-valued representation of Gal_{F^+}. Moreover, let Λ be the ring of integers of a finite extension of the quotient field of $W(k)$. The following proposition now follows easily from the definitions:

Proposition 6.1. *Let v be a place of F^+ which splits as $\tilde v\tilde v^c$ in F. Denote $\bar r_v:=\bar r\mid\mathrm{Gal}_{F_v^+}$ and $\bar\rho_{\tilde v}:=\bar\rho\mid\mathrm{Gal}_{F_{\tilde v}}$. Fix a lift $\chi_v:\mathrm{Gal}_{F_v^+}\to\Lambda^\times$ of $\mathtt{m}\circ\bar r_v$. Then*

$$R_\Lambda^{(\square),\chi_v}(\bar r_v)\cong R_\Lambda^{(\square)}(\bar\rho_{\tilde v}),\quad H^i(F_v^+,\mathfrak{g}_n^{\mathtt{der}})\cong H^i(F_{\tilde v},\mathfrak{gl}_n)\quad\text{and}\quad Z^1(F_v^+,\mathfrak{g}_n^{\mathtt{der}})\cong Z^1(F_{\tilde v},\mathfrak{gl}_n).$$

This observation allows us to define local conditions for deformations of $\bar r$ at split places by GL_n-valued local conditions. In order to make this precise, let $\Sigma\subset\mathrm{Pl}_{F^+}^{\mathrm{fin}}$ be a finite set of places and assume that any place in Σ splits as $v=\tilde v\tilde v^c$ in the extension $F\mid F^+$ (so, in particular, we fix a place $\tilde v$ above v). Moreover, assume that $\bar r$ is unramified outside Σ, i.e., factors through $\mathrm{Gal}_{F^+,\Sigma}$. We set $\tilde\Sigma:=\{\tilde v\mid v\in\Sigma\}$. Fix a character $\chi:\mathrm{Gal}_{F^+,\Sigma}\to\Lambda^\times$ lifting $\mathtt{m}\circ\bar r$. Moreover, for each $\tilde v\in\tilde\Sigma$ fix a deformation condition D_v of the GL_n-valued representation $\bar\rho_{\tilde v}$.

Definition 6.2 (Deformation problem, following [Clozel et al. 2008]). The collection

$$\mathscr{S} = (F \mid F^+, \Sigma, \tilde{\Sigma}, \Lambda, \bar{r}, \chi, \{D_v\}_{v\in\Sigma}),$$

parametrizing deformations r of $\bar{r}$ to $\mathcal{C}_\Lambda$ which fulfill $\mathtt{m}\circ r = \chi$, which are unramified outside Σ and fulfill D_v (via Proposition 6.1) at $v \in \Sigma$, defines a global deformation condition.

We end this section by a remark on the conventions for multiple framings, in which we differ from [Clozel et al. 2008]. For this, let $T \subset \Sigma$ be a nonempty subset and recall our Definition 3.17 for the multiply framed deformation functor $D_\Lambda^{\square_T,\mathcal{S}}(\bar{r})$ and its representing object $R_\Lambda^{\square_T,\mathcal{S}}(\bar{r})$. Comparing this with the functor and representing object considered in [loc. cit., Definition 2.2.7], which we denote by $D_\Lambda^{\boxdot_T,\mathcal{S}}(\bar{r})$ and $R_\Lambda^{\boxdot_T,\mathcal{S}}(\bar{r})$, we easily get the following observation:

Proposition 6.3. *$D_\Lambda^{\square_T,\mathcal{S}}(\bar{r})$ is representable if and only if $D_\Lambda^{\boxdot_T,\mathcal{S}}(\bar{r})$ is representable, and in this case we have*

$$R_\Lambda^{\square_T,\mathcal{S}}(\bar{r}) \cong R_\Lambda^{\boxdot_T,\mathcal{S}}(\bar{r})[\![X_1, \dots, X_{\#T}]\!].$$

6B. ***Automorphic forms and Hecke algebras.*** For this subsection, let us assume that the extension $F \mid F^+$ is unramified at all finite places and, in case n is even, that $\frac{n}{2}[F^+ : \mathbb{Q}]$ is even. This allows us to fix a definite unitary group H over $\mathcal{O}_{F^+}$, as considered in [Guerberoff 2011, Section 2.11] or [Geraghty 2010, Section 1.1], whose key properties we recall here:

- The extension of scalars of H to F^+ is an outer form of GL_n/F^+, which becomes isomorphic to GL_n/F after extending scalars to F.
- H is quasisplit at every finite place of F^+.
- H is totally definite, i.e., $H(F^+_\infty)$ is compact and $H(F^+_v) \cong U_n(\mathbb{R})$ for all infinite places v of F^+.
- For any finite place v of F^+ which splits as $\tilde{v}\tilde{v}^c$ in F, we can choose an isomorphism $\iota_{\tilde{v}} \colon H(F^+_v) \to \mathrm{GL}_n(F_{\tilde{v}})$ whose restriction to $H(\mathcal{O}_{F^+_v})$ provides an isomorphism $H(\mathcal{O}_{F^+_v}) \cong \mathrm{GL}_n(\mathcal{O}_{F_{\tilde{v}}})$.

Level subgroups. Let us fix a finite subset $\mathcal{T} \subset \mathrm{Pl}^{\mathrm{fin}}_{F^+}$ such that each $v \in \mathcal{T}$ splits as $\tilde{v}\tilde{v}^c$ in F. For the remainder of this section, the letter U will denote an open compact subgroup of $H(\mathbb{A}^\infty_{F^+})$. For later applications, we will be particularly interested in the choice $U_\mathcal{T} := \prod_{v\in\mathrm{Pl}^{\mathrm{fin}}_{F^+}} U_v$ with:

- If v is not split in $F \mid F^+$, then U_v is a hyperspecial maximal compact subgroup of $H(F^+_v)$.
- If $v \notin \mathcal{T}$ splits, then $U_v = H(\mathcal{O}_{F^+_v})$.
- If $v \in \mathcal{T}$, then $U_v = \iota_{\tilde{v}}^{-1}(\mathrm{Iw})$, where $\mathrm{Iw} \subset \mathrm{GL}_n(\mathcal{O}_{F_{\tilde{v}}})$ denotes the Iwahori subgroup.

We remark that in many articles (e.g., [Clozel et al. 2008]) the set $\mathcal{T}$ is enlarged by a choice of auxiliary places at which a suitable level condition is imposed. Our arguments don't require such auxiliary places.

Weights. Recall the parametrization of complex and ℓ-adic representations of unitary and general linear groups, e.g., from [Guerberoff 2011]:

- To a tuple $\boldsymbol{\omega} = (\underline{\omega}_\tau) \in (\mathbb{Z}^{n,+})^{\mathrm{Hom}(F^+,\mathbb{R})}$ we associate the representation

$$\xi_{\boldsymbol{\omega}} : H(F_\infty^+) = \prod_{\tau\in\mathrm{Hom}(F^+,\mathbb{R})} H(F_\tau^+) \cong \prod_{\tau\in\mathrm{Hom}(F^+,\mathbb{R})} U_n(\mathbb{R}) \xrightarrow{\varphi} \prod_{\tau\in\mathrm{Hom}(F^+,\mathbb{R})} \mathrm{GL}_n(W_{\underline{\omega}_\tau}) \subset \mathrm{GL}_n(W_{\boldsymbol{\omega}}),$$

where $W_{\boldsymbol{\omega}} = \otimes_\tau W_{\underline{\omega}_\tau}$ and where φ is the product of the highest weight representations $W_{\underline{\omega}_\tau}$ attached to the weight $\underline{\omega}_\tau$ (see e.g., [Bellaïche and Chenevier 2009; Guerberoff 2011; Geraghty 2010]).

- Let ℓ be a rational prime such that every place v of F^+ above ℓ splits in $F \mid F^+$ and fix for each such v a place $\tilde{v}$ of F above v. Let $\mathcal{K}$ be a finite extension of $\mathbb{Q}_\ell$ which is F-big enough and let $\boldsymbol{\omega} = (\underline{\omega}_\tau) \in (\mathbb{Z}^{n,+})^{\mathrm{Hom}(F,\mathcal{K})}$. To each $\tau \in \mathrm{Hom}(F,\mathcal{K})$ we can associate a place v of F^+ above ℓ for which we have just fixed a place $\tilde{v}$. Denote this assignment $\mathrm{Hom}(F,\mathcal{K}) \to \Omega_\ell^F$ by $\tau \mapsto w_\tau$. Let

$$\begin{aligned}\xi_{\boldsymbol{\omega}}^{\mathcal{K}} : \prod_{v\in\Omega_\ell^F} H(F_v^+) \cong \prod_{v\in\Omega_\ell^F} \mathrm{GL}_n(F_{\tilde{v}}) \\ \xrightarrow{\prod d_v} \prod_{v\in\Omega_\ell^F} \prod_{\substack{\tau\in\mathrm{Hom}(F,\mathcal{K}) \\ w_\tau=\tilde{v}}} \mathrm{GL}_n(F_{\tilde{v}}) = \prod_{\tau\in\mathrm{Hom}(F,\mathcal{K})} \mathrm{GL}_n(F_{\tilde{v}}) \\ \xrightarrow{\psi} \prod_{\tau\in\mathrm{Hom}(F,\mathcal{K})} \mathrm{GL}_n(W_{\underline{\omega}_\tau}^{\mathcal{K}}) \subset \mathrm{GL}_n(W_{\boldsymbol{\omega}}^{\mathcal{K}})\end{aligned}$$

be the representation where each d_v is the diagonal embedding, where $W_{\boldsymbol{\omega}}^{\mathcal{K}} = \otimes_\tau W_{\underline{\omega}_\tau}^{\mathcal{K}}$ and where ψ is the product of the highest weight representations $W_{\underline{\omega}_\tau}^{\mathcal{K}}$ attached to the weight $\underline{\omega}_\tau$. The representation $\xi_{\boldsymbol{\omega}}^{\mathcal{K}}$ admits an integral model over $\mathcal{O}_{\mathcal{K}}$, whose underlying finite free $\mathcal{O}_{\mathcal{K}}$-module we denote by $M_{\boldsymbol{\omega}}^{\mathcal{O}_{\mathcal{K}}}$.

Automorphic forms. We denote by

$$\mathcal{A}(H) = \bigoplus_\pi \pi^{m(\pi)}$$

the space of (complex) automorphic forms on H, which decomposes into isomorphism classes of irreducible representations of $H(\mathbb{A}_{F^+})$, each occurring with finite multiplicity $m(\pi)$ (see e.g., [Guerberoff 2011]).

Definition 6.4 (Vector-valued automorphic form). Let $\boldsymbol{\omega} \in (\mathbb{Z}^{n,+})^{\mathrm{Hom}(F^+,\mathbb{R})}$ be a weight, then we denote by $\mathcal{S}_{\boldsymbol{\omega}}$ the space of locally constant functions $f : H(\mathbb{A}_{F^+}^\infty) \to W_{\boldsymbol{\omega}}^\vee$ which fulfill

$$f(\gamma.h) = \gamma_\infty.f(h) \quad \forall h \in H(\mathbb{A}_{F^+}^\infty), \gamma \in H(F^+).$$

(We denote by γ_∞ the image of γ under the canonical embedding $H(F^+) \to H(F_\infty^+)$.) $H(\mathbb{A}_{F^+}^\infty)$ acts on $\mathcal{S}_{\boldsymbol{\omega}}$ by right translation, and for a level subgroup U we denote by $\mathcal{S}_{\boldsymbol{\omega}}(U)$ the space of U-fixed vectors.

There exists an $H(\mathbb{A}_{F^+}) = H(\mathbb{A}_{F^+,\infty}) \times H(\mathbb{A}_{F^+}^\infty)$-equivariant decomposition

$$\mathcal{A}(H) = \bigoplus_{\boldsymbol{\omega}} W_{\boldsymbol{\omega}} \otimes \mathcal{S}_{\boldsymbol{\omega}}.$$

Thus we can associate with $f \in \mathcal{S}_{\boldsymbol{\omega}}$ the automorphic representation $\langle f \rangle$ that is uniquely characterized by the condition that it contains all vectors of $W_{\boldsymbol{\omega}} \otimes f$. The main feature of the group H is the existence of avatars.

Theorem 6.5. *Let Π be a RACSDC automorphic representation of $\mathrm{GL}_n(\mathbb{A}_F)$ of weight $\boldsymbol{\omega} \in (\mathbb{Z}^{n,+})^{\mathrm{Hom}(F,\mathbb{C})}$ in the sense of* [Clozel et al. 2008, Section 4]. *Then there exists an automorphic representation π_0 of $H(\mathbb{A}_{F^+})$ such that Π is a base change of π_0:*

- *For each archimedean place v of F^+ and each place $\tilde{v}$ of F above v, we have $\pi_{0,v} \cong \xi_{\underline{\omega}_{\tilde{v}}}$.*
- *For each finite place v of F^+ which splits as $\tilde{v}\tilde{v}^c$ in F, $\Pi_{\tilde{v}}$ is the local base change of $\pi_{0,v}$.*
- *If v is a finite place of F^+ which stays inert in F and for which Π_v is unramified, then π_v has a fixed vector for a maximal hyperspecial compact subgroup of $H(F_v^+)$.*

Proof. See [Guerberoff 2011, Theorem 2.2] and [Geraghty 2010, Lemma 2.2.7]. □

Hecke algebras. We continue to consider a fixed set of places $\mathcal{T}$ as above (with corresponding level subgroup $U = U_{\mathcal{T}}$) and a weight $\boldsymbol{\omega}$. For $j \in \{1, \dots, n\}$ and for w a place of F which is split over F^+ and does not divide an element of $\mathcal{T}$, we consider the following Hecke operator (acting on $\mathcal{S}_{\boldsymbol{\omega}}(U)$):

$$T_{F_w}^{(j)} = \left[U.\iota_w^{-1} \begin{pmatrix} \varpi_{F_w} \mathbf{1}_j & 0 \\ 0 & \mathbf{1}_{n-j} \end{pmatrix} .U \right]$$

For a finite set $\mathcal{T}' \subset \mathrm{Pl}_{F^+}^{\mathrm{fin}}$ containing $\mathcal{T}$ and a subring $\mathscr{R}$ of $\mathbb{C}$ we define the Hecke algebra

$${}^{\mathscr{R}}\boldsymbol{T}_{\boldsymbol{\omega}}^{\mathcal{T}'}(U) := \mathrm{im}\big(\mathscr{R}[T_{F_w}^{(j)} \mid j \in \{1, \dots, n\}, w \in \mathrm{Pl}_F^{\mathtt{split},\mathcal{T}'}] \to \mathrm{End}_{\mathbb{C}}(\mathcal{S}_{\boldsymbol{\omega}}(U))\big),$$

where $\mathrm{Pl}_F^{\mathtt{split},\mathcal{T}'}$ denotes the set of places of F which are split over F^+ and which do not divide an element of $\mathcal{T}'$. Besides the case $\mathscr{R} = \mathbb{Z}$ we will be interested in $\mathscr{R} = \mathcal{E}_f$ (the coefficient field of an eigenform f with respect to ${}^{\mathbb{Z}}\boldsymbol{T}_{\boldsymbol{\omega}}^{\mathcal{T}}(U)$)) and in $\mathscr{R} = \mathcal{E}(U) = \prod_f \mathcal{E}_f$, where the product (i.e., the field compositum operation) runs through all eigenforms of $\mathcal{S}_{\boldsymbol{\omega}}(U)$. We note the following well-known facts: There are only finitely many (one-dimensional) eigenspaces $\mathbb{C}.f_1, \dots, \mathbb{C}.f_r$ contained in $\mathcal{S}_{\boldsymbol{\omega}}(U)$, so $\mathcal{E}(U)$ is a number field. Moreover, $\mathcal{S}_{\boldsymbol{\omega}}(U)$ admits a basis of eigenforms, i.e., we can choose the f_i such that

$$\mathcal{S}_{\boldsymbol{\omega}}(U) \cong \mathbb{C}.f_1 \oplus \cdots \oplus \mathbb{C}.f_r$$

as a $\boldsymbol{T}_{\boldsymbol{\omega}}^{\mathcal{T}}(U)$-module (see decomposition (3.1.1) of [Guerberoff 2011]). By mapping a Hecke operator to its f-eigenvalue, any eigenform $f \in \mathcal{S}_{\boldsymbol{\omega}}(U)$ gives rise to a $\mathbb{Z}$-algebra-homomorphism

$$\varphi_f : {}^{\mathbb{Z}}\boldsymbol{T}_{\boldsymbol{\omega}}^{\mathcal{T}}(U) \to \mathcal{E}(U), \quad T_{F_w}^{(j)} \mapsto a_f(T_{F_w}^{(j)})$$

and it can be shown that $\mathrm{im}(\varphi_f) \subset \mathcal{O}_{\mathcal{E}(U)}$. The form f is uniquely characterized by φ_f (up to $\mathbb{C}$-multiples).

ℓ-adic models of automorphic forms. The following is based on Section 2.3 of [Guerberoff 2011]. For this paragraph, we fix a rational prime ℓ which does not lie below $\mathcal{T}$ and such that all places of F^+ above ℓ are split in the extension $F \mid F^+$ and consider the following setup: Let $\mathcal{K}$ be a finite extension of $\mathbb{Q}_\ell$ which is F-big enough and fix an isomorphism $\iota : \overline{\mathcal{K}} \cong \mathbb{C}$. Moreover, we fix an ℓ-adic weight $\boldsymbol{\omega}$, i.e., an element of

$$(\mathbb{Z}^{n,+})_c^{\mathrm{Hom}(F,\mathcal{K})} = \{\boldsymbol{\omega} \in (\mathbb{Z}^{n,+})^{\mathrm{Hom}(F,\mathcal{K})} \mid \underline{\omega}_{\tau^c,i} = -\underline{\omega}_{\tau,n-i+1} \forall \tau \in \mathrm{Hom}(F,\mathcal{K}), i \in \{1,\dots,n\}\}.$$

Definition 6.6. For $U \subset H(\mathbb{A}_{F^+}^\infty)$ a compact subgroup and an $\mathcal{O}_\mathcal{K}$-algebra A, suppose that either the projection of U to $H(F_\ell^+)$ is contained in $H(\mathcal{O}_{F_\ell^+})$ or that A is a $\mathcal{K}$-algebra. Then we define

$$S_{\boldsymbol{\omega}}(U, A) = \{f : H(F^+)\backslash H(\mathbb{A}_{F^+}^\infty) \to A \otimes_{\mathcal{O}_K} M_{\boldsymbol{\omega}}^{\mathcal{O}_\mathcal{K}} \mid u_\ell . f(hu) = f(h) \forall u \in U, h \in H(\mathbb{A}_{F^+}^\infty)\},$$

where u_ℓ denotes the image of u under the projection map $H(\mathbb{A}_{F^+}^\infty) \to H(F_\ell^+)$.

We are primarily interested in the case that A is $\mathcal{O}_\mathcal{K}$-flat, so that we have $S_{\boldsymbol{\omega}}(U, A) \cong A \otimes_{\mathcal{O}_\mathcal{K}} S_{\boldsymbol{\omega}}(U, \mathcal{O}_\mathcal{K})$.

The main connection with complex automorphic forms is as follows (see also [Guerberoff 2011, Section 2.3]): The isomorphism ι gives rise to a bijection $\iota_*^+ : (\mathbb{Z}^{n,+})_c^{\mathrm{Hom}(F,\mathcal{K})} \cong (\mathbb{Z}^{n,+})^{\mathrm{Hom}(F,\mathbb{R})}$, and the assignment $f \mapsto (h \mapsto \theta_{\boldsymbol{\omega}}(h_\ell . f(h)))$ provides isomorphisms of $\mathbb{C}H(\mathbb{A}_{F^+}^\infty)$-modules

$$\bigcup_U S_{\boldsymbol{\omega}}(U, \mathbb{C}) \cong \mathcal{S}_{\iota_*^+(\boldsymbol{\omega})^\vee} \quad \text{and} \quad S_{\boldsymbol{\omega}}(U, \mathbb{C}) \cong \mathcal{S}_{\iota_*^+(\boldsymbol{\omega})^\vee}(U). \tag{11}$$

(Here, $\mathbb{C}$ is understood as a $\mathcal{O}_\mathcal{K}$-algebra via ι and $\iota_*^+(\boldsymbol{\omega})^\vee$ is defined by $\iota_*^+(\boldsymbol{\omega})^\vee_{\tau,i} = -\iota_*^+(\boldsymbol{\omega})^\vee_{\tau,n+1-i}$.)

For a place w not dividing ℓ, the operators $T_{F_w}^{(j)}$ also act on $S_{\boldsymbol{\omega}}(U, \mathcal{O}_\mathcal{K}) \subset S_{\boldsymbol{\omega}}(U, \mathbb{C})$, and this action commutes with the isomorphism (11). This motivates the following definition: Let $\mathcal{T}'$ be a finite set of places of F^+ containing $\mathcal{T} \cup \Omega_\ell^{F^+}$ and let $\mathscr{R}$ be a subring of $\mathcal{O}_\mathcal{K}$, then we define the Hecke algebra

$${}^{\mathscr{R}}\mathbb{T}_{\boldsymbol{\omega}}^{\mathcal{T}'}(U) = \mathrm{im}\big(q : \mathscr{R}[T_{F_w}^{(j)} \mid j \in \{1,\dots,n\}, w \in \mathrm{Pl}_F^{\mathtt{split},\mathcal{T}'}] \to \mathrm{End}_{\mathcal{O}_\mathcal{K}}(S_{\boldsymbol{\omega}}(U, \mathcal{O}_\mathcal{K}))\big),$$

where we will often abbreviate $\mathbb{T}_{\boldsymbol{\omega}}^{\mathcal{T}'}(U) = {}^{\mathcal{O}_\mathcal{K}}\mathbb{T}_{\boldsymbol{\omega}}^{\mathcal{T}'}(U)$. If $f \in S_{\boldsymbol{\omega}}(U, \mathcal{O}_\mathcal{K})$ is an eigenform for this algebra, then we see, using the compatibility with the isomorphism (11), that the eigenvalue for a Hecke operator T is given by $\iota^{-1}(a_{\tilde{f}})$, where $\tilde{f} \in \mathcal{S}_{\iota_*^+(\boldsymbol{\omega})^\vee}(U)$ is the corresponding complex automorphic form. In other words, we can interpret the map $\varphi_{\tilde{f}}$ from above as

$$\varphi_f^\ell : {}^{\mathbb{Z}}\mathbb{T}_{\boldsymbol{\omega}}^{\mathcal{T}_\ell}(U) \to \iota(\mathcal{E}(U)) \cong \mathcal{E}(U).$$

Note that we use the bold symbol $\boldsymbol{T}$ for complex Hecke algebras and the blackboard bold symbol $\mathbb{T}$ for ℓ-adic Hecke algebras.

Fixed type Hecke algebras. Fix a finite set $\tilde{\Sigma} \subset (\mathcal{T}' - \Omega_\ell^F)$ of places of F together with a tuple $\underline{\sigma} = (\sigma_v)_{v \in \tilde{\Sigma}}$, where each σ_v is a finite-dimensional complex representation of $\mathrm{GL}_n(\mathcal{O}_{F_v})$. Let ${}_{\underline{\sigma}}S_{\boldsymbol{\omega}}(U, \mathcal{O}_\mathcal{K}) \subset S_{\boldsymbol{\omega}}(U, \mathcal{O}_\mathcal{K})$ be the subspace generated by those forms f whose complex correspondents $\tilde{f}$ fulfill the following condition for all places $v \in \tilde{\Sigma}$: If π_v denotes the local component of the automorphic representation $\pi = \langle \tilde{f} \rangle$ at v,

then $\pi_v \mid \mathrm{GL}_n(\mathcal{O}_{F_v})$ contains σ_v as a subrepresentation. Note that the $T^{(j)}_{F_w}$ (for w in $\mathrm{Pl}_F^{\mathtt{split},\mathcal{T}'}$) stabilize the subspace ${}_{\underline{\sigma}}S_{\boldsymbol{\omega}}(U,\mathcal{O}_{\mathcal{K}})$, so we can define

$$ {}^{\mathscr{R}}_{\underline{\sigma}}\mathbb{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U) = \mathrm{im}\big({}_{\underline{\sigma}}q : \mathscr{R}[T^{(j)}_{F_w} \mid j\in\{1,\dots,n\}, w\in \mathrm{Pl}_F^{\mathtt{split},\mathcal{T}'}] \to \mathrm{End}_{\mathcal{O}_{\mathcal{K}}}({}_{\underline{\sigma}}S_{\boldsymbol{\omega}}(U,\mathcal{O}_{\mathcal{K}}))\big). $$

We easily see that the assignment $q(T^{(j)}_{F_w}) \mapsto {}_{\underline{\sigma}}q(T^{(j)}_{F_w})$ defines an $\mathscr{R}$-algebra surjection ${}_{\underline{\sigma}}\theta$ from ${}^{\mathscr{R}}\mathbb{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U)$ to ${}^{\mathscr{R}}_{\underline{\sigma}}\mathbb{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U)$. We note the following (for $\mathscr{R}=\mathcal{O}_{\mathcal{K}}$):

- In the same way as for ${}^{\mathcal{O}_{\mathcal{K}}}\boldsymbol{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U)$, we can check that ${}^{\mathcal{O}_{\mathcal{K}}}_{\underline{\sigma}}\mathbb{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U)$ is free and finitely generated over $\mathcal{O}_{\mathcal{K}}$.
- Assume that ${}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U)_{\mathfrak{m}} \cong \mathcal{O}_{\mathcal{K}}$ holds for any maximal ideal $\mathfrak{m}$, then ${}^{\mathcal{O}_{\mathcal{K}}}_{\underline{\sigma}}\mathbb{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U)_{\mathfrak{n}}$ is a quotient of $\mathcal{O}_{\mathcal{K}}$ for any maximal ideal $\mathfrak{n}$. By the above bullet point, it thus follows that ${}^{\mathcal{O}_{\mathcal{K}}}_{\underline{\sigma}}\mathbb{T}^{\mathcal{T}'}_{\boldsymbol{\omega}}(U)_{\mathfrak{n}}$ is isomorphic to $\mathcal{O}_{\mathcal{K}}$ for any maximal ideal $\mathfrak{n}$.

6C. ***Attaching Galois representations to automorphic forms.*** Retain all notation from above and let $\mathfrak{m} \subset {}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U)$ be a maximal ideal.

Proposition 6.7 [Guerberoff 2011, Proposition 3.1 and 3.2]. *There exists a representation*

$$ \rho_{\mathfrak{m}} : \mathrm{Gal}_F \to \mathrm{GL}_n({}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U)_{\mathfrak{m}}) $$

with the following properties, where the first two already characterize $\rho_{\mathfrak{m}}$ uniquely:

(1) *$\rho_{\mathfrak{m}}$ is unramified at all but finitely many places. If a place v of F^+ is inert and unramified in F and if U_v is a hyperspecial maximal compact subgroup of $H(F^+_v)$, then $\rho_{\mathfrak{m}}$ is unramified above v.*

(2) *If $v \in \mathrm{Pl}^{\mathrm{fin}}_{F^+}\setminus\mathcal{T}_\ell$ splits as $\tilde{v}\tilde{v}^c$ in F, then $\rho_{\mathfrak{m}}$ is unramified at $\tilde{v}$ and*

$$ \mathrm{charPoly}(\rho_{\mathfrak{m}}(\mathrm{Frob}_{\tilde{v}})) = X^n - T^{(1)}_{\tilde{v}}X^{n-1} + \cdots + (-1)^j(\boldsymbol{N}\tilde{v})^{j(j-1)/2}T^{(j)}_{\tilde{v}}X^{n-j} + \cdots + (-1)^n(\boldsymbol{N}\tilde{v})^{n(n-1)/2}T^{(n)}_{\tilde{v}}. $$

(3) *$\bar{\rho}^c_{\mathfrak{m}} \cong \bar{\rho}_{\mathfrak{m}} \otimes \bar{\epsilon}_\ell^{1-n}$.*

(4) *Fix a set $\widetilde{\Omega}^{F^+}_\ell$ of places of F such that $\widetilde{\Omega}^{F^+}_\ell \sqcup \widetilde{\Omega}^{F^+,c}_\ell = \Omega^F_\ell$ and denote by $\tilde{I}_\ell$ the set of embeddings $F \hookrightarrow \mathcal{K}$ which give rise to an element of $\widetilde{\Omega}^{F^+}_\ell$. Suppose that $w \in \widetilde{\Omega}^{F^+}_\ell$ is unramified over ℓ, that $U_{\overline{w}} = H(\mathcal{O}_{F^+,\overline{w}})$ (for $\overline{w}\in\mathrm{Pl}_{F^+}$ the place below w) and that for each $\tau\in\tilde{I}_\ell$ above w we have*

$$ \ell - 1 - n \geq \underline{\omega}_{\tau,1} \geq \cdots \geq \underline{\omega}_{\tau,n} \geq 0. $$

Then, for each open ideal $I \subset {}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U)$ there is an object $M_{\mathfrak{m},I,w}$ of $\underline{\mathrm{MF}}_{\mathcal{O}_{F_w},\mathcal{O}_{\mathcal{K}}}$ such that

$$ (\rho_{\mathfrak{m}} \otimes_{\mathcal{O}_{\mathcal{K}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U)} {}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U)/I) \mid \mathrm{Gal}_{F_w} \cong \mathtt{G}_{F_w}(M_{\mathfrak{m},I,w}). $$

If $\mathfrak{m}$ is non-Eisenstein in the sense of [Clozel et al. 2008, Definition 3.4.3], *then $\rho_{\mathfrak{m}}$ and its reduction extend to*

$$ r_{\mathfrak{m}} : \mathrm{Gal}_{F^+} \to \mathcal{G}_n({}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U)_{\mathfrak{m}}) \quad \textit{and} \quad \bar{r}_{\mathfrak{m}} : \mathrm{Gal}_{F^+} \to \mathcal{G}_n({}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U)/\mathfrak{m}). $$

Moreover, $m \circ r_{\mathfrak{m}} = \epsilon_\ell^{1-n}\delta^{\mu_{\mathfrak{m}}}_{F\mid F^+}$ for a suitable $\mu_{\mathfrak{m}} \in \mathbb{Z}/2\mathbb{Z}$, where $\delta_{F\mid F^+}$ is the nontrivial character of $\mathrm{Gal}(F \mid F^+)$.

In this way we can associate to a RACSDC automorphic representation π of $\mathrm{GL}_n(\mathbb{A}_F)$ and a finite place λ of $\mathcal{E}(U)$ a residual representation $\bar{r}_{\pi,\lambda} : \mathrm{Gal}_{F^+} \to \mathcal{G}_n(\overline{\mathbb{F}}_{\ell(\lambda)})$. Let us assume that $\bar{r}_{\pi,\lambda}$ is absolutely irreducible for all λ in a subset of $\mathrm{Pl}^{\mathrm{fin}}_{\mathcal{E}(U)}$ of Dirichlet density 1. Then the set

$$\Lambda^1_{\mathcal{E}(U)} = \{\lambda \mid \bar{\rho}_{\pi,\lambda'} \text{ is absolutely irreducible for all } \lambda' \text{ dividing } \ell(\lambda)\}$$

has also Dirichlet density 1. In this way, we get an association from π to the compatible systems of residual Galois representations $\mathcal{R}_\pi = (\bar{r}_{\pi,\lambda})_{\lambda\in\Lambda^1_{\mathcal{E}(U)}}$ and $\mathcal{R}'_\pi = (\bar{\rho}_{\pi,\lambda})_{\lambda\in\Lambda^1_{\mathcal{E}(U)}}$.

7. Consequences from modularity lifting theorems

Let us start with the following adaption of [Khare and Wintenberger 2009a, Lemma 3.6]:

Lemma 7.1. *Let k be a finite field of characteristic ℓ, G a profinite group satisfying the ℓ-finiteness condition and $\eta : G \to \mathcal{G}_n(k)$ be an absolutely irreducible continuous representation. Let $\mathcal{F}_n(G)$ be a subcategory of deformations of η in k-algebras which defines a deformation condition. Let $\eta_{\mathcal{F}} : G \to \mathrm{GL}_n(R_{\mathcal{F}})$ be the universal deformation of η in $\mathcal{F}_n(G)$. Then $R_{\mathcal{F}}$ is finite if and only if $\eta_{\mathcal{F}}(G)$ is finite.*

Proof. The proof of Lemma 3.6 of Khare–Wintenberger goes through verbatim, except that we have to refer to [Clozel et al. 2008, Lemma 2.1.12] instead of Carayol's lemma. □

7A. ***A minimal R = T-theorem.*** Our starting point is a RACSDC automorphic representation $\pi = \langle f\rangle \subset \mathcal{S}_{\boldsymbol{\omega}}(U)$ (where $U = U_S$ for a finite set of places S of F) and a place $\lambda \in \Lambda^1_{\mathcal{E}(U)}$. Fix a finite F-big enough extension $\mathcal{K}$ of $\mathcal{E}(U)_\lambda$. We abbreviate $r, \bar{r}, \rho, \bar{\rho}$ for the associated Galois representations via Proposition 6.7 for the unique maximal ideal $\mathfrak{m}$ containing $\mathcal{O}_{\mathcal{K}} \otimes_{\mathbb{Z}} \ker \varphi^\ell_f$. We assume furthermore the following:

- All places of S_ℓ split in the extension $F \mid F^+$.
- All ramification of ρ is unipotent (this can always be achieved by a finite solvable base change).
- ρ is a minimally ramified (at all places in S) and FL-crystalline (at all places dividing ℓ) lift of $\bar{\rho}$.
- $\bar{\rho}$ is absolutely irreducible.
- $\bar{\rho}(\mathrm{Gal}_{F(\zeta_\ell)})$ is adequate in the sense of Thorne [2012, Definition 2.20]:
  - $H^1(X_\ell, k_\lambda) = 0$ and $H^1(X_\ell, \mathfrak{gl}^0_n) = 0$.
  - For any simple $k_\lambda[X_\ell]$-submodule $W \subset \mathfrak{gl}_n$, there exists a semisimple element $\sigma \in X_\ell$ with eigenvalue $\alpha \in k_\lambda$ such that $\mathrm{tr}\, e_{\sigma,\alpha} W \neq 0$. (Here, $e_{\sigma,\alpha} \in \mathfrak{gl}_n$ denotes the unique idempotent in $k_\lambda[\sigma]$ with image equal to the α-eigenspace of σ.)

Let us abbreviate $R^{(\mathtt{min}),[\mathtt{crys}]} := R^{\chi,(\mathtt{min}),[\mathtt{crys}]}_{\mathcal{O}_{\mathcal{K}},S_\ell}(\bar{r})$ for the ring parametrizing fixed-determinant deformations of $\bar{r}$ which are unramified outside S_ℓ (minimally ramified in S) and [FL-crystalline at places dividing ℓ]. Moreover, let $\mathbb{T}$ (resp. $\mathbb{T}^{\mathtt{min}}$) denote the Hecke algebra ${}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}^{S_\ell}_{\boldsymbol{\omega}}(U)_{\mathfrak{m}}$ (resp. ${}^{\mathcal{O}_{\mathcal{K}}}_{\underline{\sigma}}\mathbb{T}^{S_\ell}_{\boldsymbol{\omega}}(U)_{\mathfrak{m}}$), where $\mathfrak{m}$ is the maximal ideal such that $\bar{r}_{\mathfrak{m}} \cong \bar{r}$ and where $\underline{\sigma} = (\sigma_\nu)_{\nu\in\tilde{S}}$ is defined as follows: For each $\nu \in \tilde{S}$ we can associate an inertial type τ_ν in the same way as we did just before Theorem 5.10. To each

τ_v one can associate a representation $\sigma_v = \sigma(\tau_v)$ of $K = \mathrm{GL}_n(\mathcal{O}_{F_v})$ (which is then the K-type of the $\mathrm{GL}_n(F_v)$-representation associated to an extension of τ_v to Gal_{F_v}.) To construct $\sigma(\tau)$ (see also [Shotton 2015, Section 4.6; Bellaïche and Chenevier 2009, Section 6.5.2; Schneider and Zink 1999]):

- Consider the finite group $\mathfrak{G} = \mathrm{GL}_n(\ell(v))$ and its standard Borel subgroup $\mathfrak{B} \subset \mathfrak{G}$. Then the irreducible constituents of the (complex) representation $\mathrm{ind}_{\mathfrak{B}}^{\mathfrak{G}}(1)$ are called the unipotent representations of $\mathfrak{G}$. These representations can (canonically) be parametrized by the irreducible representations of the Weyl group $\mathcal{W}(\mathfrak{G}) \cong S_n$; see e.g., [Prasad 2014, Corollary 4.4]. The irreducible representations of S_n in turn can be parametrized by partitions of n in terms of Specht modules; see [James and Kerber 1981]. In other words, we get a canonical bijection $h : \mathcal{Y}_n \cong \mathrm{Rep}(\mathfrak{G})^{\mathtt{uni}}$, where $\mathrm{Rep}(\mathfrak{G})^{\mathtt{uni}}$ denotes the set of all unipotent representations of $\mathfrak{G}$ up to isomorphism. The map h can be explicitly described in terms of induction from Levi subgroups (see [Shotton 2015, Definition 4.34]) and sends $(1, \dots, 1)$ to the trivial representation and (n) to the Steinberg representation.
- Under the unipotent ramification assumption, the set of inertial types $\mathcal{I}^{\mathtt{uni}}$ is in bijection with the set $\mathcal{Y}_n$ of partitions of n via ∇ from Section 5D.

We have a decomposition

$$\mathrm{ind}_I^K(1) \cong \mathrm{infl}_{\mathfrak{G}}^K \mathrm{ind}_{\mathfrak{B}}^{\mathfrak{G}}(1) \cong \bigoplus_{\Pi \in \mathrm{Rep}(\mathfrak{G})^{\mathtt{uni}}} m_\Pi \, \mathrm{infl}_{\mathfrak{G}}^K(\Pi),$$

where $I \subset K$ denotes the Iwahori subgroup, $\mathrm{infl}_{\mathfrak{G}}^K$ denotes the inflation along the pro-$\ell(v)$-radical of K and where the $m_\Pi \geq 1$ are suitable multiplicities. Analogous to [Bellaïche and Chenevier 2009, Remark 6.5.2(iii)], one can thus check that the assignment $\tau \mapsto \sigma(\tau)$ is described in terms of partitions as $\tau \mapsto \mathrm{infl}_{\mathfrak{G}}^K(h \circ \nabla(\tau))$. Observe that the special case $n = 2$ is precisely [loc. cit., Remark 6.5.2(iii)] and [Shotton 2015, Example 2.17].

We stress that the notions $\mathbb{T}$ and $\mathbb{T}^{\mathtt{min}}$ depend on the choice of the place λ.

Proposition 7.2. *A map $h : \mathbb{T} \to \overline{\mathbb{Q}}_\ell$ factors through $\mathbb{T}^{\mathtt{min}}$ if and only if the concatenation*

$$h' : R^{\mathtt{crys}} \to \mathbb{T} \to \overline{\mathbb{Q}}_\ell$$

factors through $R^{\mathtt{min,crys}}$.

Proof. The map h corresponds to an automorphic form $g \in S_{\boldsymbol{\omega}}(U, \mathcal{O}_{\mathcal{K}})$ such that $\bar{r}_{\langle g \rangle} \cong \bar{r}$. By Theorem 5.10 and the above, $\rho_{\langle g \rangle, v}$ (for $v \in S$) is a minimally ramified lift of $\bar{\rho}_v$ if and only if it is of type τ_v if and only if $\langle g \rangle_v$ is of type σ_v. (If $\langle g \rangle_v$ is of type σ_v, then $\rho_{\langle g \rangle, v}$ is at most as ramified as τ_v.) Thus, h factors through $\mathbb{T}^{\mathtt{min}}$ if and only if $g \in_{\underline{\sigma}} S_{\boldsymbol{\omega}}(U, \mathcal{O}_{\mathcal{K}})$ if and only if $r_{\langle g \rangle}$ is (as a lift of $\bar{r}$) minimally ramified in S if and only if the associated map $h' : R^{\mathtt{crys}} \to \overline{\mathbb{Q}}_\ell$ factors through $R^{\mathtt{min,crys}}$. □

Theorem 7.3. *$R^{\mathtt{min,crys}}$ is finite flat over $\mathcal{O}_{\mathcal{K}}$, so in particular there exists a characteristic-0 point of* $\mathrm{Spec}\, R^{\mathtt{min,crys}}$. *Moreover, we have isomorphisms*

$$R^{\mathtt{min,crys}} \cong R^{\mathtt{min,crys}}_{\mathtt{red}} \cong \mathbb{T}^{\mathtt{min}}.$$

Proof. We first remark that $R^{\mathtt{min},\mathtt{crys}}/(\ell)$ is of finite cardinality or, equivalently (by Nakayama's lemma), that $R^{\mathtt{min},\mathtt{crys}}$ is finitely generated as a $\mathcal{O}_{\mathcal{K}}$-module. This follows directly from [Barnet-Lamb et al. 2014, Theorem 2.3.2], as we know that the local deformation rings $R^{\square,\chi_v,\mathtt{crys}}(\bar{\rho}_v)$ and $R^{\square,\chi_v,\mathtt{min}}(\bar{\rho}_v)$ are smooth, hence correspond to irreducible components of $\operatorname{Spec} R^{\square,\chi_v}(\bar{\rho}_v)$ on which the local lifts ρ_v live.

Next, we remark that by Corollary 3.21 (together with the smoothness-results Lemma 5.4 and Propositions 5.6 and 3.22, the identity $\dim(\mathfrak{gl}_n^{c_v=-1}) = n(n+1)/2$ and Remark 7.5 below) there exists a presentation

$$R^{\mathtt{min},\mathtt{crys}} \cong \mathcal{O}_{\mathcal{K}}[\![X_1,\dots,X_m]\!]/(f_1,\dots,f_m)$$

for some $m \in \mathbb{N}_0$.

Using this presentation and the finiteness of $R^{\mathtt{min},\mathtt{crys}}/(\ell)$, it follows as in the proof of Theorem 3.7 of [Khare and Wintenberger 2009a] or of Lemma 2 in Böckle's appendix to [Khare 2003] that $R^{\mathtt{min},\mathtt{crys}}$ is finite flat over $\mathcal{O}_{\mathcal{K}}$, hence free and finitely generated over $\mathcal{O}_{\mathcal{K}}$. This proves the first claim.

As a second step, we remark that any morphism $f\colon R^{\mathtt{min},\mathtt{crys}} \to \overline{\mathbb{Q}}_\ell$ factors over $\mathbb{T}^{\mathtt{min}}$: By [Barnet-Lamb et al. 2014, Theorem 2.3.1], such an f factors over the nonminimal Hecke algebra $\mathbb{T}$. Therefore Proposition 7.2 applies.

Now, $R^{\mathtt{min},\mathtt{crys}}\big[\frac{1}{\ell}\big] = R^{\mathtt{min},\mathtt{crys}} \otimes_{\mathcal{O}_{\mathcal{K}}} \mathcal{K}$ is a finite $\mathcal{K}$-algebra, hence $R^{\mathtt{min},\mathtt{crys}}\big[\frac{1}{\ell}\big]$ is Artinian; see e.g., [Atiyah and Macdonald 1969, Exercise 8.3]. Therefore, $R^{\mathtt{min},\mathtt{crys}}\big[\frac{1}{\ell}\big]$ can be decomposed into a product of finitely many local Artinian rings

$$R^{\mathtt{min},\mathtt{crys}}\left[\frac{1}{\ell}\right] \cong \oplus_{\mathfrak{p}} R^{\mathtt{min},\mathtt{crys}}\left[\frac{1}{\ell}\right]_{\mathfrak{p}}$$

and by [Allen 2016, Theorem 3.1.3] the tangent space $\mathfrak{p}/\mathfrak{p}^2$ of each $R^{\mathtt{min},\mathtt{crys}}\big[\frac{1}{\ell}\big]_{\mathfrak{p}}$ vanishes. Hence, $\mathfrak{p} = \mathfrak{p}^2$, and it follows from Nakayama's lemma that $\mathfrak{p} = 0$, i.e., that $R^{\mathtt{min},\mathtt{crys}}\big[\frac{1}{\ell}\big]_{\mathfrak{p}}$ is a field. Thus, $R^{\mathtt{min},\mathtt{crys}}\big[\frac{1}{\ell}\big]$ is a finite product of fields. The same is true for $\mathbb{T}^{\mathtt{min}}\big[\frac{1}{\ell}\big]$: As $\mathbb{T}^{\mathtt{min}}\big[\frac{1}{\ell}\big]$ is finitely generated, its Jacobson radical equals its nilradical, which vanishes because $\mathbb{T}^{\mathtt{min}}$ is reduced. Hence $\mathbb{T}^{\mathtt{min}}\big[\frac{1}{\ell}\big]$ is semisimple, i.e., a product of finitely many fields.

Consider the exact sequence

$$0 \to \ker(\varphi) \to R^{\mathtt{min},\mathtt{crys}} \xrightarrow{\varphi} \mathbb{T}^{\mathtt{min}} \to 0, \tag{12}$$

where φ denotes the canonical projection. It follows from the above observation about $R^{\mathtt{min},\mathtt{crys}}\big[\frac{1}{\ell}\big]$ and $\mathbb{T}^{\mathtt{min}}\big[\frac{1}{\ell}\big]$ together with Proposition 7.2 that $\varphi\big[\frac{1}{\ell}\big]$ is an isomorphism. Moreover, as both $R^{\mathtt{min},\mathtt{crys}}$ and $\mathbb{T}^{\mathtt{min}}$ are finite free over $\mathcal{O}_{\mathcal{K}}$, (12) is a split exact sequence of free $\mathcal{O}_{\mathcal{K}}$-modules. Hence, $\ker(\varphi) = 0$ since $\varphi\big[\frac{1}{\ell}\big]$ is an isomorphism. This completes the proof of the theorem. □

Corollary 7.4. *The following is a pushout diagram*:

$$\begin{array}{ccc} R^{\mathtt{crys}} & \longrightarrow & R^{\mathtt{min},\mathtt{crys}} \\ \downarrow & & \downarrow \\ \mathbb{T} & \longrightarrow & \mathbb{T}^{\mathtt{min}} \end{array}$$

Remark 7.5. We remark that for each $\nu \in \Omega_\infty$ the local deformation ring $R^{\square,\chi_\nu}_{W(k_\lambda)}(\bar r_{\lambda,\nu})$ is formally smooth of relative dimension $d^\square_\nu = \dim(\mathfrak{b}^{\mathtt{der}}_n)$: We get from Proposition 3.22 that $R^{\square,\chi_\nu}_{W(k_\lambda)}(\bar r_{\lambda,\nu})$ is formally smooth of relative dimension $\dim((\mathfrak{g}_n)^{c_\nu=-1}) = \dim((\mathfrak{gl}_n)^{c_\nu=-1})$, where c_ν is the nontrivial element of the decomposition group at ν. By construction (see Lemma 2.1.4 and Proposition 3.4.4 of [Clozel et al. 2008]), the image of $\bar r_\lambda(c_\nu)$ is not contained in $\mathrm{GL}_n \times \mathrm{GL}_1$. Moreover,

$$\mathfrak{m} \circ \bar r_\lambda(c_\nu) = \bar\epsilon_\ell^{1-n}(c_\nu)\delta^{\mu_\mathfrak{m}}_{F\,|\,F^+}(c_\nu) = (-1)^{\mu_\mathfrak{m}+p},$$

where $p = n+1 \pmod 2 \in \mathbb{Z}/2\mathbb{Z}$, where ϵ_ℓ denotes the cyclotomic character (sending c_ν to -1), where $\delta_{F\,|\,F^+}$ denotes the nontrivial character of $\mathrm{Gal}(F \mid F^+)$ and where $\mu_\mathfrak{m}$ is a suitable element of $\mathbb{Z}/2\mathbb{Z}$. As in [Thorne 2012, Corollary 6.9], we get $\mu_\mathfrak{m} \equiv n \pmod 2$, so we have $\mathfrak{m}\circ\bar r_\lambda(c_\nu) = -1$, independent of the parity of n. Using [Clozel et al. 2008, Lemma 2.1.3], this implies $\dim((\mathfrak{gl})^{c_\nu=-1}_n) = n(n+1)/2 = \dim(\mathfrak{b}^{\mathtt{der}}_n)$.

7B. *A $\boldsymbol{T} = \boldsymbol{O}$-theorem.* Let $\mathcal{E} \supset \mathcal{E}(U)$ be a number field in $\mathbb{C}$ with ring of integers $\mathcal{O}_\mathcal{E}$. For each $\lambda \in \mathrm{Pl}^{\mathrm{fin}}_\mathcal{E}$ such that $\ell := \ell(\lambda) \gg 0$, let us fix an F-big enough extension $\mathcal{K}_\lambda$ of $\mathcal{E}_\lambda$ and let us abbreviate

$$\boldsymbol{T} := {}^{\mathcal{O}_\mathcal{E}}\boldsymbol{T}^{\mathcal{T}}_{\boldsymbol{\omega}}(U) \text{ and } \boldsymbol{T}_\lambda := \mathcal{O}_{\mathcal{K}_\lambda} \otimes_{\mathcal{O}_\mathcal{E}} \boldsymbol{T} \cong {}^{\mathcal{O}_{\mathcal{K}_\lambda}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}_\ell}(U).$$

Observe the following about the isomorphism on the right-hand side: Using that $\mathcal{S}_{\boldsymbol{\omega}}(U)$ admits a basis of eigenforms, we can embed $\boldsymbol{T}$ into a product of finitely many $\mathcal{O}_{\mathcal{E}(U)}$. Hence, $\boldsymbol{T}$ is finitely generated as a $\mathbb{Z}$-module, hence as a $\mathbb{Z}$-algebra. It follows that there exists a Sturm-like bound $C \in \mathbb{N}$ such that $\boldsymbol{T}$ is already generated by those $T^{(j)}_{F_w}$ with $\ell(w) \le C$. Hence, using the compatibility from (11), we get

$$\mathcal{O}_{\mathcal{K}_\lambda} \otimes_{\mathcal{O}_\mathcal{E}} \boldsymbol{T} \cong \mathcal{O}_{\mathcal{K}_\lambda} \otimes_{\mathcal{O}_\mathcal{E}} {}^{\mathcal{O}_\mathcal{E}}\boldsymbol{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}}(U) \cong {}^{\mathcal{O}_{\mathcal{K}_\lambda}}\mathbb{T}^{\mathcal{T}_\ell}_{\boldsymbol{\omega}_\ell}(U)$$

as long as $\ell > C$. Then we have:

Lemma 7.6. *For almost all λ (the failure set depending only on $\boldsymbol{T}$), $\boldsymbol{T}_\lambda$ decomposes as a product of finitely many complete discrete valuation rings, finite over $\mathbb{Z}_\ell$.*

Proof. First, we see that $\boldsymbol{T}$ is an order in $\boldsymbol{T} \otimes_{\mathcal{O}_\mathcal{E}} \mathcal{E} \cong k_1 \times \cdots \times k_m$, where the k_i denote suitable number fields (containing E) and the decomposition follows because $\boldsymbol{T} \otimes_{\mathcal{O}_\mathcal{E}} \mathcal{E}$ is reduced (as already remarked). Hence, $\boldsymbol{T}$ is contained in the maximal order $\oplus_{i=1}^m \mathcal{O}_{k_i}$. It follows that there exists a suitable $N \in \mathbb{N}$ such that:

- $\boldsymbol{T}\big[\frac{1}{N}\big] \cong \oplus_{i=1}^m \mathcal{O}_{k_i}\big[\frac{1}{N}\big]$.
- for any λ with $\ell(\lambda) \nmid N$, we have $\boldsymbol{T}_\lambda \cong \boldsymbol{T}\big[\frac{1}{N}\big]_\lambda := \mathcal{O}_{\mathcal{K}_\lambda} \otimes_{\mathcal{O}_\mathcal{E}} \boldsymbol{T}\big[\frac{1}{N}\big]$.

Thus, for those λ we get an isomorphism $\boldsymbol{T}_\lambda \cong \oplus_{i=1}^m \mathcal{O}_{\mathcal{K}_\lambda} \otimes_{\mathcal{O}_\mathcal{E}} \mathcal{O}_{k_i}$. As each factor itself is a product of complete discrete valuation rings (see, e.g., [Serre 1979, Chapter 2, Section 3, Theorem 1(ii)]), the lemma follows. □

Because we assumed that $\mathcal{E}$ contains all Hecke eigenvalues, in fact all the fields k_i in the above proof are equal to $\mathcal{E}$. Hence, for almost all ℓ, the above lemma implies that $\boldsymbol{T}_\lambda$ is isomorphic to a product of finitely many copies of $\mathcal{O}_{\mathcal{K}_\lambda}$. Thus, we get:

Corollary 7.7. *For almost all λ and all maximal ideals $\mathfrak{m} \subset \boldsymbol{T}_\lambda$, we have an isomorphism $\boldsymbol{T}_{\lambda,\mathfrak{m}} \cong \mathcal{O}_{\mathcal{K}_\lambda}$.*

7C. *An $R = R^{\mathtt{min}}$-theorem.* We retain all notation from the above and start with a preparatory corollary (to Corollary 5.14):

Corollary 7.8. *For almost all λ for which $\bar\rho_\lambda$ is absolutely irreducible, we have $R^{\mathtt{crys},(1,\dots,1)} = R^{\mathtt{crys}}$.*

Proof. Let $m := \max\{p \in \mathbb{N} \mid p \text{ prime}, \nu \mid p \text{ for some } \nu \in S\}$. Then, for all λ with $\ell(\lambda) > m^{n!}$, the claim follows directly from Corollary 5.14. □

Moreover, we need a congruence argument: First, recall that the Hecke algebra ${}^{\mathcal{O}_\mathcal{K}}\mathbb{T}^{S_\ell}_{\boldsymbol{\omega}}(U)$ acts semisimply on $\mathcal{S}_{\boldsymbol{\omega}}(U)$, so the space $\mathcal{S}_{\boldsymbol{\omega}}(U)$ decomposes into finitely many eigenspaces. For the following, let us consider *congruences*, by what we mean triples (H_1, H_2, ℓ), where $H_1 \neq H_2$ are two Hecke eigenspaces and where ℓ is a rational prime such that there exists an isomorphism $\bar\rho_{f_1,\lambda_1} \otimes \overline{\mathbb{F}}_\ell \cong \bar\rho_{f_2,\lambda_2} \otimes \overline{\mathbb{F}}_\ell$ for some choice of forms $f_i \in H_i$ and of places λ_i of the corresponding coefficient fields fulfilling $\ell(\lambda_i) = \ell$.

Proposition 7.9. *There exist only finitely many such congruences in $\mathcal{S}_{\boldsymbol{\omega}}(U)$.*

Proof. We easily see that a congruence (H_1, H_2, ℓ) corresponds to two distinct minimal prime ideals $\mathfrak{p}_{f_1}, \mathfrak{p}_{f_2}$ of $\boldsymbol{T}$ for which there exists a maximal ideal $\mathfrak{m} \subset \boldsymbol{T}$ which contains ℓ, $\mathfrak{p}_{f_1}$ and $\mathfrak{p}_{f_2}$. It follows from the finite flatness of $\boldsymbol{T}$ over $\mathbb{Z}$ that for given eigenforms f_1, f_2 there exist only finitely many maximal ideals containing $\mathfrak{p}_{f_1}$ and $\mathfrak{p}_{f_2}$. Thus, the claim follows immediately from the finite-dimensionality of the space of automorphic representations of given level and weight. □

Theorem 7.10. *For almost all λ for which $\bar\rho_\lambda$ is absolutely irreducible, we have*

$$R^{\mathtt{crys}} \cong R^{\mathtt{min},\mathtt{crys}}.$$

Proof. We apply the proof of Theorem 7.3, where we replace $R^{\mathtt{min},\mathtt{crys}}$ by $R^{\mathtt{crys}}$ and $\mathbb{T}^{\mathtt{min}}$ by $R^{\mathtt{min},\mathtt{crys}}$:

Let us first show that $R^{\mathtt{crys}}/(\ell)$ is of finite cardinality for almost all ℓ: By Nakayama's Lemma, this is equivalent to $R^{\mathtt{crys}}$ being finitely generated as a W-module. So consider the exact sequence

$$\mathrm{Nil}(R^{\mathtt{crys}})/(\ell) \to R^{\mathtt{crys}}/(\ell) \to R^{\mathtt{crys}}_{\mathtt{red}}/(\ell) \to 0.$$

We can assume that the nilradical $\mathrm{Nil}(R^{\mathtt{crys}})$ is finitely generated as a W-module: The filtration quotients $\mathrm{Nil}(R^{\mathtt{crys}})^i/\mathrm{Nil}(R^{\mathtt{crys}})^{i+1}$ are finitely generated $R^{\mathtt{crys}}$-modules (by Noetherianness) on which $\mathrm{Nil}(R^{\mathtt{crys}})$ operates trivially, hence finitely generated $R^{\mathtt{crys}}_{\mathtt{red}}$-modules. Assuming that $R^{\mathtt{crys}}_{\mathtt{red}}$ is finitely generated as a W-module, it follows that each filtration quotient is a finitely generated W-module. But, again by Noetherianness, $\mathrm{Nil}(R^{\mathtt{crys}})^i$ vanishes for $i \gg 0$. Hence, $\mathrm{Nil}(R^{\mathtt{crys}})$ is a finitely generated W-module. Thus, it remains to show that $R^{\mathtt{crys}}_{\mathtt{red}}$ is a finitely generated W-module. By Corollary 7.8, we can apply [Guerberoff 2011, Theorem 4.1] which yields the existence of a suitable finite extension $L^+ \mid F^+$, unramified at all places above ℓ, such that the ring $R^{(1,\dots,1),\mathtt{crys}}_{L^+,\mathtt{red}} := R^{\chi_{L^+},(1,\dots,1),\mathtt{crys}}_{W,\tilde S_\ell,\mathtt{red}}(\bar r_{L^+})$ is isomorphic to a suitable Hecke algebra $\mathbb{T}$ (acting on automorphic forms on a unitary group over $\mathbb{A}_{F^+}$), hence that $R^{(1,\dots,1),\mathtt{crys}}_{L^+,\mathtt{red}}$ is finitely generated over W. In order to use this result, we apply the approach of

[Khare and Wintenberger 2009a, proof of Proposition 3.8]: First, we remark that it is sufficient to show that the mod-ℓ reduction $\bar{r}^{(1,\dots,1),\mathtt{crys}}$ of the universal deformation

$$r^{(1,\dots,1),\mathtt{crys}}\colon G_{F^+} \to \mathcal{G}_n(R^{(1,\dots,1),\mathtt{crys}})$$

has finite image in order to deduce finiteness of $R^{(1,\dots,1),\mathtt{crys}}/(\ell)$, using Lemma 7.1. On the other hand, the image of the reduction of the universal deformation $r_{L^+}^{(1,\dots,1),\mathtt{crys}}$, parametrizing crystalline deformations of $\bar{r} \mid \mathrm{Gal}_{L^+}$ fulfilling the Taylor condition at $\tilde{S}$, is finite (as shown above). As $r^{(1,\dots,1),\mathtt{crys}} \mid G_{L^+}$ is a crystalline deformation of $\bar{r} \mid G_{L^+}$ (see Lemma 5.5 and Proposition 5.12), we get a morphism

$$\varphi\colon R_{L^+}^{(1,\dots,1),\mathtt{crys}}/(\ell) \to R^{(1,\dots,1),\mathtt{crys}}/(\ell)$$

such that $\bar{r}^{(1,\dots,1),\mathtt{crys}} \mid G_{L^+} = \varphi \circ \bar{r}_{L^+}^{(1,\dots,1),\mathtt{crys}}$. It follows that $\bar{r}^{(1,\dots,1),\mathtt{crys}} \mid G_{L^+}$ has finite image, hence (as $[L:F] < \infty$) that $\bar{r}^{(1,\dots,1),\mathtt{crys}}$ has finite image. As $R^{\mathtt{crys}}/(\ell)$ is a quotient of $R^{(1,\dots,1),\mathtt{crys}}/(\ell)$, the former is finitely generated, as claimed.

Next, we remark that by Corollary 3.21 (together with the smoothness-results Lemma 5.4, Theorem 5.1, Proposition 3.22, the identity $\dim(\mathfrak{gl}_n^{c_v=-1}) = n(n+1)/2$ and Remark 7.5 above) there exists a presentation

$$R^{\mathtt{crys}} \cong \mathcal{O}_{\mathcal{K}}[\![X_1,\dots,X_m]\!]/(f_1,\dots,f_m).$$

for some $m \in \mathbb{N}_0$.

Using this presentation and the finiteness of $R^{\mathtt{crys}}/(\ell)$, it follows as in the proof of Theorem 3.7 of [Khare and Wintenberger 2009a] or of Lemma 2 in Böckle's appendix to [Khare 2003] that $R^{\mathtt{crys}}$ is finite flat over $\mathcal{O}_{\mathcal{K}}$, hence free and finitely generated over $\mathcal{O}_{\mathcal{K}}$. This proves the first claim.

Moreover, we claim that for almost all λ, any morphism $R^{\mathtt{crys}} \to \overline{\mathbb{Q}}_\ell$ factors over $R^{\mathtt{min},\mathtt{crys}}$. Using automorphy lifting, this claim can be restated as follows: For almost all λ, the following holds: For any automorphic form g whose associated Galois representation $\rho_{g,\lambda}$ reduces to $\bar{\rho}_\lambda$, $\rho_{g,\lambda}$ is a minimally ramified lift of $\bar{\rho}_\lambda$. Now, let λ be a place such that this statement fails. Then, as there always exists a minimally ramified lift of $\bar{\rho}_{g,\lambda}$ with a corresponding automorphic form f (see Theorem 7.3), we get a congruence $({}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}_{\boldsymbol{\omega}}^{S_\ell}(U).f, {}^{\mathcal{O}_{\mathcal{K}}}\mathbb{T}_{\boldsymbol{\omega}}^{S_\ell}(U).g, \ell(\lambda))$. Thus, the claim follows from Proposition 7.9.

This completes the proof of the theorem. □

Let us close with the following corollary (to Theorem 7.10), giving a local $R = R^{\mathtt{min}}$ result:

Corollary 7.11. *For almost all* λ, $R^{\square,\chi_v,\mathtt{min}}(\bar{\rho}_{\lambda,v}) \cong R^{\square,\chi_v}(\bar{\rho}_{\lambda,v})$ *holds for any* $v \in S$.

Proof. Otherwise, there would be a nonminimal component of $R^{\square,\chi_v,\mathtt{min}}(\bar{\rho}_{\lambda,v})$. By [Barnet-Lamb et al. 2014, Theorem 4.3.1], there would be a lift of $\bar{\rho}_{\lambda,v}$ whose local representation at v lies on this component, in contradiction to Theorem 7.10. □

8. Unobstructedness for RACSDC automorphic representations

We are now in a position to state and prove our main result. For this, let π be a RACSDC automorphic representation of $\mathrm{GL}_n(\mathbb{A}_F)$ with ramification set S. To π we can attach a compatible system $\mathcal{R}_\pi =$

$(\bar{r}_\lambda)_{\lambda\in\Lambda^1_{\mathcal{E}_\Pi}}$ where $\Pi\subset\mathcal{S}_{\boldsymbol{\omega}}(U)$ (for a suitable weight $\boldsymbol{\omega}$ and level $U=U_S$). Here, $\mathcal{E}_\Pi$ denotes the number field generated by all Hecke eigenvalues of Π, $\Lambda^1_{\mathcal{E}_\Pi}\subset \mathrm{Pl}_{\mathcal{E}_\Pi}$ denotes the set of places for which $\bar{\rho}_\lambda$ is absolutely irreducible and we assume the following:

Assumption 8.1. (Irreducibility): The set $\Lambda^1_{\mathcal{E}_\Pi}\subset \mathrm{Pl}_{\mathcal{E}_\Pi}$ has Dirichlet density 1.

(No consecutive weights): The multisets of Hodge–Tate weights HT_τ of the system $\mathcal{R}_\pi$ fulfill (for all embeddings τ) the condition from Theorem 5.2: If two numbers a, b occur in HT_τ, then $|a-b|\neq 1$.

We stress that we understand the first part as a general conjecture on Galois representations attached to RACSDC representations (so, in particular, we assume that this is correct independently of the choice of F or π), while the second part puts a constraint on our choice of π. We also have the following:

Remark 8.2. The first part of Assumption 8.1 is known to hold e.g., if π is extremely regular [Barnet-Lamb et al. 2014]. Results in this direction are also contained in [Patrikis and Taylor 2015], but they are not directly applicable to our situation. We also remark that all entries in the ℓ-adic system $(\rho_{\pi,\lambda})_{\lambda\in\mathrm{Pl}_{\mathcal{E}(U)}}$ are expected (by cuspidality of π) to be absolutely irreducible and that this, using suitable modularity lifting theorems, is expected to imply absolute irreducibility of $\bar{\rho}_{\pi,\lambda}$ for almost all λ. An established result in this direction is that absolute irreducibility of the ℓ-adic system implies absolute irreducibility of $\bar{\rho}_{\pi,\lambda}$ except for a failure set of Dirichlet density 0, see [Patrikis et al. 2018].

Our main result is now as follows:

Theorem 8.3. *Presuming Assumption 8.1, there exists a subset $\Lambda^0_{\mathcal{E}_\Pi}\subset\Lambda^1_{\mathcal{E}_\Pi}$ of Dirichlet density 1 such that the functor $D^{\square_{S_\ell},\chi}_{S_\ell,W(k_\lambda)}(\bar{r}_\lambda)$ is globally unobstructed whenever $\lambda\in\Lambda^0_{\mathcal{E}_\Pi}$.*

As a first step towards the proof, let us consider the following simplifying assumption:

Assumption 8.4. (1) $F\mid F^+$ is unramified at all finite places and, in case n is even, then also $\frac{n}{2}[F^+:\mathbb{Q}]$ is even.

(2) Each place ν of F^+ which lies below S splits in $F\mid F^+$ as, say, $\tilde{\nu}\tilde{\nu}^c$. (For archimedean places, this condition is automatically fulfilled, so we can replace S by $S\sqcup\Omega_\infty$ without loss of generality.)

(3) For each place ν of F^+ which lies below S, the Weil–Deligne representation (r_ν, W_ν) attached to Π has unramified underlying Weil-representation r_ν.

Remark that the third part can be characterized as follows: The ℓ-adic representation $r_{\Pi,\lambda}$ is at ν a minimally ramified deformation of $\bar{r}_{\Pi,\lambda}$. (As the system associated to Π is compatible, this does not depend on the choice of $\lambda\in\Lambda^1_{\mathcal{E}_\Pi}$.) Now, consider the following (seemingly weaker) variation of Theorem 8.3:

Theorem 8.5. *Presuming Assumptions 8.1 and 8.4, there exists a subset $\Lambda^0_{\mathcal{E}_\Pi}\subset\Lambda^1_{\mathcal{E}_\Pi}$ of Dirichlet density 1 such that the functor $D^{\square_{S_\ell},\chi}_{S_\ell,W(k_\lambda)}(\bar{r}_\lambda)$ is globally unobstructed whenever $\lambda\in\Lambda^0_{\mathcal{E}_\Pi}$.*

Proof that Theorem 8.5 implies Theorem 8.3. It is a standard argument (see, e.g., the proof of [Clozel et al. 2008, Theorem 4.4.2]) that there exists a finite solvable extension $F_1^+\mid F^+$ of totally real fields such that

the extension $F_1 = F_1^+.F \mid F_1^+$ and the compatible family associated to the base change Π_{F_1} of Π to F_1 fulfill Assumption 8.4. Thus, referring to Lemma 4.8 and eliminating the finitely many places λ which divide the index $[F_1^+ : F^+]$, we see that Assumption 8.4 can be included in the statement of Theorem 8.3 without causing loss of generality. □

Consequently, the remainder of this section is devoted to the proof of Theorem 8.5. For better comprehension, let us give an overview of the strategy of the proof: We want to arrange for a situation where the framework of Section 4 is applicable, i.e., we want to consider suitable field extensions $L^+_{(\lambda)}$ for as many λ as possible such that Theorem 4.2 implies the vanishing of the dual Selmer groups of the base changed functors $D^{\square_{S_\ell},\chi}_{S_\ell,W(k_\lambda)}(\bar r_\lambda \mid G_{L^+_{(\lambda)}})$. This application of Theorem 4.2 happens in Theorem 8.12 below. By a careful choice of the extensions $L^+_{(\lambda)}$, we ensure that the potential unobstructedness arguments of Section 4 apply and yield the vanishing of the dual Selmer groups of the non-base changed functor. The local parts of the unobstructedness-condition then follow directly from the material in Section 5B, allowing us to conclude the statement of Theorem 8.5. The crucial property we have to impose on the extension $L^+_{(\lambda)}$ is *procurability* (Definition 8.7), i.e., that the deformation ring $R^{\chi\mid \mathrm{Gal}_{L^+},\mathtt{crys}}_{S_\ell,\mathcal{O}_{\mathcal{K}_\lambda}}(\bar r_\lambda \mid \mathrm{Gal}_{L^+})$ is isomorphic to $\mathcal{O}_{\mathcal{K}_\lambda}$ (corresponding to condition $(R = T)$ in Section 4). It is the content of Theorem 8.8 that for a set of places of Dirichlet density 1 we can find such suitable procurable extensions. This, in turn, is established by studying the seemingly weaker condition of $\star$-procurability (see the list ($\star_1$)–($\star_5$) below), which is proved to imply procurability almost everywhere (see Claim 1 below). By an argument based on Chebotarev's density theorem (and postponed to the Appendix), we can conclude that for a density-1 set we can find such $\star$-procurable extensions of 2-power degree.

8A. ***Proof of Theorem 8.5.*** Let us begin with some preparatory definitions.

Definition 8.6. A totally real, finite extension L^+ of F^+ is called *preadmissible* if the extension $L^+ \mid F^+$ is Galois and solvable and if $L := F.L^+$ is unramified over L^+ at every finite place.

We remark that these conditions are designed to capture the following: If L^+ is preadmissible, then there exists a unitary group H over L^+ (as considered in Section 6B) and a unitary avatar Π_L on $H(\mathbb{A}_{L^+})$ of the base change π_L of π to L.

For the following, let $\mathcal{E}$ be a number field containing $\mathcal{E}(U)$ and let L^+ be preadmissible.

Definition 8.7. A place $\lambda \in \Lambda^1_{\mathcal{E}}$ is L^+-*procurable* if the following two conditions are fulfilled:

(P.1) The restriction of $\bar\rho_\lambda$ to Gal_L remains absolutely irreducible.

(P.2) There exists an L-big enough extension $\mathcal{K}_\lambda$ of $\mathcal{E}_\lambda$ such that there is an isomorphism

$$R^{\chi,\mathtt{crys}}_{S_\ell,\mathcal{O}_{\mathcal{K}_\lambda}}(\bar r_\lambda \mid G_{L^+}) := R^{\chi\mid \mathrm{Gal}_{L^+},\mathtt{crys}}_{S_\ell,\mathcal{O}_{\mathcal{K}_\lambda}}(\bar r_\lambda \mid \mathrm{Gal}_{L^+}) \cong \mathcal{O}_{\mathcal{K}_\lambda}. \tag{13}$$

We remark that the first condition is rather harmless and affects only a failure set of Dirichlet density 0; see Assumption 8.1. We also remark that in the second condition, we consider $\bar r_\lambda$ as a representation with values in the residue field $k_{\mathcal{O}_{\mathcal{K}_\lambda}}$ of $\mathcal{O}_{\mathcal{K}_\lambda}$ instead of k_λ.

With respect to a preadmissible extension L^+, define $\mathrm{Proc}(L^+) \subset \Lambda^1_{\mathcal{E}}$ as the subset of those λ which are L^+-procurable.

Since there is a lack of $R=T$-theorems for mod-ℓ representations where the unitary group is associated to an extension F/F^+ in which places above ℓ do not split, we need to work around it by the following chain of extensions of F^+:

Theorem 8.8. *There exists a nested sequence $F^+ = L_0^+ \subset L_1^+ \subset \cdots$ of preadmissible extensions of F^+ such that*

$$\lim_{i\to\infty} \delta\Big(\bigcup_{j=1}^{i} \mathrm{Proc}(L_j^+)\Big) = 1, \tag{14}$$

where $\delta(\Delta)$ denotes the density of those rational primes q for which each $\lambda \in \mathrm{Pl}_{\mathcal{E}}$ above q fulfills $\lambda \in \Delta$.

Proof. Let us first introduce another notation: With respect to a preadmissible extension L^+, we say that $\lambda \in \Lambda^1_{\mathcal{E}}$ is L^+-$\star$-*procurable*, if the following list is met (where, as usual, we abbreviate $\ell = \ell(\lambda)$):

($\star_1$) ℓ is not divisible by any element of S.

($\star_2$) ℓ is unramified in the extension $L \mid \mathbb{Q}$.

($\star_3$) All places of L above S_ℓ are split in the extension $L \mid L^+$.

($\star_4$) The base change π_L of π to L remains cuspidal.

($\star_5$) If $\nu \in \mathrm{Pl}_A$ lies above S, then π_L admits a nontrivial Iwahori fix-vector.

Of particular difficulty will be proving that there are sufficiently many λ that fulfill condition ($\star_3$); this will be postponed to the Appendix.

As above, this defines a subset $\mathrm{Proc}^\star(L^+) \subset \Lambda^1_{\mathcal{E}}$. (Observe that condition ($\star_4$) does not depend on λ, but we intentionally include it in the list. So, if Π_L fails to be cuspidal, we have $\mathrm{Proc}^\star(L^+) = \varnothing$.)

Claim 1: $\mathrm{Proc}^\star(L^+) - \mathrm{Proc}(L^+)$ is finite.

Proof of Claim 1. We can suppose that $\mathrm{Proc}^\star(L^+)$ is not empty (otherwise the claim is trivially true), so in particular that π_L is a RACSDC representation and there exists a unitary group and an avatar Π_L over L. Now, for each $\lambda \in \mathrm{Proc}^\star(L^+)$ we pick an L-big enough field extension $\mathcal{K}_\lambda$ of $\mathcal{E}_\lambda$. We consider the complex Hecke algebra ${}^{\mathcal{O}_\mathcal{E}}\boldsymbol{T}^S_{\boldsymbol{\omega}}(U)$ and the ℓ-adic model $\mathbb{T} := {}^{\mathcal{O}_\mathcal{E}}\mathbb{T}^{S_\ell}_{\boldsymbol{\omega}}(U)$.

Write $\Pi_L = \langle f \rangle$ for the unitary avatar of the base change of π to L and for a suitable choice $f \in \mathcal{S}_{\boldsymbol{\omega}}(U)$. We see that $\bar\rho_\lambda \mid \mathrm{Gal}_L$ equals the reduction of the representation attached to the maximal ideal $\mathfrak{m} = \ker(\varphi_{f^{(\lambda)}}) \subset \mathbb{T}$ by Proposition 6.7, where $f^{(\lambda)}$ is the ℓ-adic model of f.

Recalling that we presume Assumption 8.1, we see that the conditions at the beginning of Section 7A hold for almost all of these choice of $L \mid L^+$, $\ell = \ell(\lambda)$, U, $\boldsymbol{\omega}$, $\mathcal{E}(U)$, $\mathcal{K}_\lambda$ and $\mathfrak{m}$. The main issue is the adequateness of $\bar\rho(\mathrm{Gal}_{F(\zeta_\ell)})$, which follows from [Barnet-Lamb et al. 2014, Proposition 2.1.2] as long as $\ell > 2(n+1)$. Thus, the desired isomorphism (13) follows for almost all λ in $\mathrm{Proc}^\star(L^+)$ by Corollary 7.7. This completes the proof of the claim. □

Consequently, it suffices to show that there exists a nested sequence $F^+ = L_0^+ \subset L_1^+ \subset \cdots$ of preadmissible extensions of F^+ such that (14) holds with Proc* instead of Proc. For the construction of these extensions, we define the set

$$\Theta_F := \{d \in \mathbb{N} \mid \sqrt{d} \notin F, \text{ the base change } \pi \rightsquigarrow \pi_{F(\sqrt{d})} \text{ remains cuspidal}\}.$$

By [Arthur and Clozel 1989, Theorem 4.2] there exists a finite extension E of F such that for any extension K' of F we have the following implication: If $E \cap K' = F$, then the base change of Π to K' remains cuspidal. This implication remains true after replacing E by its Galois closure, so we can assume that $E \mid F$ is Galois. Therefore this set is not empty, so choose a $d_1 \in \Theta_F$ and take $L_1^+ = F^+(\sqrt{d_1})$.

Claim 2: L_1^+ is preadmissible.

Proof of Claim 2. The extension $L_1^+ \mid F^+$ is automatically Galois and solvable because $[L_1^+ : F^+] = 2$. Thus we are left to check that $L_1 \mid L_1^+$ is unramified everywhere. This follows from e.g., [Marcus 1977, Chapter 4, Exercise 10]. □

Claim 3: $\delta(\text{Proc}^\star(F_1^+)) \geq \frac{1}{2}$.

Proof of Claim 3. We check which λ fail the list ($\star_1$)–($\star_5$):

- Concerning ($\star_1$) and ($\star_2$), we have to exclude the finitely many places λ for which $\ell(\lambda)$ is not coprime to S or ramifies in $L_1^+ \mid \mathbb{Q}$.
- By an estimation based on Chebotarev's density theorem (postponed as Lemma A to the appendix), the density of those ℓ which fulfill the condition that all primes of L_1^+ above ℓ are split in the extension $L_1 \mid L_1^+$ is at least $\frac{1}{2}$.
- Condition ($\star_4$) is universally fulfilled by our choice of L_1^+.
- Concerning condition ($\star_5$), we remark that by local-global compatibility (see [Chenevier and Harris 2013, Theorem 1.4] and the references therein) π_L admits an Iwahori-fixed vector if $\rho \mid \text{Gal}_L$ has unipotent ramification at v [Wedhorn 2008, (4.3.6) Proposition]. Thus, condition ($\star_5$) follows immediately from Assumption 8.4. □

For the next tower step we take $F_2^+ = F_1^+(\sqrt{d_2})$ for some $d_2 \in \Theta_{F_2}$. It is again easy to check that $\Theta_{F_2} \neq \varnothing$ and that F_2^+ is preadmissible. As in the proof of Claim 3, the statement of Lemma A implies $\delta(\text{Proc}^\star(F_2^+)) \geq \frac{3}{4}$. Iterating this construction of quadratic extensions we end up with a nested sequence of preadmissible fields F_j^+ such that

$$\delta\left(\bigcup_{j=1}^{i} \text{Proc}^\star(L_j^+)\right) \geq \delta(\text{Proc}^\star(L_i^+)) \geq 1 - \frac{1}{2^i} \xrightarrow{i\to\infty} 1.$$

Together with Claim 1, this concludes the proof of Theorem 8.8. □

We now give a slight variant of the above:

Definition 8.9. With regard to a preadmissible extension L^+ of F^+, we say that $\lambda \in \Lambda_{\mathcal{E}}^1$ is L^+-♯-*procurable* if the restriction of $\bar{\rho}_\lambda$ to Gal_L (with $L = F.L^+$) remains absolutely irreducible and if there is an isomorphism

$$R_\lambda^{\square,\chi,\mathtt{crys}}(L^+) \cong W(k_\lambda)[\![x_1,\dots,x_u]\!], \tag{15}$$

where $R_\lambda^{\square,\chi,\mathtt{crys}}(L^+) = R_{S_\ell,W(k_\lambda)}^{\square,\chi,\mathtt{crys}}(\bar{r}_\lambda \mid G_{L^+})$ and $u = \dim(\mathfrak{g}_n^{\mathtt{der}}) = n^2$. The set of all λ which are L^+-♯-procurable is denoted by $\mathrm{Proc}^\sharp(L^+)$.

Corollary 8.10. *There exists a nested sequence $F^+ = L_0^+ \subset L_1^+ \subset \cdots$ of preadmissible extensions of F^+ such that*

$$\lim_{i\to\infty} \delta\left(\bigcup_{j=1}^{i} \mathrm{Proc}^\sharp(L_j^+)\right) = 1.$$

Proof. For $i \in \mathbb{N}$, denote $\Delta_i = \bigcup_{j\le i} \mathrm{Proc}(L_j^+)$. Also fix for each $\lambda \in \Delta_i$ some $j \le i$ such that $\lambda \in \mathrm{Proc}(L_j^+)$. Denote the corresponding field extension from the proof of Theorem 8.8 by $L_{(\lambda)} = L_{(\lambda)}^+.F$. By Theorem 8.8, for such a $\lambda \in \Delta_i$ we have the identity (13) for a suitable extension $\mathcal{O}_{\mathcal{K}_\lambda}$ of $W(k_\lambda)$. The third part of Proposition 3.18 then yields

$$R_{S_\ell,\mathcal{O}_{\mathcal{K}_\lambda}}^{\square,\chi,\mathtt{crys}}(\bar{r}_\lambda \mid G_{L_{(\lambda)}^+}) \cong \mathcal{O}_{\mathcal{K}_\lambda}[\![x_1,\dots,x_u]\!].$$

Thus, we can use Lemma 3.7 (and, if necessary, Remark 3.11) to deduce the desired isomorphism (15). □

Corollary 8.11. *There exists a subset $\Lambda_{\mathcal{E}}^2 \subset \Lambda_{\mathcal{E}}^1$ of Dirichlet density 1 such that for each $\lambda \in \Lambda_{\mathcal{E}}^2$ there exists a finite, totally real extension $L_{(\lambda)}^+$ of F and an isomorphism*

$$R_{S_\ell,W(k_\lambda)}^{\square_{S_\ell},\chi,\mathtt{crys}}(\bar{r}_\lambda \mid G_{L_{(\lambda)}^+}) \cong W(k_\lambda)[\![x_1,\dots,x_{w(\lambda)}]\!]$$

with $w(\lambda) = (n^2+1).\#S_\ell - 1$.

Proof. This follows directly from Proposition 3.18. □

Next, we will apply the framework of Section 4 to the attained λ.

Theorem 8.12. *There exists a cofinite subset $\Lambda_{\mathcal{E}}^3 \subset \Lambda_{\mathcal{E}}^2$ such that the following holds: Let $\lambda \in \Lambda_{\mathcal{E}}^3$ and $L_{(\lambda)}^+$ the corresponding extension from Corollary 8.11. Then the functors*

$$D_{S_\ell,W(k_\lambda)}^{\square_{S_\ell},\chi,\mathtt{min}}(\bar{r}_\lambda \mid G_{L_{(\lambda)}^+}) \quad \textit{and} \quad D_{S_\ell,W(k_\lambda)}^{\square_{S_\ell},\chi}(\bar{r}_\lambda \mid G_{L_{(\lambda)}^+})$$

have vanishing dual Selmer group.

Proof. We start with the `min`-case. When applying the framework, we take for `sm` the condition parametrizing arbitrary deformations, for `crys` the condition parametrizing FL-crystalline deformations (see Section 5C) and for `min` the condition parametrizing minimally ramified deformations (see Section 5D). Moreover, we take $\chi = \epsilon_\ell^{1-n}\delta_{F\mid F^+}^{n \pmod 2}$. Let us now check the following list of conditions (and we abbreviate $L^+ = L_{(\lambda)}^+$ as we check this for a fixed $\lambda \in \Lambda_{\mathcal{E}}^2$):

(**sm/k**): As we took for $\mathtt{sm}$ the unrestricted deformation condition, we have to check that for each $v \in \Omega_\ell$ the functor $D^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+})$ is representable and that the representing object is formally smooth of relative dimension

$$d_v^{\square,\mathtt{sm}} = \dim(\mathfrak{g}_n^{\mathtt{der}})([L_{\tilde{v}} : \mathbb{Q}_\ell] + 1) = n^2([L_{\tilde{v}} : \mathbb{Q}_\ell] + 1) = n^2([L_v^+ : \mathbb{Q}_\ell] + 1).$$

(This also amounts to the vanishing of the error terms δ_v in Theorem 4.2.)

Check: Representability was already remarked in Section 3. For the remaining claim, we first refer to Proposition 6.1 in order to get an isomorphism

$$R^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+}) \cong D^{\square,\chi_v}_{W(k_\lambda)}(\bar{\rho}_{\lambda,v} \mid G_{L_{\tilde{v}}}).$$

Now the claim follows from Theorem 5.2.

(**crys**): For each $v \in \Omega_\ell$, the subfunctor

$$D^{\square,\chi_v,\mathtt{crys}}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+}) \hookrightarrow D^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+})$$

is relatively representable and the representing object is formally smooth of relative dimension $d_v^{\square,\mathtt{crys}} = \dim(\mathfrak{g}_n^{\mathtt{der}}) + (\dim(\mathfrak{g}_n^{\mathtt{der}}) - \dim(\mathfrak{b}_n^{\mathtt{der}}))[L_v^+ : \mathbb{Q}_\ell]$, where $\mathfrak{b}_n$ denotes the Lie algebra of a Borel subgroup of $\mathcal{G}_n$.

Check: By definition, we have

$$R^{\square,\chi_v,\mathtt{crys}}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+}) \cong D^{\square,\chi_v,\mathtt{crys}}_{W(k_\lambda)}(\bar{\rho}_{\lambda,v} \mid G_{L_{\tilde{v}}^+}).$$

Thus, the claim follows from Lemma 5.4.

(**min**): For each $v \in S$, the subfunctor

$$D^{\square,\chi_v,\mathtt{min}}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+}) \hookrightarrow D^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+})$$

is relatively representable and the representing object is formally smooth of relative dimension $d_v^{\square,\mathtt{min}} = \dim(\mathfrak{g}_n^{\mathtt{der}})$.

Check: Again, by definition, we have

$$R^{\square,\chi_v,\mathtt{min}}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+}) \cong D^{\square,\chi_v,\mathtt{min}}_{W(k_\lambda)}(\bar{\rho}_{\lambda,v} \mid G_{L_{\tilde{v}}^+}).$$

Thus, the claim follows from Proposition 5.6.

(∞): For each $v \in \Omega_\infty$, the local deformation ring $R^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+})$ is formally smooth of relative dimension $d_v^{\square} = \dim(\mathfrak{b}_n^{\mathtt{der}})$.

Check: This was already used, see Remark 7.5.

(**Presentability**): Consider the ring

$$R^{\mathtt{loc},\mathtt{min}}(L^+) := \widehat{\bigotimes}_{v \in S_\ell} \tilde{R}_v(L^+)$$

with $\tilde{R}_v(L^+) = D^{\square,\chi_v,\mathtt{min}}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+}))$ if $v \in S$ and $D^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_{\lambda,v} \mid G_{L_v^+}))$ otherwise. Then, there exists a presentation

$$R^{\square_{S_\ell},\chi}_{S_\ell,W(k_\ell)}(\bar{r}_\lambda \mid G_{L^+}) \cong R^{\mathtt{loc},\mathtt{min}}(L^+)[\![X_1, \dots, X_a]\!]/(f_1, \dots, f_b)$$

with $a - b = (\#S_\ell - 1).\dim(\mathfrak{g}_n^{\mathtt{ab}})$.

Check: This is the content of Proposition 3.20, but we have to check Assumption 3.19. As $\mathfrak{g}_n^{\mathtt{der}} = \mathfrak{gl}_n$, this condition holds by Corollary 3.25 for almost all λ.

($\boldsymbol{R} = \boldsymbol{T}$): The ring $R^{\square_{S_\ell},\chi,\mathtt{min},\mathtt{crys}}_{S_\ell,W(k_\ell)}(\bar{r}_\lambda \mid G_{L^+})$ is formally smooth of relative dimension $r_0 = \dim(\mathfrak{g}).\#S_\ell - \dim(\mathfrak{g}^{\mathtt{ab}})$.

Check: This follows from Corollary 8.11.

We see that the general requirements of Theorem 4.2 are met, so let us check the additional requirements of part 2 of Theorem 4.2:

- The condition $\ell \gg 0$ can be achieved by leaving out finitely many λ.
- The vanishing of $H^0(\mathrm{Gal}_{L^+}, \mathfrak{g}_n^{\mathtt{der},\vee})$ can be checked by observing
$$H^0(\mathrm{Gal}_{L^+}, \mathfrak{g}_n^{\mathtt{der},\vee}) \subset H^0(\mathrm{Gal}_L, \mathfrak{g}_n^{\mathtt{der},\vee}) \cong H^0(\mathrm{Gal}_L, \mathfrak{gl}_n^{\vee}),$$
as the adjoint representation of Gal_L on $\mathfrak{g}_n^{\mathtt{der}}$ (via $\bar{r}_\lambda$) corresponds to the adjoint representation of Gal_L on $\mathfrak{gl}_n$ (via $\bar{\rho}_\lambda$); see [Clozel et al. 2008, Section 2.1]. Thus, the desired vanishing follows for almost all λ by Corollary 3.25.
- For $\nu \in S$, $\dim(L_{\lambda,\nu}) = h^0(\mathrm{Gal}_{L_\nu^+}, \mathfrak{g}_n^{\mathtt{der}})$: As ν is split, Proposition 6.1 yields $h^0(\mathrm{Gal}_{L_\nu^+}, \mathfrak{g}_n^{\mathtt{der}}) = h^0(\mathrm{Gal}_{L_{\tilde{\nu}}}, \mathfrak{gl}_n^{\mathtt{der}})$, where the action on $\mathfrak{gl}_n$ is via $\bar{\rho}_{\lambda,\tilde{\nu}}$. The claim thus follows from [Clozel et al. 2008, Corollary 2.4.21].

The finitely many exclusions which occurred in the above items are now the places we must exclude from $\Lambda^2_{\mathcal{E}}$ to get $\Lambda^3_{\mathcal{E}}$. This finishes the first part, i.e., that $D^{\square_{S_\ell},\chi,\mathtt{min}}_{S_\ell,W(k_\lambda)}(\bar{r}_\lambda \mid G_{L^+_{(\lambda)}})$ has vanishing dual Selmer group. Concerning the second statement (i.e., the claimed vanishing of the non-minimal dual Selmer group) we first note that on each level $L^+_{(\lambda)}$ we can apply the $R = R^{\mathtt{min}}$-result of Corollary 7.11, yielding the desired vanishing except for a finite failure set. In other words, fix a place λ', then we have
$$D^{\square_{S_\ell},\chi,\mathtt{min}}_{S_\ell,W(k_\lambda)}(\bar{r}_\lambda \mid G_{L^+_{(\lambda)}}) = D^{\square_{S_\ell},\chi}_{S_\ell,W(k_\lambda)}(\bar{r}_\lambda \mid G_{L^+_{(\lambda)}})$$
for all λ with $L^+_{(\lambda)} = L^+_{(\lambda')}$, except for a finite failure set $\mathfrak{F}_{\lambda'}$. We should check that the occurrence of these failure sets at each step in the tower of field extensions does not disturb the desired result. For this, recall that the $L^+_{(\lambda)}$ show up in the tower $F^+ = L_0^+ \subset L_1^+ \subset \cdots$ and that, by the first statement, we have
$$\lim_{i\to\infty} \delta\{\lambda \mid D^{\square_{S_\ell},\chi,\mathtt{min}}_{S_\ell,W(k_\lambda)}(\bar{r}_\lambda \mid G_{L^+_{(\lambda)}}) \text{ has vanishing dual Selmer group, } L^+_{(\lambda)} \subset L_i^+\} = 1.$$
But this clearly implies
$$\lim_{i\to\infty} \delta\{\lambda \mid D^{\square_{S_\ell},\chi}_{S_\ell,W(k_\lambda)}(\bar{r}_\lambda \mid G_{L^+_{(\lambda)}}) \text{ has vanishing dual Selmer group, } L^+_{(\lambda)} \subset L_i^+, \lambda \notin \mathfrak{F}_\lambda\} = 1,$$
completing the proof. □

Proof of Theorem 8.5. The " has vanishing dual Selmer group" part of Theorem 8.5 follows immediately from Theorem 8.12 and the potential unobstructedness result of Lemma 4.8, as each $[L^+_{(\lambda)} : F^+]$ is a power of 2 (and k_λ has odd characteristic for $\lambda \in \Lambda^2_{\mathcal{E}}$). It remains to show that for all $\nu \in \Omega_\ell^{F^+}$ the local

lifting ring $R^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_{\lambda,v})$ is relatively smooth. By Theorem 5.2 we know that

$$\lim_{i\to\infty} \delta\{\lambda \mid R^{\square,\chi_v}_{W(k_\lambda)}(\bar{r}_\lambda \mid \mathrm{Gal}_{L^+_{(\lambda)},\nu'}) \text{ is unobstructed for all } \nu' \in \Omega^{L^+_{(\lambda)}}_{\ell(\lambda)},\ L^+_{(\lambda)} \subset L^+_i\}$$
$$= \lim_{i\to\infty} \delta\{\lambda \mid \text{any } \nu' \in \Omega^{L^+_{(\lambda)}}_{\ell(\lambda)} \text{ is split in the extension } L_{(\lambda)} \mid L^+_{(\lambda)},\ L^+_{(\lambda)} \subset L^+_i\} = 1.$$

Using Proposition 4.5 and Corollary 4.4, the claim follows. □

Appendix: A lemma on prime densities in non-Galois extensions

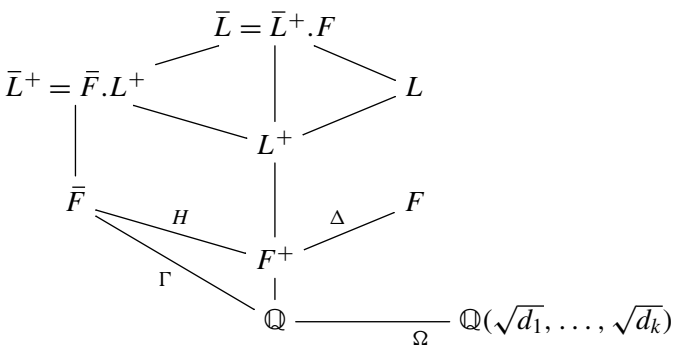
Let us consider a CM-field F with totally real subfield F^+ and we denote by $L^+ = F^+(\sqrt{d_1}, \ldots, \sqrt{d_k})$ the totally real extension of F^+ of degree 2^k, obtained by adjoining the square roots of some choice of elements $d_1, \ldots, d_k \in \mathbb{N}$ such that each d_i is a nonsquare in the Galois closure $\bar{F}^+$ of F^+. Set $L = L^+.F$. Then we have:

Lemma A. *Let $\Xi_{\mathbb{Q}}$ be the set of all those rational primes ℓ with the following property: For any place $\wp$ of L^+,*

$$[\wp \textit{ divides } \ell] \Rightarrow [\wp \textit{ splits in } L \mid L^+].$$

Then the density $\delta(\Xi_{\mathbb{Q}})$ of $\Xi_{\mathbb{Q}}$ in the set of all rational primes is at least $1 - 1/2^k$.

Proof. Consider the following diagram of fields:



with corresponding Galois groups $\Delta = \mathbb{Z}/2\mathbb{Z}$, $\Omega = (\mathbb{Z}/2\mathbb{Z})^k$ and Γ, H (for which we don't make an assumption). By our initial assumption that the d_i are not squares we have

$$\mathrm{Gal}(\bar{L}^+ \mid \mathbb{Q}) \cong \Gamma \times \Omega \quad \text{and, hence,} \quad \mathrm{Gal}(\bar{L} \mid \mathbb{Q}) \cong \Gamma \times \Omega \times \Delta.$$

Now, let $\mathfrak{P}$ be a place of $\bar{L}$ with corresponding Frobenius element $(\gamma, \omega, \delta) \in \mathrm{Gal}(\bar{L} \mid \mathbb{Q})$. As Ω and Δ are abelian, the conjugacy class of $\mathfrak{P}$ can be written as $\{(u\gamma u^{-1}, \omega, \delta) \mid u \in \Gamma\}$ and consists precisely of the Frobenii of the places of L lying over the same rational prime p as $\mathfrak{P}$. Let $\wp$ be the place of L below $\mathfrak{P}$. Its Frobenius element is given by

$$(\gamma, \omega, \delta)^{e_{\gamma,\omega,\delta}} \in H \times \{1\} \times \Delta = \mathrm{Gal}(\bar{L} \mid L^+)$$

for $e_{\gamma,\omega,\delta}$ minimal such that $(\gamma, \omega, \delta)^{e_{\gamma,\omega,\delta}} \in H \times \{1\} \times \Delta$. The condition that $\wp$ splits in $L \mid L^+$ then amounts precisely to $(\gamma, \omega, \delta)^{e_{\gamma,\omega,\delta}} \in H \times \{1\} \times \{1\}$, or, written in a more sophisticated way, that $q((\gamma, \omega, \delta)^{e_{\gamma,\omega,\delta}}) = 1$, where

$$q : \mathrm{Gal}(\bar{L} \mid L^+) \to \mathrm{Gal}(\bar{L} \mid L^+)/\mathrm{Gal}(\bar{L} \mid \bar{L}^+)$$

is the quotient map. If $\omega \neq 1$, we clearly must have $2 \mid e_{\gamma,\omega,\delta}$, which implies that $\wp$ splits in $L \mid L^+$. It is also important to note that the condition $\omega \neq 1$ is kept intact by conjugation inside $\mathrm{Gal}(\bar{L} \mid \mathbb{Q})$. Now, set

$$\Xi^* = \{(\gamma, \omega, \delta) \in \mathrm{Gal}(\bar{L} \mid \mathbb{Q}) \mid q((\gamma, \omega, \delta)^{e_{\gamma,\omega,\delta}}) = 1\}$$

and consider the subset $\Xi \subset \Xi^*$ which consists of those $g \in \Xi^*$ for which the complete conjugacy class is contained in Ξ^*, i.e., $\Xi = \{g \in \Xi^* \mid \langle g \rangle \subset \Xi^*\}$. We can give another characterization of this set, Ξ is the union of all conjugacy classes $\langle g \rangle \subset \mathrm{Gal}(\bar{L} \mid \mathbb{Q})$ with the following property: If $\boldsymbol{P}_g$ denotes the set of all places $\mathfrak{P}$ of $\bar{L}$ such that $\mathrm{Frob}_{\mathfrak{P}} \in \langle g \rangle$, then for any place $\wp$ of L^+ the following hold:

$$[\exists \mathfrak{P} \in \mathbf{P}_g \text{ such that } \mathfrak{P} \text{ divides } \wp] \Rightarrow [\wp \text{ splits in } L \mid L^+].$$

Then we have

$$\#\Xi \geq \#\{(\gamma, \omega, \delta) \in \mathrm{Gal}(\bar{L} \mid \mathbb{Q}) \mid \omega \neq 1\} = (2^k - 1).2.\#\Gamma.$$

As $\Xi_{\mathbb{Q}} = \{\ell \in \mathrm{Pl}_{\mathbb{Q}} \mid \exists g \in \Xi \text{ such that } \mathfrak{P} \mid \ell \text{ for all } \mathfrak{P} \in \mathbf{P}_g\}$, it follows from Chebotarev's density theorem that

$$\delta(\Xi_{\mathbb{Q}}) \geq \frac{(2^k - 1).2.\#\Gamma}{\mathrm{Gal}(\bar{L} \mid \mathbb{Q})} = 1 - \frac{1}{2^k}. \qquad \square$$

Acknowledgements

The main part of this article grew out of my PhD thesis [Guiraud 2016] at the University of Heidelberg. I would like to thank my doctoral advisor Gebhard Böckle for his support and help, throughout my PhD studies and thereafter. Moreover, I would like to thank Enno Nagel for his help during the preparation of the final version of this article. Finally, I would like to thank the Graduiertenkolleg Heidelberg (Stipendium nach dem Landesgraduiertenfördergesetz), MATCH (The Mathematics Center Heidelberg), the DAAD (Promos Stipendium) and the DFG (FG1920) for the funding received during my PhD studies and thereafter.

References

[Allen 2016] P. B. Allen, "Deformations of polarized automorphic Galois representations and adjoint Selmer groups", *Duke Math. J.* **165**:13 (2016), 2407–2460. MR Zbl

[Arthur and Clozel 1989] J. Arthur and L. Clozel, *Simple algebras, base change, and the advanced theory of the trace formula*, Ann. Math. Stud. **120**, Princeton Univ. Press, 1989. MR Zbl

[Atiyah and Macdonald 1969] M. F. Atiyah and I. G. Macdonald, *Introduction to commutative algebra*, Addison-Wesley, Reading, MA, 1969. MR Zbl

[Balaji 2012] S. Balaji, *G-valued potentially semi-stable deformation rings*, Ph.D. thesis, University of Chicago, 2012, Available at https://search.proquest.com/docview/1346024943.

[Barnet-Lamb et al. 2014] T. Barnet-Lamb, T. Gee, D. Geraghty, and R. Taylor, "Potential automorphy and change of weight", *Ann. of Math.* (2) **179**:2 (2014), 501–609. MR Zbl

[Bate et al. 2005] M. Bate, B. Martin, and G. Röhrle, "A geometric approach to complete reducibility", *Invent. Math.* **161**:1 (2005), 177–218. MR Zbl

[Bate et al. 2010] M. Bate, B. Martin, G. Röhrle, and R. Tange, "Complete reducibility and separability", *Trans. Amer. Math. Soc.* **362**:8 (2010), 4283–4311. MR Zbl

[Bellaïche 2009] J. Bellaïche, "An introduction to Bloch and Kato's conjecture", preprint, Clay Math. Inst. Summer School, 2009, Available at https://tinyurl.com/bellbloch.

[Bellaïche and Chenevier 2009] J. Bellaïche and G. Chenevier, *Families of Galois representations and Selmer groups*, Astérisque **324**, Soc. Math. France, Paris, 2009. MR Zbl

[Bleher and Chinburg 2003] F. M. Bleher and T. Chinburg, "Deformations with respect to an algebraic group", *Illinois J. Math.* **47**:3 (2003), 899–919. MR Zbl

[Böckle 2007] G. Böckle, "Presentations of universal deformation rings", pp. 24–58 in *L-functions and Galois representations* (Durham, UK, 2014), edited by D. Burns et al., Lond. Math. Soc. Lecture Note Ser. **320**, Cambridge Univ. Press, 2007. MR Zbl

[Böckle 2013] G. Böckle, "Deformations of Galois representations", pp. 21–115 in *Elliptic curves, Hilbert modular forms and Galois deformations*, edited by H. Darmon et al., Birkhäuser, Basel, 2013. MR Zbl

[Chenevier 2011] G. Chenevier, "On the infinite fern of Galois representations of unitary type", *Ann. Sci. École Norm. Sup.* (4) **44**:6 (2011), 963–1019. MR Zbl

[Chenevier and Harris 2013] G. Chenevier and M. Harris, "Construction of automorphic Galois representations, II", *Camb. J. Math.* **1**:1 (2013), 53–73. MR Zbl

[Clozel et al. 2008] L. Clozel, M. Harris, and R. Taylor, "Automorphy for some l-adic lifts of automorphic mod l Galois representations", *Publ. Math. Inst. Hautes Études Sci.* **108** (2008), 1–181. MR Zbl

[Conrad et al. 1999] B. Conrad, F. Diamond, and R. Taylor, "Modularity of certain potentially Barsotti–Tate Galois representations", *J. Amer. Math. Soc.* **12**:2 (1999), 521–567. MR Zbl

[Deligne 1973] P. Deligne, "Formes modulaires et représentations de GL(2)", pp. 55–105 in *Modular functions of one variable, II* (Antwerp, 1972), edited by P. Deligne and W. Kuyk, Lecture Notes in Math **349**, Springer, Berlin, 1973. MR Zbl

[Deligne and Serre 1974] P. Deligne and J.-P. Serre, "Formes modulaires de poids 1", *Ann. Sci. École Norm. Sup.* (4) **7** (1974), 507–530. MR Zbl

[EGA IV$_1$ 1964] A. Grothendieck, "Eléments de géométrie algébrique, IV: Étude locale des schémas et des morphismes de schémas, I", *Inst. Hautes Études Sci. Publ. Math.* **20** (1964), 5–259. MR Zbl

[Fontaine and Laffaille 1982] J.-M. Fontaine and G. Laffaille, "Construction de représentations p-adiques", *Ann. Sci. École Norm. Sup.* (4) **15**:4 (1982), 547–608. MR Zbl

[Freyd 1964] P. Freyd, *Abelian categories: an introduction to the theory of functors*, Harper & Row, New York, 1964. MR Zbl

[Fulton and Harris 1991] W. Fulton and J. Harris, *Representation theory: a first course*, Grad. Texts in Math. **129**, Springer, 1991. MR Zbl

[Gamzon 2016] A. Gamzon, "Unobstructed Hilbert modular deformation problems", *J. Théor. Nombres Bordeaux* **28**:1 (2016), 221–236. MR Zbl

[Gee 2011] T. Gee, "Automorphic lifts of prescribed types", *Math. Ann.* **350**:1 (2011), 107–144. MR Zbl

[Gee and Kisin 2014] T. Gee and M. Kisin, "The Breuil–Mézard conjecture for potentially Barsotti–Tate representations", *Forum Math. Pi* **2** (2014), art. id. e1. MR Zbl

[Geraghty 2010] D. J. Geraghty, *Modularity lifting theorems for ordinary Galois representations*, Ph.D. thesis, Harvard University, 2010, Available at https://search.proquest.com/docview/612773827.

[Gouvêa 2001] F. Q. Gouvêa, "Deformations of Galois representations", pp. 233–406 in *Arithmetic algebraic geometry* (Park City, UT, 1999), edited by B. Conrad and K. Rubin, IAS/Park City Math. Ser. **9**, Amer. Math. Soc., Providence, RI, 2001. MR Zbl

[Guerberoff 2011] L. Guerberoff, "Modularity lifting theorems for Galois representations of unitary type", *Compos. Math.* **147**:4 (2011), 1022–1058. MR Zbl

[Guiraud 2016] D.-A. Guiraud, *A framework for unobstructedness of Galois deformation rings*, Ph.D. thesis, Universität Heidelberg, 2016, Available at http://www.ub.uni-heidelberg.de/archiv/20248.

[Harris et al. 2008] J. M. Harris, J. L. Hirst, and M. J. Mossinghoff, *Combinatorics and graph theory*, 2nd ed., Springer, 2008. MR Zbl

[Hatley 2015] J. Hatley, *Obstruction criteria for modular deformation problems*, Ph.D. thesis, University of Massachusetts Amherst, 2015, Available at https://tinyurl.com/hatleyphd.

[James and Kerber 1981] G. James and A. Kerber, *The representation theory of the symmetric group*, Encycl. Math. Appl. **16**, Addison-Wesley, Reading, MA, 1981. MR Zbl

[Jannsen 1989] U. Jannsen, "On the l-adic cohomology of varieties over number fields and its Galois cohomology", pp. 315–360 in *Galois groups over* $\mathbb{Q}$ (Berkeley, 1987), edited by Y. Ihara et al., Math. Sci. Res. Inst. Publ. **16**, Springer, 1989. MR Zbl

[Khare 2003] C. Khare, "On isomorphisms between deformation rings and Hecke rings", *Invent. Math.* **154**:1 (2003), 199–222. MR Zbl

[Khare and Wintenberger 2009a] C. Khare and J.-P. Wintenberger, "On Serre's conjecture for 2-dimensional mod p representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$", *Ann. of Math.* (2) **169**:1 (2009), 229–253. MR Zbl

[Khare and Wintenberger 2009b] C. Khare and J.-P. Wintenberger, "Serre's modularity conjecture, II", *Invent. Math.* **178**:3 (2009), 505–586. MR Zbl

[Kisin 2007] M. Kisin, "Modularity of 2-dimensional Galois representations", pp. 191–230 in *Current developments in mathematics, 2005*, edited by D. Jerison et al., Int. Press, Somerville, MA, 2007. MR Zbl

[Levin 2013] B. W. A. Levin, *G-valued flat deformations and local models*, Ph.D. thesis, Stanford University, 2013.

[Marcus 1977] D. A. Marcus, *Number fields*, Springer, 1977. MR Zbl

[Matsumura 1970] H. Matsumura, *Commutative algebra*, Benjamin, New York, 1970. MR Zbl

[Mauger 2000] D. Mauger, *Algèbre de Hecke quasi-ordinaire universelle d'un groupe réductif*, Ph.D. thesis, Université Sorbonne Paris Nord, 2000, Available at https://tel.archives-ouvertes.fr/tel-00005938.

[Mazur 1989] B. Mazur, "Deforming Galois representations", pp. 385–437 in *Galois groups over* $\mathbb{Q}$ (Berkeley, 1987), edited by Y. Ihara et al., Math. Sci. Res. Inst. Publ. **16**, Springer, 1989. MR Zbl

[Mazur 1997] B. Mazur, "An introduction to the deformation theory of Galois representations", pp. 243–311 in *Modular forms and Fermat's last theorem* (Boston, 1995), edited by G. Cornell et al., Springer, 1997. MR Zbl

[Neukirch et al. 2008] J. Neukirch, A. Schmidt, and K. Wingberg, *Cohomology of number fields*, 2nd ed., Grundlehren der Math. Wissenschaften **323**, Springer, 2008. MR Zbl

[Patrikis and Taylor 2015] S. Patrikis and R. Taylor, "Automorphy and irreducibility of some l-adic representations", *Compos. Math.* **151**:2 (2015), 207–229. MR Zbl

[Patrikis et al. 2018] S. T. Patrikis, A. W. Snowden, and A. J. Wiles, "Residual irreducibility of compatible systems", *Int. Math. Res. Not.* **2018**:2 (2018), 571–587. MR Zbl

[Prasad 2014] D. Prasad, "Notes on representations of finite groups of Lie type", preprint, 2014. arXiv

[Ramakrishna 1993] R. Ramakrishna, "On a variation of Mazur's deformation functor", *Compositio Math.* **87**:3 (1993), 269–286. MR Zbl

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

[Sernesi 2006] E. Sernesi, *Deformations of algebraic schemes*, Grundlehren der Math. Wissenschaften **334**, Springer, 2006. MR Zbl

[Serre 1979] J.-P. Serre, *Local fields*, Grad. Texts in Math. **67**, Springer, 1979. MR Zbl

[Serre 1998] J.-P. Serre, "The notion of complete reducibility in group theory", part II in "Moursund lectures", lecture notes, Univ. Oregon Math. Dept., 1998. arXiv

[Serre 2000] J.-P. Serre, *Local algebra*, Springer, 2000. MR Zbl

[Shimura 1971] G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Publ. Math. Soc. Japan **11**, Iwanami Shoten, Tokyo, 1971. MR Zbl

[Shotton 2015] J. Shotton, *The Breuil–Mézard conjecture when l is not equal to p*, Ph.D. thesis, Imperial College London, 2015, Available at https://doi.org/10.25560/25747.

[Shotton 2018] J. Shotton, "The Breuil–Mézard conjecture when $l \neq p$", *Duke Math. J.* **167**:4 (2018), 603–678. MR Zbl

[Sprang 2013] J. Sprang, "A universal deformation ring with unexpected Krull dimension", *Math. Z.* **275**:1-2 (2013), 647–652. MR Zbl

[Taylor 2008] R. Taylor, "Automorphy for some l-adic lifts of automorphic mod l Galois representations, II", *Publ. Math. Inst. Hautes Études Sci.* **108** (2008), 183–239. MR Zbl

[Thorne 2012] J. Thorne, "On the automorphy of l-adic Galois representations with small residual image", *J. Inst. Math. Jussieu* **11**:4 (2012), 855–920. MR Zbl

[Tilouine 1996] J. Tilouine, *Deformations of Galois representations and Hecke algebras*, Narosa, New Delhi, 1996. MR Zbl

[Wedhorn 2008] T. Wedhorn, "The local Langlands correspondence for GL(n) over p-adic fields", pp. 237–320 in *School on Automorphic Forms on* GL(n) (Trieste, Italy, 2000), edited by L. Göttsche et al., ICTP Lect. Notes **21**, Abdus Salam Int. Cent. Theoret. Phys., Trieste, Italy, 2008. MR Zbl

[Weston 2004] T. Weston, "Unobstructed modular deformation problems", *Amer. J. Math.* **126**:6 (2004), 1237–1252. MR Zbl

[Yamagami 2004] A. Yamagami, "On the unobstructedness of the deformation problems of residual modular representations", *Tokyo J. Math.* **27**:2 (2004), 443–455. MR Zbl

Communicated by Brian Conrad
Received 2017-06-30 Revised 2019-08-13 Accepted 2020-02-10

guiraud@mailbox.org *Baden-Baden, Germany*

msp

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1381

The Hilbert scheme of hyperelliptic Jacobians and moduli of Picard sheaves

Andrea T. Ricolfi

Let C be a hyperelliptic curve embedded in its Jacobian J via an Abel–Jacobi map. We compute the scheme structure of the Hilbert scheme component of Hilb_J containing the Abel–Jacobi embedding as a point. We relate the result to the ramification (and to the fibres) of the Torelli morphism $\mathcal{M}_g \to \mathcal{A}_g$ along the hyperelliptic locus. As an application, we determine the scheme structure of the moduli space of Picard sheaves (introduced by Mukai) on a hyperelliptic Jacobian.

Introduction

Main result. In this paper we study the deformation theory of a smooth *hyperelliptic* curve C of genus $g \geq 2$, embedded in its Jacobian $J = (\mathrm{Pic}^0 C, \Theta_C)$ via an Abel–Jacobi map

$$\mathsf{aj}\colon C \hookrightarrow J.$$

We work over an algebraically closed field k of characteristic different from 2. Our aim is to compute the scheme structure of the Hilbert scheme component

$$\mathrm{Hilb}_{C/J} \subset \mathrm{Hilb}_J$$

containing the point defined by aj. It is well-known that the embedded deformations of C into J are parametrised by translations of C, and that they are *obstructed* as long as $g \geq 3$ (see the next section for more details). In other words $\mathrm{Hilb}_{C/J}$ is *singular*, with reduced underlying variety isomorphic to J. The tangent space dimension to the Hilbert scheme has been computed in [Lange and Sernesi 2004; Griffiths 1967]. The result is

$$\dim_k H^0(C, N_C) = 2g - 2.$$

MSC2010: primary 14C05; secondary 14H40, 14K10.
Keywords: Jacobian, Torelli morphism, Hilbert schemes, Picard sheaves, Fourier–Mukai transform.

Therefore, as $\dim J = g$, the nonreduced structure of $\mathrm{Hilb}_{C/J}$ along J is accounted for (up to first order) by $g-2$ extra tangents. By homogeneity of the Jacobian, it is natural to expect a decomposition

$$\mathrm{Hilb}_{C/J} = J \times R_g$$

for some artinian scheme R_g with embedding dimension $g-2$. As we shall see, this is precisely what happens, and R_g turns out to be the "smallest" (in the sense of Lemma 3.3) artinian scheme with the required embedding dimension. More precisely, let

$$R_g = \operatorname{Spec} k[s_1, \ldots, s_{g-2}]/\mathfrak{m}^2, \tag{0-1}$$

where $\mathfrak{m} = (s_1, \ldots, s_{g-2})$ is the maximal ideal of the origin. The main result of this paper (proved in Theorem 3.6 in the main body) is the following.

Theorem 1. *Let C be a hyperelliptic curve of genus $g \geq 2$ over a field k of characteristic different from 2, and let J be its Jacobian. Then there is an isomorphism of k-schemes*

$$\mathrm{Hilb}_{C/J} \cong J \times R_g,$$

where R_g is the artinian scheme (0-1).

Interpretation. Let $\mathcal{M}_g$ be the moduli stack of smooth curves of genus g, and let $\mathcal{A}_g$ be the moduli stack of principally polarised abelian varieties of dimension g. The Torelli morphism

$$\tau_g \colon \mathcal{M}_g \to \mathcal{A}_g$$

sends a curve C to its Jacobian $J = \mathrm{Pic}^0 C$, principally polarised by the Theta divisor Θ_C. One can interpret the artinian scheme R_g as the fibre of τ_g over a hyperelliptic point $[J, \Theta_C] \in \mathcal{A}_g$. This makes explicit the link between the *ramification* of τ_g along the hyperelliptic locus (in other words, the failure of the infinitesimal Torelli property) and the singularities of the Hilbert scheme $\mathrm{Hilb}_{C/J}$ (in other words, the *obstructions* to deform C in J). We come back to this in Section 3B.

Moduli of Picard sheaves. As an application of our result, in Section 4 we compute the scheme structure of certain moduli spaces of *Picard sheaves* on a hyperelliptic Jacobian J. Mukai introduced these spaces as an application of his Fourier transform; he completed their study in the nonhyperelliptic case [Mukai 1981; 1987], leaving open the hyperelliptic one.

Let F be the Fourier–Mukai transform of a line bundle $\xi = \mathscr{O}_C(dp_0)$, where $p_0 \in C$ and we assume $1 \leq d \leq g-1$ to ensure that F is a simple sheaf on J. Let $M(F)$ be the connected component of the moduli space of simple sheaves containing the point $[F]$. Mukai proved that $M(F)_{\mathrm{red}} = \hat{J} \times J$, the isomorphism being given by the family of twists and translations of F [Mukai 1987, Example 1.15]. Under the same assumptions of Theorem 1, we prove the following (see Theorem 4.2 in the main body of the text).

Theorem 2. *There is an isomorphism of k-schemes*

$$M(F) \cong \hat{J} \times J \times R_g.$$

Enumerative geometry of abelian varieties. A motivation for understanding the *scheme structure* of classical moduli spaces such as the Hilbert scheme (Theorem 1) and the moduli space of Picard sheaves (Theorem 2) comes from the subject of enumerative geometry of abelian varieties.

For instance, the Hilbert scheme of curves (in a 3-fold) is the main player in Donaldson–Thomas theory — see, for instance, [Bryan et al. 2018] for an exhaustive treatment (including several interesting conjectures) of the Enumerative Geometry of curves on abelian surfaces and 3-folds. Understanding the scheme structure (or even the closed points!) of the Hilbert scheme of curves on a 3-fold is very often a hopeless problem. Of course, Donaldson–Thomas theory has developed several sophisticated tools to deal with the lack of an explicit description of the Hilbert scheme; however, this paper shows that, at least for an arbitrary *Abel–Jacobi curve*, the Hilbert scheme can be described completely. Thus an immediate corollary of Theorem 1 is the explicit description of the Donaldson–Thomas theory of an Abel–Jacobi curve; see Section 3C.

On the other hand, it is conceivable that the theory of *Picard sheaves*, arising as a direct application of the Fourier–Mukai transform, could be exploited to aim for a deeper understanding of the intersection theory and cohomology of Jacobians, and possibly their compactifications. Having at one's disposal global results such as Theorem 2 might allow one to treat the *whole* moduli space (the universal Jacobian over the moduli space of curves) at once in developing a theory of *tautological rings* for (possibly compactified, universal) Jacobians, by combining Fourier–Mukai techniques with suitable analogues of the intersection theoretic calculations carried out in [Pagani et al. 2018].

Conventions. We work over an algebraically closed field k of characteristic $p \neq 2$. All curves are smooth and proper over k, they are (geometrically) connected, and their Jacobians are principally polarised by the Theta divisor.

1. Ramification of Torelli and the Hilbert scheme

In this section we provide the framework for where the problem tackled in this paper naturally lives in.

1A. ***Deformations of Abel–Jacobi curves.*** The following theorem was proved in the stated form by Lange and Sernesi, but see also the work of Griffiths [1967].

Theorem 1.1 [Lange and Sernesi 2004, Theorem 1.2]. *Let C be a smooth curve of genus $g \geq 3$:*

(i) *If C is nonhyperelliptic, then $\mathrm{Hilb}_{C/J}$ is smooth of dimension g.*

(ii) *If C is hyperelliptic, then $\mathrm{Hilb}_{C/J}$ is irreducible of dimension g and everywhere nonreduced, with Zariski tangent space of dimension $2g-2$.*

In both cases, the only deformations of C in J are translations.

The statement of Theorem 1.1 is proved over $\mathbb{C}$ in [Lange and Sernesi 2004], but it holds over algebraically closed fields k of arbitrary characteristic. To see this, we need Collino's extension of the Ran–Matsusaka criterion for Jacobians to an arbitrary field, which we state here for completeness.

Theorem 1.2 [Collino 1984]. *Let X be an abelian variety of dimension g over an algebraically closed field k. Let D be an effective 1-cycle generating X and let $\Theta \subset X$ be an ample divisor such that $D \cdot \Theta = g$. Then (X, Θ, D) is a Jacobian triple.*

Proof of Theorem 1.1. Let $C \to \operatorname{Spec} k$ be a smooth curve of genus g and fix an Abel–Jacobi map $C \hookrightarrow J$. Consider the normal bundle exact sequence

$$0 \to T_C \to T_J|_C \to N_C \to 0.$$

Since we have a canonical identification $T_J|_C = H^1(C, \mathscr{O}_C) \otimes_k \mathscr{O}_C$, the induced cohomology sequence is

$$0 \to H^1(C, \mathscr{O}_C) \to H^0(C, N_C) \xrightarrow{\partial} H^1(C, T_C) \xrightarrow{\sigma} H^1(C, \mathscr{O}_C)^{\otimes 2}. \tag{1-1}$$

Since $H^0(C, N_C)$ is the tangent space to the Hilbert scheme, and $\dim_k H^1(C, \mathscr{O}_C) = g$, it is clear that $\mathrm{Hilb}_{C/J}$ is smooth of dimension g if and only if $\partial = 0$, if and only if σ is injective. The map σ factors through the subspace $\operatorname{Sym}^2 H^1(C, \mathscr{O}_C)$, and its dual is the multiplication map

$$\mu_C \colon \operatorname{Sym}^2 H^0(C, K_C) \to H^0(C, K_C^2),$$

where K_C is the canonical line bundle of C. For a modern, fully detailed proof of the identification $\sigma^\vee = \mu_C$, we refer the reader to [Landesman 2019, Theorem 4.3]. By a theorem of Max Noether [Arbarello et al. 1985, Chapter III, Section 2], the map μ_C is surjective if and only if C is nonhyperelliptic (see also [Griffiths 1967; Andreotti 1958] for different proofs). If C is hyperelliptic, the quotient $H^0(C, N_C)/H^1(C, \mathscr{O}_C) = \operatorname{Im}\partial$ has dimension $g-2$, as shown directly in [Oort and Steenbrink 1980, Section 2] by choosing appropriate bases of differentials. This proves part (i) of Theorem 1.1, along with the count $h^0(C, N_C) = 2g-2$ (and the nonreducedness statement) of part (ii). So in the nonhyperelliptic case, $\mathrm{Hilb}_{C/J}$ is smooth of dimension g.

To finish the proof of part (ii), suppose C is hyperelliptic, and let $D \subset J$ be a closed 1-dimensional k-subscheme defining a point of $\mathrm{Hilb}_{C/J}$. Then D is represented by the *minimal cohomology class*

$$\frac{\Theta_C^{g-1}}{(g-1)!}$$

on J. This implies at once that D generates J, and that $D \cdot \Theta_C = g$. Therefore, by Theorem 1.2, $(\operatorname{Pic}^0 D, \Theta_D)$ and (J, Θ_C) are isomorphic as principally polarised abelian varieties. By Torelli's theorem, this implies (using also that C is hyperelliptic) that D is a translate of C. Thus $\mathrm{Hilb}_{C/J}$ is irreducible of dimension g, and its k-points coincide with those of J. The result follows. □

Corollary 1.3. *Let J be the Jacobian of a nonhyperelliptic curve C. Then the family of translations of C inside J induces an isomorphism*

$$J \xrightarrow{\sim} \mathrm{Hilb}_{C/J}.$$

Proof. The natural morphism $h\colon J \to \mathrm{Hilb}_{C/J}$ is proper (since J is proper and the Hilbert scheme is proper, hence separated), injective on points and tangent spaces — since the tangent map at $0 \in J$ is the map $\mathrm{d}h\colon H^1(C, \mathscr{O}_C) \hookrightarrow H^0(C, N_C)$ in the sequence (1-1). Thus h is a closed immersion, in particular it is unramified. However, the proof of Theorem 1.1 shows that $h\colon J \to \mathrm{Hilb}_{C/J}$ is bijective and, since C is nonhyperelliptic, $\mathrm{d}h$ is an isomorphism. Thus h is an isomorphism. □

Remark 1.4. If C is a generic complex curve of genus at least 3, its 1-cycle on J is not algebraically equivalent to the cycle of $-C$ by a famous theorem of Ceresa [1983]. Here $-C$ is the image of C under the automorphism $-1\colon J \to J$. Therefore the Hilbert scheme Hilb_J contains another component $\mathrm{Hilb}_{-C/J}$, disjoint from $\mathrm{Hilb}_{C/J}$ and still isomorphic to J.

1B. ***Torelli problems.*** Consider the Torelli morphism

$$\tau_g\colon \mathcal{M}_g \to \mathcal{A}_g$$

from the stack of nonsingular curves of genus g to the stack of principally polarised abelian varieties, sending a curve to its (canonically polarised) Jacobian. The *infinitesimal Torelli problem* asks whether the Torelli morphism is an immersion. It is well-known that τ_g is ramified along the hyperelliptic locus; this is again Noether's theorem, stating that μ_C, the *codifferential* of τ_g at $[C] \in \mathcal{M}_g$, is not surjective. So, even though τ_g is injective on geometric points by Torelli's theorem, it is not an immersion.

To sum up, we have the following. Let C be an arbitrary smooth curve of genus $g \geq 3$, and let J be its Jacobian. Then the following conditions are equivalent:

(i) C is hyperelliptic.

(ii) $\mathrm{Hilb}_{C/J}$ is singular at $[\mathrm{aj}\colon C \hookrightarrow J]$.

(iii) The embedded deformations of C into J are obstructed.

(iv) $\tau_g\colon \mathcal{M}_g \to \mathcal{A}_g$ is ramified at $[C]$.

(v) Infinitesimal Torelli fails at C.

The *local Torelli problem* for curves, studied by Oort and Steenbrink [1980], asks whether the morphism

$$t_g\colon M_g \to A_g$$

between the coarse moduli spaces is an immersion. These schemes do not represent the corresponding moduli functors, so the local structure of t_g is not (directly) linked with deformation theory of curves and their Jacobians. However, introducing suitable level structures, one replaces the normal varieties M_g and A_g with smooth varieties

$$M_g^{(n)} \quad \text{and} \quad A_g^{(n)}$$

that are *fine* moduli spaces for the corresponding moduli problem, and are étale over $\mathcal{M}_g$ and $\mathcal{A}_g$, respectively.

Let $p \geq 0$ be the characteristic of the base field. Oort and Steenbrink show that t_g is an immersion if $p = 0$. The answer to the local Torelli problem is also affirmative if $p > 2$, at almost all points of M_g. More precisely, t_g is an immersion at those points in M_g representing curves C such that $\operatorname{Aut} C$ has no elements of order p [Oort and Steenbrink 1980, Corollary 3.2]. Finally, t_g is *not* an immersion if $p = 2$ and $g \geq 5$ [loc. cit., Corollary 5.3].

2. Moduli spaces with level structures

In this section we introduce the moduli spaces of curves and abelian varieties we will be working with throughout.

2A. ***Level structures.*** Let S be a scheme. An abelian scheme over S is a group scheme $X \to S$ which is smooth and proper and has geometrically connected fibres. We let $\hat{X} \to S$ denote the dual abelian scheme. A polarisation on $X \to S$ is an S-morphism $\lambda\colon X \to \hat{X}$ such that its restriction to every geometric point $s \in S$ is of the form

$$\phi_{\mathscr{L}}\colon X_s \to \hat{X}_s, \quad x \mapsto \mathrm{t}_x^*\mathscr{L} \otimes \mathscr{L}^\vee,$$

for some ample line bundle $\mathscr{L}$ on X_s. Here and in what follows, t_x is the translation $y \mapsto x + y$ by the element $x \in X_s$. We say λ is *principal* if it is an isomorphism.

Fix an integer $n > 0$ and an abelian scheme $X \to S$ of relative dimension g. Multiplication by n is an S-morphism of group schemes

$$[n]\colon X \to X,$$

and we denote its kernel by X_n. Assuming n is not divisible by p, we have that X_n is an étale group scheme over S, locally isomorphic in the étale topology to the constant group scheme $(\mathbb{Z}/n\mathbb{Z})^{2g}$. One has $\hat{X}_n = X_n^D$, where the superscript D denotes the Cartier dual of a finite group scheme. Then any principal polarisation λ on X induces a skew-symmetric bilinear form

$$E_n\colon X_n \times_S X_n \xrightarrow{\mathrm{id}\times\lambda} X_n \times_S X_n^D \xrightarrow{e_n} \mu_n,$$

where e_n is the Weil pairing. The group $\mathbb{Z}/n\mathbb{Z}$ is Cartier dual to μ_n. We endow $(\mathbb{Z}/n\mathbb{Z})^g \xrightarrow{\sim} \mu_n^g$ with the standard symplectic structure, given by the $2g \times 2g$ matrix

$$\begin{pmatrix} 0 & \mathbb{1}_g \\ -\mathbb{1}_g & 0 \end{pmatrix}.$$

Definition 2.1 [Oort and Steenbrink 1980]. A (symplectic) level-n structure on a principally polarised abelian scheme $(X/S, \lambda)$ is a symplectic isomorphism

$$\alpha\colon (X_n, E_n) \xrightarrow{\sim} (\mathbb{Z}/n\mathbb{Z})^{2g}.$$

A level-n structure on a smooth proper curve $\mathcal{C} \to S$ is a level structure on its Jacobian $\mathrm{Pic}^0(\mathcal{C}/S) \to S$.

Curves with level structure are represented by pairs (C, α). We consider (C, α) and (C', α') as being isomorphic if there is an isomorphism $u\colon C \xrightarrow{\sim} C'$ such that the induced isomorphism $J(u)\colon J' \xrightarrow{\sim} J$ between the Jacobians takes α' to α. An isomorphism between (X, λ, α) and (X', λ', α') is an isomorphism $(X', \lambda') \xrightarrow{\sim} (X, \lambda)$ of principally polarised abelian schemes, taking α' to α.

Remark 2.2. If C is a curve of genus $g \geq 3$ with trivial automorphism group, and α is a level structure on C, then (C, α) is not isomorphic to $(C, -\alpha)$. On the other hand, if J denotes the Jacobian of C, one has that (J, Θ_C, α) and $(J, \Theta_C, -\alpha)$ are isomorphic, because the automorphism $-1\colon J \to J$, defined globally on J, identifies the two pairs.

2A1. *Choice of level.* As indicated by Theorem 2.3 below, moduli spaces of curves and abelian varieties with level structure are well behaved when the condition $(p, n) = 1$ is met. For later purposes, we need to strengthen the condition $(p, n) = 1$. Note that $p = \operatorname{char} k$ is fixed, as well as the genus g. However, we are free to choose $n \geq 3$, and the condition we require is that the order of the symplectic group

$$|\mathrm{Sp}(2g, \mathbb{Z}/n\mathbb{Z})| = n^{g^2} \cdot \prod_{i=1}^{g} (n^{2i} - 1)$$

is not divisible by p. In particular, this implies $(p, n) = 1$. From now on,

$$n \text{ is fixed in such a way that } p \text{ does not divide } |\mathrm{Sp}(2g, \mathbb{Z}/n\mathbb{Z})|. \tag{2-1}$$

This condition implies that the symplectic group $\mathrm{Sp}(2g, \mathbb{Z}/n\mathbb{Z})$ acts freely and transitively on the set of symplectic level-n structures on a smooth curve defined over k. This will be used in the proof of Lemma 2.5.

2B. ***Moduli spaces.*** Let $\mathscr{M}_g^{(n)}$ be the functor $\mathrm{Sch}_k^{\mathrm{op}} \to$ Sets sending a k-scheme S to the set of S-isomorphism classes of curves of genus g with level-n structure. Similarly, let $\mathscr{A}_g^{(n)}$ be the functor sending S to the set of S-isomorphism classes of principally polarised abelian schemes of relative dimension g over S equipped with a level-n structure.

Theorem 2.3. *If $n \geq 3$ and $(p, n) = 1$, the functors $\mathscr{M}_g^{(n)}$ and $\mathscr{A}_g^{(n)}$ are represented by smooth quasiprojective varieties $M_g^{(n)}$ and $A_g^{(n)}$ of dimensions $3g - 3$ and $g(g+1)/2$ respectively.*

Proof. For the statement about $\mathscr{M}_g^{(n)}$ we refer to [Popp 1977], whereas the one about $\mathscr{A}_g^{(n)}$ is [Mumford 1965, Theorem 7.9]. □

Consider the morphism

$$j_n\colon M_g^{(n)} \to A_g^{(n)} \tag{2-2}$$

sending a curve with level structure to its Jacobian, as usual principally polarised by the Theta divisor. The map j_n is generically of degree two onto its image, essentially because of Remark 2.2. To link it back to $t_g\colon M_g \to A_g$, Oort and Steenbrink form the geometric quotient

$$V^{(n)} = M_g^{(n)}/\Sigma,$$

where

$$\Sigma\colon M_g^{(n)} \to M_g^{(n)} \tag{2-3}$$

is the involution sending $[D, \beta] \mapsto [D, -\beta]$. Note that Σ is the identity if $g \leq 2$. The map j_n factors through a morphism

$$\iota\colon V^{(n)} \to A_g^{(n)},$$

which turns out to be injective on geometric points [Oort and Steenbrink 1980, Lemma 1.11]. In fact, we need the following stronger statement:

Theorem 2.4 [Oort and Steenbrink 1980, Theorem 3.1]. *If $g \geq 2$ and $\operatorname{char} k \neq 2$ then ι is an immersion.*

Oort and Steenbrink use this result crucially to solve the local Torelli problem as we recalled in Section 1B. For us, it is not important to have the statement of local Torelli (which strictly speaking only holds globally in characteristic 0); all we need in our argument is Theorem 2.4, which is why we require the base field k to have characteristic $p \neq 2$.

The following result was proven in [Deligne and Mumford 1969, Proposition 5.8] in greater generality. We give a short proof here for the sake of completeness.

Lemma 2.5. *The maps $\varphi\colon M_g^{(n)} \to \mathcal{M}_g$ and $\psi\colon A_g^{(n)} \to \mathcal{A}_g$ forgetting the level structure are étale.*

Proof. We start by showing that φ is flat. Choose an atlas for $\mathcal{M}_g$, that is, an étale surjective map $a\colon U \to \mathcal{M}_g$ from a scheme. Form the fibre square

$$\begin{array}{ccc} V & \xrightarrow{b} & M_g^{(n)} \\ \downarrow & \square & \downarrow{\scriptstyle\varphi} \\ U & \xrightarrow[a]{} & \mathcal{M}_g \end{array}$$

and pick a point $u \in U$, with image $y = a(u) \in \mathcal{M}_g$. The fibre $V_u \subset V$ is contained in $b^{-1}\varphi^{-1}(y)$, which is étale over $\varphi^{-1}(y)$ because b is étale. In particular, since $\varphi^{-1}(y)$ is finite, the same is true for V_u. Therefore $V \to U$ is a map of smooth varieties with fibres of the same dimension (zero); by "miracle flatness" [EGA IV$_3$ 1966, Proposition 15.4.2], it is flat; therefore φ is flat. On the other hand, the geometric fibres of φ are the symplectic groups $\mathrm{Sp}(2g, \mathbb{Z}/n\mathbb{Z})$, and they are reduced by our choice of n (see (2-1) in Section 2A1). Hence φ is smooth of relative dimension zero, that is, étale. The same argument applies to the map ψ, with the symplectic group replaced by $\mathrm{Sp}(2g, \mathbb{Z}/n\mathbb{Z})/\pm\mathbb{1}$. □

Remark 2.6. The maps $M_g^{(n)} \to M_g$ and $A_g^{(n)} \to A_g$ down to the coarse moduli schemes are still finite Galois covers, but they are not étale.

By Lemma 2.5, we can identity the tangent space to a point $[C, \alpha] \in M_g^{(n)}$ with the tangent space to its image $[C] \in \mathcal{M}_g$ under φ, and similarly on the abelian variety side. Moreover, the cartesian diagram

$$\begin{array}{ccc} M_g^{(n)} & \xrightarrow{j_n} & A_g^{(n)} \\ {\scriptstyle\varphi}\downarrow & \square & \downarrow{\scriptstyle\psi} \\ \mathcal{M}_g & \xrightarrow[\tau_g]{} & \mathcal{A}_g \end{array} \tag{2-4}$$

allows us to identify the map

$$\sigma\colon H^1(C, T_C) \to \operatorname{Sym}^2 H^1(C, \mathscr{O}_C),$$

already appeared in (1-1), with the tangent map of j_n at a point $[C, \alpha]$. As we already mentioned, in [Oort and Steenbrink 1980, Section 2] it is shown that if C is hyperelliptic the kernel of σ has dimension $g-2$.

3. Proof of the main theorem

3A. ***Proof of Theorem 1.*** Let C be a hyperelliptic curve of genus $g \geq 3$ and let J be its Jacobian. Fix an Abel–Jacobi embedding $C \hookrightarrow J$ and let

$$H := \operatorname{Hilb}_{C/J}$$

be the Hilbert scheme component containing such embedding as a point. Let

$$\begin{array}{ccc} \mathcal{Z} & \xhookrightarrow{\iota} & H \times J \\ \downarrow & \swarrow{\scriptstyle \mathrm{pr}_1} & \\ H & & \end{array}$$

be the universal family over the Hilbert scheme.

Lemma 3.1. *The restriction morphism*

$$\iota^*\colon \operatorname{Pic}^0(H \times J/H) \to \operatorname{Pic}^0(\mathcal{Z}/H)$$

is an isomorphism of abelian schemes over H.

Proof. We use the *critère de platitude par fibres* [EGA IV$_3$ 1966, Théorème 11.3.10] in the following special case: suppose given a scheme S and an S-morphism $f\colon X \to Y$ such that

(a) X/S is finitely presented and flat,

(b) Y/S is locally of finite type, and

(c) $f_s\colon X_s \to Y_s$ is flat for each $s \in S$. Then f is flat.

Applying this to $(S, f) = (H, \iota^*)$, we conclude that ι^* is flat. But $\operatorname{Pic}^0(H \times J/H)$ is isomorphic, over H, to the constant abelian scheme $H \times J$, and ι^* is an isomorphism on each fibre over H. Therefore it is a flat, unramified and bijective morphism, hence an isomorphism. □

Let α be a fixed level-n structure on J, with $n \geq 3$ chosen as in Section 2A1. Form the constant level structure α_H on the abelian scheme $H \times J \to H$. Transferring the level structure α_H from $H \times J$ to $\mathrm{Pic}^0(\mathcal{Z}/H)$ using the isomorphism ι^* of Lemma 3.1, we can now regard $\mathcal{Z} \to H$ as a family of Abel–Jacobi curves with level-n structure. Since $M_g^{(n)}$ is a *fine* moduli space for these objects, we obtain a morphism

$$f\colon H \to M_g^{(n)}. \tag{3-1}$$

Note that the topological image of f is just the point $x \in M_g^{(n)}$ corresponding to $[C, \alpha]$. The tangent map $\mathrm{d}f$ at the point $[C] \in H$ is the connecting homomorphism

$$\partial\colon H^0(C, N_C) \to H^1(C, T_C),$$

already appeared in (1-1).

Our next goal is to view the Hilbert scheme H over a suitable artinian scheme R_g. Recall the Torelli type morphism j_n introduced in (2-2). We define

$$R_g \subset M_g^{(n)}$$

to be the scheme-theoretic fibre of j_n over the moduli point $[J, \alpha] \in A_g^{(n)}$. Let $y \in V^{(n)}$ be the image of the point $x = [C, \alpha]$ under the quotient map

$$M_g^{(n)} \to V^{(n)} = M_g^{(n)}/\Sigma,$$

where Σ is the involution first appeared in (2-3). During the proof of [Oort and Steenbrink 1980, Corollary 3.2] it is shown that one can choose local coordinates $t_1, \ldots, t_{3g-3}$ around x such that $\Sigma^* t_i = t_i$ if $i = 1, \ldots, 2g-1$ and $\Sigma^* t_i = -t_i$ if $i = 2g, \ldots, 3g-3$. Oort–Steenbrink deduce that

$$\hat{\mathscr{O}}_y = \hat{\mathscr{O}}_x^{\Sigma} = k[\![t_1, \ldots, t_{2g-1}, t_{2g}^2, t_{2g}t_{2g+1}, \ldots, t_{3g-3}^2]\!]. \tag{3-2}$$

Since we have a factorisation

$$j_n\colon M_g^{(n)} \to V^{(n)} \xhookrightarrow{\ \iota\ } A_g^{(n)}$$

where ι is an *immersion* by Theorem 2.4, we deduce from (3-2) that

$$R_g = \mathrm{Spec}\, k[s_1, \ldots, s_{g-2}]/\mathfrak{m}^2,$$

where $\mathfrak{m} = (s_1, \ldots, s_{g-2}) \subset k[s_1, \ldots, s_{g-2}]$. For instance, R_3 is the scheme of dual numbers $k[s]/s^2$, and if $g = 4$ we get the triple point $k[s,t]/(s^2, st, t^2)$.

Recall the cohomology sequence

$$0 \to H^1(C, \mathscr{O}_C) \to H^0(C, N_C) \xrightarrow{\ \partial\ } H^1(C, T_C) \xrightarrow{\ \sigma\ } H^1(C, \mathscr{O}_C)^{\otimes 2}, \tag{3-3}$$

where σ factors through $\mathrm{Sym}^2 H^1(C, \mathscr{O}_C)$, the tangent space of $\mathcal{A}_g$ at $[J, \Theta_C]$. Since C is hyperelliptic, the image of ∂ has dimension $g - 2 > 0$. In other words, the differential $\partial = \mathrm{d}f$, where f was defined

in (3-1), does not vanish at the point $[C] \in H$. Thus f is not scheme-theoretically constant, although $x = [C, \alpha] \in M_g^{(n)}$ is the only point in the image. On the other hand, the composition

$$j_n \circ f \colon H \to M_g^{(n)} \to A_g^{(n)}$$

is the constant morphism since its differential is identically zero. Indeed the composition

$$\sigma \circ \partial \colon H^0(C, N_C) \to H^1(C, T_C) \to \operatorname{Sym}^2 H^1(C, \mathscr{O}_C)$$

vanishes by exactness of (3-3). So the image point $[J, \alpha]$ does not deform even at first order, and we conclude that f factors through the scheme-theoretic fibre of j_n. This gives us a morphism

$$\pi \colon H \to R_g. \tag{3-4}$$

We will exploit the following technical lemma:

Lemma 3.2 [Kollár 1996, Lemma 1.10.1]. *Let R be the spectrum of a local ring, $p \colon U \to V$ a morphism over R, with $U \to R$ flat and proper. If the restriction $p_0 \colon U_0 \to V_0$ of p over the closed point $0 \in R$ is an isomorphism, then p is an isomorphism.*

Recall that $J = H_{\mathrm{red}}$, so we have a closed immersion $J \hookrightarrow H$ (with empty complement). Consider the closed point $0 \in J$ corresponding to C. Let us fix a regular sequence $f_1, \dots, f_g$ in the maximal ideal of $\mathscr{O}_{J,0}$. Choose lifts $\tilde{f}_i \in \mathscr{O}_{H,0}$ along the natural surjection $\mathscr{O}_{H,0} \twoheadrightarrow \mathscr{O}_{J,0}$, for $i = 1, \dots, g$. Then we consider the zero scheme

$$i \colon S_g = Z(\tilde{f}_1, \dots, \tilde{f}_g) \hookrightarrow H, \tag{3-5}$$

the largest artinian scheme supported at $0 \in H$. We next show that the composition

$$\rho = \pi \circ i \colon S_g \hookrightarrow H \to R_g \tag{3-6}$$

is an isomorphism, where π is defined in (3-4). We will need the following lemma:

Lemma 3.3. *Let $\ell \colon k[x_1, \dots, x_d]/\mathfrak{m}^2 \twoheadrightarrow B$ be a surjection of local Artin k-algebras such that the differential $\mathrm{d}\ell$ is an isomorphism. Then ℓ is an isomorphism.*

Proof. Since $\mathrm{d}\ell$ is an isomorphism by assumption, B has embedding dimension d, hence it can be written as a quotient $k[x_1, \dots, x_d]/I$, so that its maximal ideal is $\mathfrak{m}_B = \mathfrak{m}/I$. Starting from the surjection ℓ, it is then clear that $\mathfrak{m}^2 \subset I$, and we have to show the other inclusion. This follows from the chain of isomorphisms

$$\mathfrak{m}/\mathfrak{m}^2 \xrightarrow{\sim} \mathfrak{m}_B/\mathfrak{m}_B^2 = \frac{\mathfrak{m}/I}{(\mathfrak{m}/I)^2} = \frac{\mathfrak{m}/I}{\mathfrak{m}^2/I \cap \mathfrak{m}^2} = \frac{\mathfrak{m}/\mathfrak{m}^2}{I/\mathfrak{m}^2},$$

where the first isomorphism is $(\mathrm{d}\ell)^\vee$. □

Lemma 3.4. *The tangent map $\mathrm{d}\rho \colon T_{S_g} \to T_{R_g}$ is an isomorphism.*

Proof. The kernel of $H^1(C, T_C) \to H^1(C, \mathscr{O}_C)^{\otimes 2}$, namely the image of $\partial\colon H^0(C, N_C) \to H^1(C, T_C)$, is the tangent space T_{R_g} to the artinian scheme R_g, as the latter is by definition the fibre of j_n. We then have a direct sum decomposition $T_0 H = T_0 J \oplus T_{R_g}$. The intersection of S_g and J inside H is the reduced origin $0 \in J$, so the linear subspace $T_{S_g} \subset T_0 H$ intersects $T_0 J$ trivially, which implies that the tangent map

$$\mathrm{d}\rho\colon T_{S_g} \subset T_0 J \oplus T_{R_g} \to T_{R_g}$$

is injective. On the other hand, the inclusion $T_{S_g} \subset T_0 H$ is cut out by independent linear functions, again because $T_{S_g} \cap T_0 J = (0)$. It follows that the linear inclusion $T_{S_g} \subset T_0 H$ has codimension equal to $\dim T_0 J = g$, thus

$$\dim T_{S_g} = \dim T_0 H - g = g - 2 = \dim T_{R_g}.$$

The result follows. □

Corollary 3.5. *The map $\rho\colon S_g \to R_g$ of (3-6) is an isomorphism.*

Proof. The map ρ is proper, injective on points and, by Lemma 3.4, injective on tangent spaces. Then it is a closed immersion; in fact, by Lemma 3.4 again, it is an isomorphism on tangent spaces, so by Lemma 3.3 it is an isomorphism. □

The corollary yields a section of π,

$$s = i \circ \rho^{-1}\colon R_g \xrightarrow{\sim} S_g \hookrightarrow H,$$

which finally allows us to prove the main result of this paper.

Theorem 3.6. *Let C he a hyperelliptic curve of genus $g \geq 2$, and let J be its Jacobian. Then there is an isomorphism of schemes*

$$J \times R_g \xrightarrow{\sim} H.$$

Proof. If $g = 2$, the Hilbert scheme is nonsingular because $\partial\colon H^0(C, N_C) \to H^1(C, T_C)$, the connecting homomorphism in (1-1), vanishes. If $g \geq 3$, consider the translation action $\mu\colon J \times H \to H$ by J on the Hilbert scheme and the composition

$$J \times R_g \xrightarrow{\mathrm{id}_J \times s} J \times H \xrightarrow{\mu} H,$$

viewed as a morphism over the artinian scheme R_g. Since it restricts to the identity id_J over the closed point of R_g, by Lemma 3.2 it must be an isomorphism. □

3B. ***Relation between Hilbert scheme and Torelli.*** Let $z = [J, \Theta_C]$ be a point in the image of the Torelli morphism $\tau_g\colon \mathcal{M}_g \to \mathcal{A}_g$. The fibre of τ_g over $\operatorname{Spec} k(z) \to \mathcal{A}_g$ is, topologically, just a point, by Torellli's theorem. This point is scheme-theoretically reduced if C is nonhyperelliptic. However, thanks to the cartesian diagram (2-4), what we can observe is that $\tau_g^{-1}(z) = \mathcal{M}_g \times_{\mathcal{A}_g} \operatorname{Spec} k(z) \subset \mathcal{M}_g$ is the artinian scheme R_g when z represents a hyperelliptic Jacobian. Theorem 3.6 thus fully develops in a qualitative form the idea already present in [Lange and Sernesi 2004], namely that understanding the ramification

(the fibres) of the Torelli morphism is equivalent to understanding the singularities of the Hilbert scheme; what the present work shows is that these singularities are controlled by the artinian scheme R_g.

The results proved so far essentially show the following:

Proposition 3.7. *Let C be a smooth curve of genus $g \geq 2$, and let J be its Jacobian. Then $\tau_g^{-1}([J, \Theta_C])$ is isomorphic to the largest closed subscheme of* $\mathrm{Hilb}_{C/J}$ *supported at* $[\mathsf{aj}\colon C \hookrightarrow J]$.

Proof. In the nonhyperelliptic case, we have $\tau_g^{-1}([J, \Theta_C]) \cong \operatorname{Spec} k$, because τ_g is unramified at $[C]$. The result then follows because $J \to \mathrm{Hilb}_{C/J}$ is an isomorphism (by Corollary 1.3). In the hyperelliptic case we get, using Corollary 3.5,

$$S_g \xrightarrow{\sim} R_g = \tau_g^{-1}([J, \Theta_C]),$$

where $S_g \subset \mathrm{Hilb}_{C/J}$, introduced in (3-5), is precisely the largest subscheme of the Hilbert scheme supported at $[\mathsf{aj}\colon C \hookrightarrow J]$. □

3C. ***Donaldson–Thomas invariants for Jacobians.*** Let C be a smooth complex projective curve of genus 3. One can study the "C-local Donaldson–Thomas invariants" of the abelian 3-fold $J = \mathrm{Pic}^0 C$. As explained in [Ricolfi 2018a; 2018b], these invariants are completely determined by the "BPS number" of the curve,

$$n_C = \nu_H(\mathscr{I}_C) \in \mathbb{Z},$$

in the sense that their generating function is equal to the rational function

$$n_C \cdot q^{-2}(1+q)^4.$$

Here $\nu_H\colon \mathrm{Hilb}_{C/J} \to \mathbb{Z}$ is the Behrend function of the Hilbert scheme. The Behrend function attached to a general finite type $\mathbb{C}$-scheme X is an invariant of the singularities of X. It was introduced in [Behrend 2009] and is now a key tool in Donaldson–Thomas theory. For a smooth scheme Y one has that ν_Y is the constant $(-1)^{\dim Y}$, and moreover $\nu_{X\times Y} = \nu_X \cdot \nu_Y$ for two complex schemes X and Y. While for nonhyperelliptic C we have $n_C = -1$ (because the Hilbert scheme is a copy of the smooth 3-fold J), the structure result

$$\mathrm{Hilb}_{C/J} = J \times \operatorname{Spec} \mathbb{C}[s]/s^2$$

in the hyperelliptic case yields $n_C = -2$, because the scheme of dual numbers has Behrend function $\nu_{R_3} = 2$.

4. An application to moduli spaces of Picard sheaves

Mukai [1981] introduced his celebrated Fourier transform, and gave an application to the moduli space of Picard sheaves on Jacobians of curves. We now review his results on nonhyperelliptic Jacobians and extend them to the hyperelliptic case. We assume that the base field k is, as ever, algebraically closed of characteristic different from 2.

We let $\Phi\colon \mathrm{D}^b(\hat{J}) \to \mathrm{D}^b(J)$ be the Fourier transform with kernel the Poincaré line bundle $\mathscr{P} \in \mathrm{Pic}(\hat{J} \times J)$. If $\hat{\mathsf{p}}\colon \hat{J} \times J \to \hat{J}$ and $\hat{\mathsf{p}}\colon \hat{J} \times J \to J$ are the projections, by definition one has

$$\Phi(\mathscr{E}) = R\mathsf{p}_*(\hat{\mathsf{p}}^*\mathscr{E} \otimes \mathscr{P}).$$

We will denote by $\Phi^i(\mathscr{E})$ the i-th cohomology sheaf of the complex $\Phi(\mathscr{E})$.

Let $p_0 \in C$ be a point on a smooth curve of genus $g \geq 2$. Let us form the line bundle $\xi = \mathscr{O}_C(dp_0)$. From now on we view it as a sheaf on $\hat{J}$ by pushing it forward along the Abel–Jacobi map $\mathsf{aj}\colon C \hookrightarrow J$ followed by the identification of J with its dual. Applying his Fourier transform, Mukai constructs

$$F = \Phi^1(\mathsf{aj}_*\xi), \tag{4-1}$$

a *Picard sheaf* of rank $g-d-1$ living on J. Assume $1 \leq d \leq g-1$, so that by [Mukai 1981, Lemma 4.9] we know that F is simple (that is, $\mathrm{End}_{\mathscr{O}_J}(F) = k$), and

$$\dim \mathrm{Ext}^1_{\mathscr{O}_J}(F, F) = \begin{cases} 2g & \text{if } C \text{ is not hyperelliptic,} \\ 3g-2 & \text{if } C \text{ is hyperelliptic.} \end{cases} \tag{4-2}$$

Let Spl_J be the moduli space of simple coherent sheaves on J, and let $M(F) \subset \mathrm{Spl}_J$ be the connected component containing the point corresponding to F. It is shown in [loc. cit., Theorem 4.8] that if $g = 2$ or C is nonhyperelliptic, the morphism

$$f\colon \hat{J} \times J \to M(F), \quad (\eta, x) \mapsto \mathsf{t}_x^*F \otimes \mathscr{P}_\eta, \tag{4-3}$$

is an isomorphism. By (4-2), the space $M(F)$ is reduced precisely when C has genus 2 or is nonhyperelliptic. For C hyperelliptic, f turns out to be an isomorphism onto the reduction $M(F)_{\mathrm{red}} \subsetneq M(F)$, as Mukai showed [1987, Example 1.15].

Remark 4.1. The moduli space $M(F)$ is a priori only an algebraic space. But an algebraic space is a scheme if and only if its reduction is a scheme. Therefore $M(F)$ is a scheme because of the isomorphism $\hat{J} \times J \cong M(F)_{\mathrm{red}}$.

The following result, which can be seen as a corollary of Theorem 3.6, completes the study of Picard sheaves on Jacobians considered by Mukai, namely those of rank $g-d-1$, with $d \leq g-1$.

Theorem 4.2. *Let C be a hyperelliptic curve of genus $g \geq 2$. Let J be its Jacobian and F a Picard sheaf as above. Then, as schemes,*

$$M(F) = \hat{J} \times J \times R_g.$$

Proof. The case $g = 2$ is already covered by Mukai's tangent space calculation. By Theorem 3.6, it is enough to exhibit an isomorphism $\hat{J} \times H \xrightarrow{\sim} M(F)$, where as usual $H \subset \mathrm{Hilb}_J$ is the Hilbert scheme component containing the Abel–Jacobi point $[C]$. We will do this by extending the morphism (4-3)

defined by Mukai, that is, completing the diagram

$$\begin{array}{ccc} \hat{J} \times J & \xrightarrow{\sim} & M(F)_{\mathrm{red}} \\ \downarrow & & \downarrow \\ \hat{J} \times H & \overset{\phi}{\dashrightarrow} & M(F) \end{array} \tag{4-4}$$

and showing that the extension ϕ is an isomorphism. Recall that via the identification $J = H_{\mathrm{red}}$ we can identify a k-valued point $x \in J(k)$ with a k-valued point of H. Also, for any such point $x \in J \subset H$, we will use the notation $x + p_0$ for the point on the Abel–Jacobi curve $\mathsf{t}_x C \subset J$ obtained by translating $p_0 \in C \subset J$ via the automorphism $\mathsf{t}_x : J \to J$. Let

$$\mathcal{Z} \overset{\iota}{\hookrightarrow} H \times J \to H$$

be the universal family of the Hilbert scheme: the fibre of $\mathcal{Z} \to H$ over $\operatorname{Spec} k(x) \hookrightarrow H$ is the subscheme $\mathsf{t}_x C \subset J$, and ι, the universal Abel–Jacobi map, restricts to $\mathsf{aj} \circ \mathsf{t}_{-x} : \mathsf{t}_x C \to C \hookrightarrow \{x\} \times J$ over the point $x \in H$. We now construct a section σ of $\mathcal{Z} \to H$ restricting to the divisor dp_0 on C (in other words, a "universal" version of ξ). If $q : H \to J$ denotes the projection (forgetting the nonreduced structure) and $u : J \to J$ is the composition $\mathsf{t}_{dp_0} \circ [d]$, the section σ is the map

$$\sigma : H \xrightarrow{(1_H, q)} H \times J \xrightarrow{1_H \times u} H \times J, \quad x \mapsto (x, d(x + p_0)).$$

Here we view $d(x + p_0)$ as a degree d divisor on the translated Abel–Jacobi curve $\mathsf{t}_x C \subset J$, in particular the image of σ clearly lands inside $\mathcal{Z}$. Let $\mathscr{L} = \mathscr{O}_{\mathcal{Z}}(\sigma)$ be the associated line bundle on the total space $\mathcal{Z}$. Then, by construction, restricting $\mathscr{L}$ to a fibre of $\mathcal{Z} \to H$ we get

$$\mathscr{L}|_{\mathsf{t}_x C} = \mathscr{O}_{\mathsf{t}_x C}(d(x + p_0)) = \mathsf{t}_{-x}^* \xi. \tag{4-5}$$

If we consider the pushforward $\iota_* \mathscr{L}$ to $H \times J$, using (4-5) we obtain

$$(\iota_* \mathscr{L})|_{x \times J} = (\mathsf{aj} \circ \mathsf{t}_{-x})_* (\mathscr{L}|_{\mathsf{t}_x C}) = \mathsf{aj}_* \xi. \tag{4-6}$$

Note that $\mathscr{L}$ is flat over H (because $\mathcal{Z} \to H$ is flat), therefore the same is true for $\iota_* \mathscr{L}$. Since taking the Fourier–Mukai transform commutes with base change, (4-6) yields

$$\Phi^1(\iota_* \mathscr{L})|_{x \times J} = \Phi^1(\mathsf{aj}_* \xi) = F. \tag{4-7}$$

Now we consider the following diagram:

$$\begin{array}{ccccccc} & & (\hat{J} \times J) \times J & \xrightarrow{\sim} & (J \times J) \times J & \xrightarrow{m \times \mathrm{id}_J} & J \times J & \xrightarrow{\mathrm{pr}_1} & J \\ & & \downarrow{\scriptstyle i} & & \downarrow & & \downarrow & & \\ \hat{J} \times J & \xleftarrow{\mathrm{pr}_{13}} & (\hat{J} \times H) \times J & \xrightarrow{\sim} & (J \times H) \times J & \xrightarrow{\mu \times \mathrm{id}_J} & H \times J & & \end{array}$$

where m and μ are the translation actions by J on J and H respectively. The Fourier–Mukai transform $\Phi^1(\iota_* \mathscr{L})$ lives on $H \times J$ and is flat over H, by flatness of $\iota_* \mathscr{L}$. By (4-7), we know that the families of

sheaves $\Phi^1(\iota_*\mathscr{L})|_{J\times J}$ and $\mathrm{pr}_1^*\,F$ (both flat over J) define the same morphism $J\to M(F)$, namely the constant morphism hitting the point $[F]$. Since Mukai's morphism $\hat{J}\times J\to M(F)$, defined in (4-3), corresponds (after identifying J with its dual) to the family of sheaves

$$(m\times \mathrm{id}_J)^*\,\mathrm{pr}_1^*\,F\otimes(\mathrm{pr}_{13}\circ i)^*\mathscr{P},$$

it follows that the family

$$(\mu\times \mathrm{id}_J)^*\Phi^1(\iota_*\mathscr{L})\otimes \mathrm{pr}_{13}^*\,\mathscr{P}$$

defines an extension $\phi\colon \hat{J}\times H\to M(F)$, completing diagram (4-4). We know that ϕ is an isomorphism around $[\xi]\mapsto[F]$. Indeed, ϕ is precisely the morphism constructed by Mukai [1987, Proposition 1.12], where he proves that $M(\xi)$ and $M(F)$ are isomorphic along a Zariski open subset. The construction is homogeneous, in the sense that ϕ does not depend on the initial point $[\xi]\in M(\xi)$. Therefore ϕ is globally an isomorphism, as claimed. □

Remark 4.3. The connected component $M(\xi)$ of the moduli space of simple sheaves containing the point $[\xi]$ is the relative Picard variety $\mathrm{Pic}^d(\mathcal{Z}/H)$, which can be identified with $\hat{J}\times H$ by Lemma 3.1. It is possible to adapt the proof of [Mukai 1987, Proposition 1.12] to show that the birational map

$$\mathrm{Pic}^d(\mathcal{Z}/H)\dashrightarrow M(F)$$

is everywhere defined (and an isomorphism), giving an immediate proof of Theorem 4.2. We preferred to present the argument above, because the construction makes the isomorphism $\phi\colon \hat{J}\times H\to M(F)$ arise directly, as a "thickening" of Mukai's isomorphism $\hat{J}\times J\to M(F)_{\mathrm{red}}$. Moreover the argument makes explicit use of (the properties of) the Fourier–Mukai transform.

Acknowledgements

I want to thank Alberto Collino for generously sharing his insight and ideas on the problem. I owe an intellectual debt to Richard Thomas, who patiently explained me how to make part of the argument work. I also thank Martin Gulbrandsen, Michał Kapustka, Aaron Landesman and Filippo Viviani for helpful discussions, and the anonymous referees for suggesting several improvements.

References

[Andreotti 1958] A. Andreotti, "On a theorem of Torelli", *Amer. J. Math.* **80**:4 (1958), 801–828. MR Zbl

[Arbarello et al. 1985] E. Arbarello, M. Cornalba, P. A. Griffiths, and J. Harris, *Geometry of algebraic curves, I*, Grundlehren der Math. Wissenschaften **267**, Springer, 1985. MR Zbl

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

[Bryan et al. 2018] J. Bryan, G. Oberdieck, R. Pandharipande, and Q. Yin, "Curve counting on abelian surfaces and threefolds", *Algebr. Geom.* **5**:4 (2018), 398–463. MR Zbl

[Ceresa 1983] G. Ceresa, "C is not algebraically equivalent to C^- in its Jacobian", *Ann. of Math.* (2) **117**:2 (1983), 285–291. MR Zbl

[Collino 1984] A. Collino, "A new proof of the Ran–Matsusaka criterion for Jacobians", *Proc. Amer. Math. Soc.* **92**:3 (1984), 329–331. MR Zbl

[Deligne and Mumford 1969] P. Deligne and D. Mumford, "The irreducibility of the space of curves of given genus", *Inst. Hautes Études Sci. Publ. Math.* **36** (1969), 75–109. MR Zbl

[EGA IV$_3$ 1966] A. Grothendieck, "Eléments de géométrie algébrique, IV: Étude locale des schémas et des morphismes de schémas, III", *Inst. Hautes Études Sci. Publ. Math.* **28** (1966), 5–255. MR Zbl

[Griffiths 1967] P. A. Griffiths, "Some remarks and examples on continuous systems and moduli", *J. Math. Mech.* **16** (1967), 789–802. MR Zbl

[Kollár 1996] J. Kollár, *Rational curves on algebraic varieties*, Ergebnisse der Mathematik (3) **32**, Springer, 1996. MR Zbl

[Landesman 2019] A. Landesman, "The infinitesimal Torelli problem", preprint, 2019. arXiv

[Lange and Sernesi 2004] H. Lange and E. Sernesi, "On the Hilbert scheme of a Prym variety", *Ann. Mat. Pura Appl.* (4) **183**:3 (2004), 375–386. MR Zbl

[Mukai 1981] S. Mukai, "Duality between $D(X)$ and $D(\hat{X})$ with its application to Picard sheaves", *Nagoya Math. J.* **81** (1981), 153–175. MR Zbl

[Mukai 1987] S. Mukai, "Fourier functor and its application to the moduli of bundles on an abelian variety", pp. 515–550 in *Algebraic geometry* (Sendai, Japan, 1985), edited by T. Oda, Adv. Stud. Pure Math. **10**, North-Holland, Amsterdam, 1987. MR Zbl

[Mumford 1965] D. Mumford, *Geometric invariant theory*, Ergebnisse der Mathematik (2) **34**, Springer, 1965. MR Zbl

[Oort and Steenbrink 1980] F. Oort and J. Steenbrink, "The local Torelli problem for algebraic curves", pp. 157–204 in *Journées de géometrie algébrique d'Angers* (Angers, France, 1979), edited by A. Beauville, Sijthoff & Noordhoff, Alphen aan den Rijn, Netherlands, 1980. MR Zbl

[Pagani et al. 2018] N. Pagani, A. T. Ricolfi, and J. van Zelm, "Pullbacks of universal Brill–Noether classes via Abel–Jacobi morphisms", 2018. To appear in *Math. Nachr.* arXiv

[Popp 1977] H. Popp, *Moduli theory and classification theory of algebraic varieties*, Lecture Notes in Math. **620**, Springer, 1977. MR Zbl

[Ricolfi 2018a] A. T. Ricolfi, "The DT/PT correspondence for smooth curves", *Math. Z.* **290**:1-2 (2018), 699–710. MR Zbl

[Ricolfi 2018b] A. T. Ricolfi, "Local contributions to Donaldson–Thomas invariants", *Int. Math. Res. Not.* **2018**:19 (2018), 5995–6025. MR Zbl

Communicated by Ravi Vakil
Received 2018-09-14 Revised 2019-12-31 Accepted 2020-02-10

aricolfi@sissa.it *Max Planck Institut für Mathematik, Bonn, Germany*
Current address: *Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy*

msp

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1399

Endomorphism algebras of geometrically split abelian surfaces over $\mathbb{Q}$

Francesc Fité and Xavier Guitart

We determine the set of geometric endomorphism algebras of geometrically split abelian surfaces defined over $\mathbb{Q}$. In particular we find that this set has cardinality 92. The essential part of the classification consists in determining the set of quadratic imaginary fields M with class group $\mathrm{C}_2 \times \mathrm{C}_2$ for which there exists an abelian surface A defined over $\mathbb{Q}$ which is geometrically isogenous to the square of an elliptic curve with CM by M. We first study the interplay between the field of definition of the geometric endomorphisms of A and the field M. This reduces the problem to the situation in which E is a $\mathbb{Q}$-curve in the sense of Gross. We can then conclude our analysis by employing Nakamura's method to compute the endomorphism algebra of the restriction of scalars of a Gross $\mathbb{Q}$-curve.

1. Introduction

Let A be an abelian variety of dimension $g \geq 1$ defined over a number field k of degree d. Let us denote by $A_{\overline{\mathbb{Q}}}$ its base change to $\overline{\mathbb{Q}}$. We refer to $\mathrm{End}(A_{\overline{\mathbb{Q}}})$, the $\mathbb{Q}$-algebra spanned by the endomorphisms of A defined over $\overline{\mathbb{Q}}$, as the $\overline{\mathbb{Q}}$-endomorphism algebra of A. For a fixed choice of g and d, it is conjectured that the set of possibilities for $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ is finite. A slightly stronger form of this conjecture, applying to endomorphism rings of abelian varieties over number fields, has been attributed to Coleman in [Bruin et al. 2006].

Hereafter, let A denote an abelian surface defined over $\mathbb{Q}$. In the case that A is geometrically simple (that is, $A_{\overline{\mathbb{Q}}}$ is simple), the previous conjecture stands widely open. If A is principally polarized and has CM it has been shown by Murabayashi and Umegaki [2001] that $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ is one of 19 possible quartic CM fields. However, narrowing down to a finite set the possible quadratic real fields and quaternion division algebras over $\mathbb{Q}$ which occur as $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ for some A has escaped all attempts of proof. See also [Orr and Skorobogatov 2018] for recent more general results which prove Coleman's conjecture for CM abelian varieties.

In the present paper, we focus on the case that A is geometrically split, that is, the case in which $A_{\overline{\mathbb{Q}}}$ is isogenous to a product of elliptic curves, which we will assume from now on. Let $\mathcal{A}$ be the set of possibilities for $\mathrm{End}(A_{\overline{\mathbb{Q}}})$, where A is a geometrically split abelian surface over $\mathbb{Q}$.

Let us briefly recall how scattered results in the literature ensure the finiteness of $\mathcal{A}$ (we will detail the arguments in Section 4). Indeed, if $A_{\overline{\mathbb{Q}}}$ is isogenous to the product of two nonisogenous elliptic curves, then the finiteness (and in fact the precise description) of the set of possibilities for $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ follows

MSC2010: primary 11G18; secondary 11G15, 14K22.
Keywords: products of CM elliptic curves, Coleman's conjecture, endomorphism algebras, singular abelian surfaces.

from [Fité et al. 2012, Proposition 4.5]. If, on the contrary, $A_{\overline{\mathbb{Q}}}$ is isogenous to the square of an elliptic curve, then the finiteness of the set of possibilities for $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ was established by Shafarevich [1996] (see also [González 2011] for the determination of the precise subset corresponding to modular abelian surfaces). In the present work, we aim at an effective version of Shafarevich's result. Our starting point is [Fité and Guitart 2018a, Theorem 1.4], which we recall in our particular setting.

Theorem 1.1 [Fité and Guitart 2018a]. *If A is an abelian surface defined over $\mathbb{Q}$ such that $A_{\overline{\mathbb{Q}}}$ is isogenous to the square of an elliptic curve $E/\overline{\mathbb{Q}}$ with complex multiplication (CM) by a quadratic imaginary field M, then the class group of M is 1, C_2, or $\mathrm{C}_2 \times \mathrm{C}_2$.*

It should be noted that several other works can be used to see that, in the situation of the theorem, the exponent of the class group of M divides 2 (see [Schütt 2007; Kani 2011], for example).

While it is an easy observation that an abelian surface A as in the theorem can be found for each quadratic imaginary field M with class group 1 or C_2 (see [Fité and Guitart 2018a, Remark 2.20] and also Section 4), the question whether such an A exists for each of the fields M with class group $\mathrm{C}_2 \times \mathrm{C}_2$ is far from trivial. The aforementioned results are thus not sufficient for the determination of the set $\mathcal{A}$. The main contribution of this article is the following theorem.

Theorem 1.2. *Let M be a quadratic imaginary field with class group $\mathrm{C}_2 \times \mathrm{C}_2$. There exists an abelian surface defined over $\mathbb{Q}$ such that $A_{\overline{\mathbb{Q}}}$ is isogenous to the square of an elliptic curve $E/\overline{\mathbb{Q}}$ with CM by M if and only if the discriminant of M belongs to the set*

$$\begin{aligned}\big\{&-84, -120, -132, -168, -228, -280, -372, -408, -435,\\ &-483, -520, -532, -595, -627, -708, -795, -1012, -1435\big\}.\end{aligned} \tag{1-1}$$

The only imaginary quadratic fields with class group $\mathrm{C}_2 \times \mathrm{C}_2$ whose discriminant does not belong to (1-1) are

$$\mathbb{Q}(\sqrt{-195}),\quad \mathbb{Q}(\sqrt{-312}),\quad \mathbb{Q}(\sqrt{-340}),\quad \mathbb{Q}(\sqrt{-555}),\quad \mathbb{Q}(\sqrt{-715}),\quad \mathbb{Q}(\sqrt{-760}). \tag{1-2}$$

With Theorem 1.2 at hand, the determination of the set $\mathcal{A}$ follows as a mere corollary (see Section 4 for the proof).

Corollary 1.3. *The set $\mathcal{A}$ of $\overline{\mathbb{Q}}$-endomorphism algebras of geometrically split abelian surfaces over $\mathbb{Q}$ is made of*:

(i) $\mathbb{Q} \times \mathbb{Q}$, $\mathbb{Q} \times M$, $M_1 \times M_2$, *where M, M_1 and M_2 are quadratic imaginary fields of class number* 1;

(ii) $\mathrm{M}_2(\mathbb{Q})$, $\mathrm{M}_2(M)$, *where M is a quadratic imaginary field with class group* 1, C_2, *or* $\mathrm{C}_2 \times \mathrm{C}_2$ *and distinct from those listed in* (1-2).

In particular, the set $\mathcal{A}$ has cardinality 92.

The paper is organized in the following manner. In Section 2 we attach a c-representation ϱ_V of degree 2 to an abelian surface A defined over $\mathbb{Q}$ such that $A_{\overline{\mathbb{Q}}}$ is isogenous to the square of an elliptic curve $E/\overline{\mathbb{Q}}$ with CM by M. It is well known that E is a $\mathbb{Q}$-curve and that one can associate a 2-cocycle c_E to E.

A c-representation is essentially a representation up to scalar and it is thus a notion closely related to that of projective representation. In the case of the c-representation ϱ_V attached to A, the scalar that measures the failure of ϱ_V to be a proper representation is precisely the 2-cocycle c_E. Choosing the language of c-representations instead of that of projective representations has an unexpected payoff: the tensor product of a c-representation ϱ and its contragradient c-representation ϱ^* is again a proper representation. We show that $\varrho_V \otimes \varrho_V^*$ coincides with the representation of $G_{\mathbb{Q}}$ on the 4-dimensional M-vector space $\mathrm{End}(A_{\overline{\mathbb{Q}}})$. This representation has been studied in detail in [Fité and Sutherland 2014] and the tensor decomposition of $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ is exploited in Theorems 2.20 and 2.27 to obtain obstructions on the existence of A. These obstructions extend to the general case those obtained in [Fité and Guitart 2018a, §3.1, §3.2] under very restrictive hypotheses. The c-representation point of view also allows us to understand in a unified manner what we called *group theoretic* and *cohomological* obstructions in [Fité and Guitart 2018a]. It should be noted that one can define analogues of ϱ_V in other more general situations. For example, a parallel construction in the context of geometrically isotypic abelian varieties potentially of GL_2-type has been exploited in [Fité and Guitart 2019] to determine a tensor factorization of their Tate modules. This can be used to deduce the validity of the Sato–Tate conjecture for them in certain cases.

In Section 3, we describe a method of Nakamura to compute the endomorphism algebra of the restriction of scalars of certain Gross $\mathbb{Q}$-curves (see Definition 2.9 below for the precise definition of these curves). Then we apply this method to all Gross $\mathbb{Q}$-curves with CM by a field M of class group $\mathrm{C}_2 \times \mathrm{C}_2$. This computation plays a key role in the proof of Theorem 1.2, both in proving the existence of the abelian surfaces for the fields M different from those listed in (1-2), and in proving the nonexistence for the fields of (1-2).

In Section 4 we culminate the proofs of Theorem 1.2 and Corollary 1.3 by assembling together the obstructions and existence results from Sections 2 and 3. We essentially show that we can use the results of Section 2 to reduce to the case of Gross $\mathbb{Q}$-curves, and then deal with this case using the results of Section 3.

Notations and terminology. For k a number field, we will work in the category of abelian varieties up to isogeny over k. Note that isogenies become invertible in this category. Given an abelian variety A defined over k, the set of endomorphisms $\mathrm{End}(A)$ of A defined over k is endowed with a $\mathbb{Q}$-algebra structure. More generally, if B is an abelian variety defined over k, we will denote by $\mathrm{Hom}(A, B)$ the $\mathbb{Q}$-vector space of homomorphisms from A to B that are defined over k. We note that for us $\mathrm{End}(A)$ and $\mathrm{Hom}(A, B)$ denote what some other authors call $\mathrm{End}^0(A)$ and $\mathrm{Hom}^0(A, B)$. We will write $A \sim B$ to mean that A and B are isogenous over k. If L/k is a field extension, then A_L will denote the base change of A from k to L. In particular, we will write $A_L \sim B_L$ if A and B become isogenous over L, and we will write $\mathrm{Hom}(A_L, B_L)$ to refer to what some authors write as $\mathrm{Hom}_L(A, B)$.

2. c-representations and k-curves

The goal of this section is to obtain obstructions to the existence of abelian surfaces defined over $\mathbb{Q}$ such that $\mathrm{End}(A_{\overline{\mathbb{Q}}}) \simeq \mathrm{M}_2(M)$, where M is a quadratic imaginary field. To this purpose, we analyze the interplay between the k-curves and c-representations that arise from them.

2A. ***c-representations: general definitions.*** Let V be a vector space of finite dimension over a field k and let G be a finite group. We say that a map

$$\varrho_V : G \to \mathrm{GL}(V)$$

is a c-representation (of the group G) if $\varrho_V(1) = 1$ and there exists a map

$$c_V : G \times G \to k^\times$$

such that for every $\sigma, \tau \in G$ one has

$$\varrho_V(\sigma)\varrho_V(\tau) = \varrho_V(\sigma\tau)c_V(\sigma, \tau). \tag{2-1}$$

Remark 2.1. The following properties follow easily from the definition:

(i) We have

$$\varrho_V(\sigma^{-1}) = \varrho_V(\sigma)^{-1}c_V(\sigma^{-1}, \sigma) \quad \text{and} \quad \varrho_V(\sigma^{-1}) = \varrho_V(\sigma)^{-1}c_V(\sigma, \sigma^{-1}).$$

In particular, $c_V(\sigma, \sigma^{-1}) = c_V(\sigma^{-1}, \sigma)$.

(ii) If $c_V(\,\cdot\,,\,\cdot\,) = 1$, the notion of c-representation corresponds to the usual notion of representation.

Let V and W be c-representations of the group G. Let $T = \mathrm{Hom}(V, W)$ denote the space of k-linear maps from V to W. A homomorphism of c-representations from V to W is a k-linear map $f \in T$ such that

$$f(v) = \varrho_W(\sigma)(f(\varrho_V(\sigma)^{-1}v))$$

for every $v \in V$ and $\sigma \in G$.

Consider now the map

$$\varrho_T : G \to \mathrm{GL}(\mathrm{Hom}(V, W)),$$

defined by

$$(\varrho_T(\sigma)f)(v) = \varrho_W(\sigma)(f(\varrho_V(\sigma)^{-1}v)).$$

Proposition 2.2. *The map ϱ_T together with the map $c_T : G \times G \to k^\times$ defined by $c_T = c_V^{-1} \cdot c_W$ equip T with the structure of a c-representation.*

Before proving the proposition we show a particular case. In the case that W is k equipped with the trivial action of G, let us write $V^* = T$ and $\varrho^* = \varrho_T$. In this case, $\varrho^*(\sigma)$ is the inverse transpose of $\varrho_V(\sigma)$. The assertion of the proposition is then immediate from (2-1).

The following two lemmas, whose proof is straightforward, imply the proposition.

Lemma 2.3. *The maps*

$$\varrho_\otimes : G \to \mathrm{GL}(V \otimes W),$$

defined by $\varrho_\otimes(\sigma)(v \otimes w) = \varrho_V(\sigma)(v) \otimes \varrho_W(\sigma)(w)$ and $c_\otimes = c_V \cdot c_W$ endow $V \otimes W$ with a structure of c-representation.

Lemma 2.4. *The map*

$$\phi : W \otimes V^* \to T$$

defined by $\phi(w \otimes f)(v) = f(v)w$ is an isomorphism of c-representations.

Corollary 2.5. *When $V = W$, the c-representation T is in fact a representation.*

2B. *k-curves: general definitions.* We briefly recall some definitions and results regarding $\mathbb{Q}$-curves and, more generally, k-curves with complex multiplication. More details can be found in [Fité and Guitart 2018a, §2.1] and the references therein (especially [Quer 2000; Ribet 1992; Nakamura 2004]).

Let $E/\overline{\mathbb{Q}}$ be an elliptic curve and let k be a number field, whose absolute Galois group we denote by G_k.

Definition 2.6. We say that E is a *k-curve* if for every $\sigma \in G_k$ there exists an isogeny $\mu_\sigma : {}^\sigma E \to E$.

Definition 2.7. We say that E is a *Ribet k-curve* if E is a k-curve and the isogenies μ_σ can be taken to be compatible with the endomorphisms of E, in the sense that the diagram

$$\begin{array}{ccc} {}^\sigma E & \xrightarrow{\mu_\sigma} & E \\ \downarrow{\scriptstyle {}^\sigma\varphi} & & \downarrow{\scriptstyle \varphi} \\ {}^\sigma E & \xrightarrow{\mu_\sigma} & E \end{array} \tag{2-2}$$

commutes for all $\sigma \in G_k$ and all $\varphi \in \mathrm{End}(E)$.

Remark 2.8. (i) Observe that if E does not have CM, then E is a k-curve if and only if it is a Ribet k-curve. If E has CM (say by a quadratic imaginary field M), it is well known that E is isogenous to all of its Galois conjugates and hence it is always a k-curve; it is a Ribet k-curve if and only if $M \subseteq k$; see [Silverman 1994, Theorem 2.2].

(ii) We warn the reader that in the present paper we are using a slightly different terminology from that of [Fité and Guitart 2018a]: as in [Fité and Guitart 2018a] the only relevant notion was that of a Ribet k-curve, we called Ribet k-curves simply k-curves.

Let K be a number field containing k. We say that an elliptic curve E/K is a k-curve *defined over K* (resp. a Ribet k-curve *defined over K*) if $E_{\overline{\mathbb{Q}}}$ is a k-curve (resp. a Ribet k-curve). We will say that E is *completely defined over K* if, in addition, all the isogenies $\mu_\sigma : {}^\sigma E \to E$ can be taken to be defined over K.

Definition 2.9. Let H denote the Hilbert class field of M and let E/H be an elliptic curve with CM by M. We say that E is a *Gross $\mathbb{Q}$-curve* if E is completely defined over H.

The next proposition characterizes the existence of Gross $\mathbb{Q}$-curves and Ribet M-curves with CM by M defined over the Hilbert class field H.

Proposition 2.10. *Let M be a quadratic imaginary field and let D denote its discriminant. Then*:

(i) *There exists a Ribet M-curve E^* with CM by M and completely defined over H.*

(ii) *There exists a Gross $\mathbb{Q}$-curve E^* with CM by M (and completely defined over H) if and only if D is not of the form*

$$D = -4p_1 \dots p_{t-1}, \tag{2-3}$$

where $t \geq 2$ and $p_1, \dots, p_{t-1}$ are primes congruent to 1 modulo 4.

The first part of the previous proposition is a weaker form of [Shimura 1971, Proposition 5, p. 521] (see also [Nakamura 2001, Remark 1]). For the second part, we refer to [Gross 1980, §11; Nakamura 2004, Proposition 5]. Discriminants of the form (2-3) are called *exceptional*.

Suppose from now on that E is a k-curve defined over K with CM by an imaginary quadratic field M. Fix a system of isogenies $\{\mu_\sigma : {}^\sigma E \to E\}_{\sigma\in G_k}$. By enlarging K if necessary, we can always assume that K/k is Galois and that E is completely defined over K. We will equip $\mathrm{End}(E)$ with the following action. For $\sigma \in \mathrm{Gal}(K/k)$ and $\varphi \in \mathrm{End}(E)$ define

$$\sigma \star \varphi = \mu_\sigma \circ {}^\sigma\varphi \circ \mu_\sigma^{-1}.$$

Note that if E is a Ribet k-curve, then this action is trivial. If we regard M as a $\mathrm{Gal}(K/k)$-module by means of the natural Galois action (which is actually the trivial action when k contains M) and $\mathrm{End}(E)$ endowed with the action defined above, then the identification of $\mathrm{End}(E)$ with M becomes a $\mathrm{Gal}(K/k)$-equivariant isomorphism. The map

$$c_E^K : \mathrm{Gal}(K/k)\times\mathrm{Gal}(K/k) \to M^\times, \quad (\sigma,\tau)\mapsto \mu_{\sigma\tau}\circ{}^\sigma\mu_\tau^{-1}\circ\mu_\sigma^{-1}$$

satisfies the condition

$$(\varrho\star c_E^K(\sigma,\tau))\cdot c_E^K(\varrho\sigma,\tau)^{-1}\cdot c_E^K(\varrho,\sigma\tau)\cdot c_E^K(\varrho,\sigma)^{-1}=1, \tag{2-4}$$

for $\varrho,\sigma,\tau\in\mathrm{Gal}(K/k)$, and is then a 2-cocycle.[1] Denote the cohomology class in $H^2(\mathrm{Gal}(K/k), M^\times)$ corresponding to c_E^K by γ_E^K. The class γ_E^K only depends on the K-isogeny class of E.

The next result is a consequence of Weil's descent criterion, extended to varieties up to isogeny by Ribet [1992, §8].

Theorem 2.11 (Ribet–Weil). *Suppose that E is a Ribet k-curve completely defined over K (and hence $M\subseteq k$). Let L be a number field with $k\subseteq L\subseteq K$, and consider the restriction map*

$$\mathrm{res} : H^2(\mathrm{Gal}(K/k), M^\times)\to H^2(\mathrm{Gal}(K/L), M^\times).$$

If $\mathrm{res}(\gamma_E^K)=1$, *there exists an elliptic curve C/L such that $E\sim C_K$.*

2C. ***M*-curves from squares of CM elliptic curves.** Let M be a quadratic imaginary field. Let A be an abelian surface defined over $\mathbb{Q}$ such that $A_{\overline{\mathbb{Q}}}$ is isogenous to E^2, where E is an elliptic curve defined over $\overline{\mathbb{Q}}$ with CM by M. Let $K/\mathbb{Q}$ denote the minimal extension over which

$$\mathrm{End}(A_{\overline{\mathbb{Q}}})\simeq\mathrm{End}(A_K).$$

By the theory of complex multiplication, K contains the Hilbert class field H of M. Note also that $K/\mathbb{Q}$ is Galois and the possibilities for $\mathrm{Gal}(K/\mathbb{Q})$ can be read from [Fité et al. 2012, Table 8]. For our purposes,

[1] Actually, this is the inverse of the cocycle considered in [Fité and Guitart 2018a], but this does not affect any of the results that we will use.

it is enough to recall that

$$\mathrm{Gal}(K/M) \simeq \begin{cases} \mathrm{C}_r & \text{for } r \in \{1, 2, 3, 4, 6\}, \\ \mathrm{D}_r & \text{for } r \in \{2, 3, 4, 6\}, \\ A_4,\ S_4. \end{cases} \tag{2-5}$$

Here, C_r denotes the cyclic group of r elements, D_r denotes the dihedral group of $2r$ elements, and A_4 (resp. S_4) stands for the alternating (resp. symmetric) group on 4 letters.

We can (and do) assume that E is in fact defined over K, and then we have that $A_K \sim E^2$. For $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ we have that $({}^\sigma E)^2 \sim {}^\sigma A_K = A_K \sim E^2$. Therefore, Poincaré's decomposition theorem implies that E is a $\mathbb{Q}$-curve completely defined over K.

For the purposes of this article, we need to consider the following (slightly more general) situation: Let N/M be a Galois subextension of K/M, and let E^* be a Ribet M-curve which is completely defined over N and such that $E_{\overline{\mathbb{Q}}} \sim E^*_{\overline{\mathbb{Q}}}$. Observe that there always exist N and E^* satisfying these conditions, for example by taking $N = K$ and $E^* = E$; but in Sections 2D and 2E we will exploit certain situations where $N \subsetneq K$ and $E^* \neq E$.

Then we can consider two cohomology classes: the class γ_E^K attached to E, and the class $\gamma_{E^*}^N$ attached to E^*. We recall the following key result about γ_E^K, which is a particular case of [Fité and Guitart 2018a, Corollary 2.4].

Theorem 2.12. *The cohomology class γ_E^K is 2-torsion.*

Denote by U the set of roots of unity of M and put $P = M^\times/U$. The same argument of [Fité and Guitart 2018a, Proof of Theorem 2.14] proves the following decomposition of the 2-torsion of $H^2(\mathrm{Gal}(K/M), M^\times)$:

$$H^2(\mathrm{Gal}(K/M), M^\times)[2] \simeq H^2(\mathrm{Gal}(K/M), U)[2] \times \mathrm{Hom}(\mathrm{Gal}(K/M), P/P^2). \tag{2-6}$$

If $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$ this particularizes to

$$H^2(\mathrm{Gal}(K/M), M^\times)[2] \simeq H^2(\mathrm{Gal}(K/M), \{\pm 1\}) \times \mathrm{Hom}(\mathrm{Gal}(K/M), P/P^2). \tag{2-7}$$

For $\gamma \in H^2(\mathrm{Gal}(K/M), M^\times)[2]$ we will denote by $(\gamma_\pm, \overline{\gamma})$ its components under the isomorphism (2-7); we will refer to $\gamma_\pm$ as the sign component and to $\overline{\gamma}$ as the degree component.

In order to study the relation between γ_E^K and $\gamma_{E^*}^N$, define L/K to be the smallest extension such that E_L^* and E_L are isogenous. Since all the endomorphisms of E are defined over K, this is also the smallest extension L/K such that $\mathrm{Hom}(E_L^*, E_L) = \mathrm{Hom}(E^*_{\overline{\mathbb{Q}}}, E_{\overline{\mathbb{Q}}})$. The extension $L/\mathbb{Q}$ is Galois. Indeed, for $\sigma \in G_\mathbb{Q}$ put $L' = {}^\sigma L$ and let $\beta_\sigma : {}^\sigma E^* \to E^*$ and $\mu_\sigma : {}^\sigma E \to E$ be isogenies defined over N and over K respectively; then, if $\phi : E_L^* \to E_L$ is an isogeny defined over L we find that $\mu_\sigma \circ {}^\sigma\phi \circ \beta_\sigma^{-1}$ is an isogeny defined over L' between $E^*_{L'}$ and $E_{L'}$, so that $L \subseteq L'$ and therefore $L = L'$.

One can also characterize L/K as the minimal extension such that

$$\mathrm{Hom}(E^*_{\overline{\mathbb{Q}}}, A_{\overline{\mathbb{Q}}}) \simeq \mathrm{Hom}(E_L^*, A_L).$$

Denote by

$$\inf_N^K : H^2(\mathrm{Gal}(N/M), M^\times) \to H^2(\mathrm{Gal}(K/M), M^\times)$$

the inflation map in Galois cohomology.

Lemma 2.13. *Suppose that $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$. Then*

$$\inf_N^K(\gamma_{E^*}^N) = w \cdot \gamma_E^K,$$

for some $w \in H^2(\mathrm{Gal}(K/M), \{\pm 1\})$.

Proof. Since $E_L \sim (E_*)_L$ we have that

$$\inf_N^L(\gamma_{E^*}^N) = \inf_K^L(\gamma_E^K). \tag{2-8}$$

Now consider the following piece of the inflation–restriction exact sequence

$$H^1(\mathrm{Gal}(L/K), M^\times) \xrightarrow{t} H^2(\mathrm{Gal}(K/M), M^\times) \xrightarrow{\inf_K^L} H^2(\mathrm{Gal}(L/M), M^\times). \tag{2-9}$$

Equality (2-8) implies that $\inf_N^K(\gamma_{E^*}^N)$ and γ_E^K have the same image under the inflation map $\inf_K^L$, and thus

$$\inf_N^K(\gamma_{E^*}^N) = t(v) \cdot \gamma_E^K$$

for some $v \in H^1(\mathrm{Gal}(L/K), M^\times)$. If $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$ we have that

$$H^1(\mathrm{Gal}(L/K), M^\times) \simeq \mathrm{Hom}(\mathrm{Gal}(L/K), \{\pm 1\})$$

and therefore $t(v)$ belongs to $H^2(\mathrm{Gal}(K/M), \{\pm 1\})$. □

Observe that from Theorem 2.12 one cannot deduce that the class $\gamma_{E^*}^N$ is 2-torsion, since A_N is not isogenous to $(E^*)^2$ in general. By Lemma 2.13, what we do deduce is that $\inf_N^K(\gamma_{E^*}^N)^2 = 1$. Therefore, once again by the inflation–restriction exact sequence

$$H^1(\mathrm{Gal}(K/N), M^\times) \xrightarrow{t} H^2(\mathrm{Gal}(N/M), M^\times) \xrightarrow{\inf_N^K} H^2(\mathrm{Gal}(K/M), M^\times) \tag{2-10}$$

we have that

$$(\gamma_{E^*}^N)^2 = t(\mu) \quad \text{for some } \mu \in H^1(\mathrm{Gal}(K/N), M^\times). \tag{2-11}$$

The following technical lemma will be used in Section 2E below.

Lemma 2.14. *Suppose that N/M is abelian and that $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$. Let $c_{E^*}^N$ be a cocycle representing the class $\gamma_{E^*}^N$. Then $c_{E^*}^N(\sigma, \tau) = \pm c_{E^*}^N(\tau, \sigma)$ for all $\sigma, \tau \in \mathrm{Gal}(N/M)$.*

Proof. Since $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$ we have that

$$H^1(\mathrm{Gal}(K/N), M^\times) = \mathrm{Hom}(\mathrm{Gal}(K/N), \{\pm 1\}). \tag{2-12}$$

By (2-11) and (2-12) we can suppose that there exists a map $d : \mathrm{Gal}(N/M) \to M^\times$ such that

$$c_{E^*}^N(\sigma, \tau)^2 = d(\sigma) d(\tau) d(\sigma\tau)^{-1} \cdot t(\mu)(\sigma, \tau),$$

where $t(\mu)(\sigma,\tau)\in\{\pm 1\}$. Therefore

$$c^N_{E^*}(\sigma,\tau)^2=\pm d(\sigma)d(\tau)d(\sigma\tau)^{-1}=\pm d(\sigma)d(\tau)d(\tau\sigma)^{-1}=\pm c^N_{E^*}(\tau,\sigma)^2.$$

We see that $c^N_{E^*}(\sigma,\tau)/c^N_{E^*}(\tau,\sigma)$ is a root of unity in M, and hence is equal to ± 1. □

2D. ***c-representations from squares of CM elliptic curves.*** Keep the notations from Section 2C. We will denote by V the M-module $\mathrm{Hom}(E^*_L, A_L)$. Fix a system of isogenies $\{\mu_\sigma : {}^\sigma E^*\to E^*\}_{\sigma\in\mathrm{Gal}(L/M)}$. We do not have a natural action of $\mathrm{Gal}(L/M)$ on V, but the next lemma says that we can use the chosen system of isogenies to define a c-action on V.

Lemma 2.15. *The map*

$$\varrho_V:\mathrm{Gal}(L/M)\to \mathrm{GL}(V)$$

defined by

$$\varrho_V(f)={}^\sigma f\circ\mu_\sigma^{-1}\qquad \text{for } \sigma\in\mathrm{Gal}(L/M),\ f\in V$$

and the 2*-cocycle* $c^L_{E^*}$ *endow the module* V *with a structure of a c-representation.*

Proof. This is tautological:

$$\varrho_V(\sigma)\varrho_V(\tau)(f)={}^{\sigma\tau}f\circ{}^\sigma\mu_\tau^{-1}\circ\mu_\sigma^{-1}={}^{\sigma\tau}f\circ\mu_{\sigma\tau}^{-1}\cdot c^L_{E^*}(\sigma,\tau)=\varrho_V(\sigma\tau)(f)c^L_{E^*}(\sigma,\tau).$$ □

Let now R denote the M-module $\mathrm{End}(A_K)$. It is equipped with the natural Galois conjugation action of $\mathrm{Gal}(L/M)$, which factors through $\mathrm{Gal}(K/M)$ and which we sometimes will write as $\varrho_R(\sigma)(\psi)={}^\sigma\psi$. Let T denote $\mathrm{Hom}(V,V)$, equipped with the c-representation structure given by Lemma 2.15 and Proposition 2.2. Note that by Corollary 2.5, we know that T is actually a $M[\mathrm{Gal}(L/M)]$-module.

Lemma 2.16. *The map*

$$\Phi: R\to T\simeq V\otimes V^*,\quad \Phi(\psi)(f)=\psi\circ f,\qquad \text{for } f\in V,\ \psi\in\mathrm{End}(A_K)$$

is an isomorphism of c-representations (*and thus of* $M[\mathrm{Gal}(L/M)]$*-modules*).

Proof. The fact that Φ is a morphism of c-representations is straightforward:

$$\begin{aligned}\varrho_T(\sigma)(\Phi({}^{\sigma^{-1}}\psi))(f)&=\varrho_V(\sigma)(\Phi({}^{\sigma^{-1}}\psi)(\varrho_V(\sigma)^{-1}(f)))\\&=\varrho_V(\sigma)({}^{\sigma^{-1}}\psi\circ\varrho_V(\sigma^{-1})(f)c^L_{E^*}(\sigma^{-1},\sigma)^{-1})\\&=\psi\circ f\circ{}^\sigma\mu^{-1}_{\sigma^{-1}}\mu_\sigma^{-1}c^L_{E^*}(\sigma^{-1},\sigma)^{-1}\\&=\Phi(\psi)(f),\end{aligned}$$

where we have used Remark 2.1 in the second and last equalities. The lemma follows by noting that Φ is clearly injective and that both R and T have dimension 4 over M. □

We now describe the $M[\mathrm{Gal}(K/M)]$-module structure of R. It follows from (2-5) that the order r of an element $\sigma\in\mathrm{Gal}(K/M)$ is 1, 2, 3, 4, or 6.

Lemma 2.17. $\mathrm{Tr}\,\varrho_R(\sigma)=2+\zeta_r+\bar\zeta_r$, *where* ζ_r *is a primitive* r*-th root of unity.*

Remark 2.18. This lemma is proven in [Fité and Sutherland 2014, Proposition 3.4] under the strong running hypothesis of that paper: in our setting that hypothesis would say that there exists E^* defined over M such that $A_{\overline{\mathbb{Q}}} \sim E^{*2}_{\overline{\mathbb{Q}}}$ (i.e., that N can be taken to be M, in the notation of the previous section).

Proof. We claim that $\mathrm{Tr}(\varrho_R) \in M$ is in fact rational. Let us postpone the proof of this claim until the end of the proof of the lemma. Assuming it, we have that

$$\mathrm{Tr}_{M/\mathbb{Q}}(\mathrm{Tr}(\varrho_R(\sigma))) = 2\,\mathrm{Tr}(\varrho_R)(\sigma). \tag{2-13}$$

But if $\varrho_{R_{\mathbb{Q}}}$ is the representation afforded by R regarded as an 8-dimensional module over $\mathbb{Q}$, we have

$$\mathrm{Tr}_{M/\mathbb{Q}}(\mathrm{Tr}(\varrho_R(\sigma))) = \mathrm{Tr}(\varrho_{R_{\mathbb{Q}}})(\sigma) = 2(2+\zeta_r+\bar{\zeta}_r), \tag{2-14}$$

where the last equality is [Fité et al. 2012, Proposition 4.9]. The comparison of (2-13) and (2-14) concludes the proof of the lemma.

We turn now to prove the rationality of $\mathrm{Tr}\,\varrho_R$. We first recall the aforementioned proof (that of [Fité and Sutherland 2014, Proposition 3.4]) which uses the fact that we can choose E^* to be defined over M. In this case, we have that V is an $M[\mathrm{Gal}(L/M)]$-module, that $\mathrm{Tr}(\varrho_{V^*})$ is a sum of roots of unity so that $\mathrm{Tr}(\varrho_{V^*}) = \overline{\mathrm{Tr}(\varrho_V)}$, and hence that $\mathrm{Tr}(\varrho_R) = \mathrm{Tr}(\varrho_V)\cdot\overline{\mathrm{Tr}\,\varrho_V}$ belongs to $\mathbb{Q}$.

For the general case, assume that $\mathrm{Tr}\,\varrho_R$ does not belong to $\mathbb{Q}$. Since it is a sum of roots of unity of orders diving either 4 or 6, then M would be $\mathbb{Q}(i)$ or $\mathbb{Q}(\sqrt{-3})$, but then we could take a model of E^* defined over M, and by the above paragraph, the trace $\mathrm{Tr}\,\varrho_R$ would be rational, which is a contradiction. □

2E. ***Obstructions.*** Keep the notations from Sections 2C and 2D. Let S denote the normal subgroup of $\mathrm{Gal}(K/M)$ generated by the square elements. In this section, we make the following hypotheses.

Hypothesis 2.19. (i) *There exists a Ribet M-curve E^* with CM by M completely defined over N, where N/M is the subextension of K/M fixed by S.*

(ii) $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$.

Let $\sigma \in \mathrm{Gal}(K/M)$ be an element of order $r \in \{4, 6\}$. Let

$$\bar{\cdot} : \mathrm{Gal}(K/M) \to \mathrm{Gal}(N/M) \simeq \mathrm{Gal}(K/M)/S \tag{2-15}$$

denote the natural projection map. Note that $\mathrm{Gal}(N/M)$ is a group of exponent dividing 2.

Theorem 2.20. *Under Hypothesis 2.19, we have*:

(i) *If $r = 4$, then $2c^N_{E^*}(\bar{\sigma}, \bar{\sigma})$ belongs to $\pm(M^\times)^2$.*

(ii) *If $r = 6$, then $3c^N_{E^*}(\bar{\sigma}, \bar{\sigma})$ belongs to $\pm(M^\times)^2$.*

Proof. First of all, note that E^* is completely defined over N. Thus we can, and do, assume that $c^L_{E^*}$ is the inflation of $c^N_{E^*}$. Let $s \in \mathrm{Gal}(L/M)$ be a lift of σ. By Hypothesis 2.19(ii), we have that $[L : K] \leq 2$.

Therefore, the order of s divides $2r$. We then have

$$\varrho_V(s)^{2r} = \varrho_V(s^2)^r c_{E^*}^N(\bar{\sigma},\bar{\sigma})^r = \varrho_V(s^{2r}) c_{E^*}^N(\bar{\sigma},\bar{\sigma})^r = c_{E^*}^N(\bar{\sigma},\bar{\sigma})^r, \tag{2-16}$$

where we have used that $c_{E^*}^N(\bar{\sigma}^{2e},\bar{\sigma}^{2e'}) = 1$ for any pair of integers e, e'. Let α and β be the eigenvalues of $\varrho_V(s)$. By (2-16), we have that $\alpha^{2r} = c_{E^*}^N(\bar{\sigma},\bar{\sigma})^r$, from which we deduce that $\omega_r\alpha^2 = c_{E^*}^N(\bar{\sigma},\bar{\sigma}) \in M^\times$, where ω_r is a (not necessarily primitive) r-th root of unity.

Since the eigenvalues of $\varrho_{V^*}(s)$ are $1/\alpha$ and $1/\beta$, by Lemmas 2.17 and 2.16 we have that

$$2 + \zeta_r + \bar{\zeta}_r = (\alpha+\beta)\Big(\frac{1}{\alpha}+\frac{1}{\beta}\Big); \text{ equivalently, } \alpha^2+\beta^2 = (\zeta_r+\bar{\zeta}_r)\alpha\beta. \tag{2-17}$$

This means that α/β satisfies the r-th cyclotomic polynomial and thus, by reordering α and β if necessary, we have that $\alpha = \beta\zeta_r$.

Combining this with (2-17), we get

$$(2+\zeta_r+\bar{\zeta}_r)c_{E^*}^N(\bar{\sigma},\bar{\sigma}) = (2+\zeta_r+\bar{\zeta}_r)\omega_r\alpha^2 = (2+\zeta_r+\bar{\zeta}_r)\alpha\beta\omega_r\zeta_r = (\alpha+\beta)^2\omega_r\zeta_r.$$

Since the left-hand side is in $M^\times$, the fact that $\alpha+\beta \in M^\times$ tells us that $\omega_r\zeta_r \in M^\times$. If $\omega_r\zeta_r$ is not rational, then $M = \mathbb{Q}(\zeta_r)$, which contradicts Hypothesis 2.19(ii). If $\omega_r\zeta_r \in \mathbb{Q}$, since it is a root of unity, it must be equal to ± 1 and thus we get

$$\pm(2+\zeta_r+\bar{\zeta}_r)c_{E^*}^N(\bar{\sigma},\bar{\sigma}) = (\alpha+\beta)^2.$$

Therefore, $(2+\zeta_r+\bar{\zeta}_r)c_{E^*}^N(\bar{\sigma},\bar{\sigma})$ belongs to $\pm(M^\times)^2$. □

Remark 2.21. It follows from the above proof that if $r = 4$, then any lift $s \in \mathrm{Gal}(L/M)$ of σ has order $2r = 8$. Indeed, if the order of s was r, then arguing as in (2-16), we would obtain $\varrho_V(s)^r = c_{E^*}^N(\bar{\sigma},\bar{\sigma})^{r/2}$, from which we would infer $\omega_{r/2}\alpha^2 = c_{E^*}^N(\bar{\sigma},\bar{\sigma})$, for some (not necessarily primitive) $r/2$-th root of unity. We could then run the same argument as above, but since $\omega_{r/2}\zeta_r$ is never rational, we would deduce now that $M = \mathbb{Q}(i)$. Note that if $r = 6$ it can certainly happen that $\omega_{r/2}\zeta_r \in \mathbb{Q}$.

Until the end of this section, we make the following additional assumption on M.

Hypothesis 2.22. (i) $\mathrm{Gal}(K/M) \simeq \mathrm{D}_4$ *or* D_6.

(ii) $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$.

Hypothesis 2.22(i) implies that N/M is a biquadratic extension. By Proposition 2.10(i), there exists a Ribet M-curve E^* with CM by M completely defined over the Hilbert class field H of M. Using [Fité and Guitart 2018a, Theorem 2.14], it is immediate to see that $H \subseteq N$, so that Hypothesis 2.22 implies Hypothesis 2.19.

The next two propositions describe the structure of the group $\mathrm{Gal}(L/M)$.

Proposition 2.23. *If* $\mathrm{Gal}(K/M) \simeq \mathrm{D}_4$, *then* $\mathrm{Gal}(L/M)$ *is isomorphic to either the dihedral group* D_8; *the generalized dihedral group* QD_8 *of order* 16; *or the generalized quaternion group* Q_{16}.[2]

[2]The gap identification numbers of QD_8 and Q_{16} are $\langle 16, 8\rangle$ and $\langle 16, 9\rangle$, respectively.

Proof. If $\mathrm{Gal}(K/M) \simeq \mathrm{D}_4$, then by Remark 2.21 we have that any element of $\mathrm{Gal}(L/M)$ projecting onto an element of $\mathrm{Gal}(K/M)$ of order 4 must have order 8. The groups of order 16 with a quotient isomorphic to D_4 satisfying the previous property are those in the statement of the proposition. □

Proposition 2.24. *If* $\mathrm{Gal}(K/M) \simeq \mathrm{D}_6$, *there exists a Ribet* M*-curve* E^* *completely defined over* N *with CM by* M *such that* $E \sim E^*_K$ *and hence* $L = K$ *and* $\mathrm{Gal}(L/M) \simeq \mathrm{D}_6$.

Proof. Recall the cohomology class $\gamma_E^K \in H^2(\mathrm{Gal}(K/M), M^\times)[2]$ attached to E and consider the restriction map

$$\mathrm{res} : H^2(\mathrm{Gal}(K/M), M^\times) \to H^2(\mathrm{Gal}(K/N), M^\times).$$

We will first see that $\gamma = \mathrm{res}\gamma_E^K$ is trivial. Recall the decomposition (2-7) of the 2-torsion cohomology classes into degree and sign components

$$H^2(\mathrm{Gal}(K/N), M^\times)[2] \simeq H^2(\mathrm{Gal}(K/N), \{\pm 1\}) \times \mathrm{Hom}(\mathrm{Gal}(K/N), P/P^2),$$

and the notation $\gamma_\pm$ (resp. $\overline{\gamma}$) for the sign component (resp. degree component) of γ. Since $\mathrm{Gal}(K/N) \simeq \mathrm{C}_3$ is the subgroup of $\mathrm{Gal}(K/M)$ generated by the squares, we have that $\overline{\gamma}$ is trivial. Since

$$H^2(\mathrm{Gal}(K/N), \{\pm 1\}) \simeq H^2(\mathrm{C}_3, \{\pm 1\}) = 0,$$

we see that $\gamma_\pm$ is also trivial. By Theorem 2.11, there exists an elliptic curve E^* defined over N such that $E^*_K \sim E$. To see that E^* is completely defined over N, on the one hand, note that since $M \neq \mathbb{Q}(i), \mathbb{Q}(\sqrt{-3})$, then E^* and any Galois conjugate ${}^\sigma E^*$ of it are isogenous over a quadratic extension of N. On the other hand, since $E^*_K \sim E$ and E is completely defined over K, we have that the smallest field of definition of $\mathrm{Hom}(E^*_{\overline{\mathbb{Q}}}, {}^\sigma E^*_{\overline{\mathbb{Q}}})$ is contained in K. Since K/N is a cubic extension, we deduce that E^* and ${}^\sigma E^*$ are in fact isogenous over N. □

Corollary 2.25. *If* $\mathrm{Gal}(K/M) \simeq \mathrm{D}_r$ *for* $r = 4$ *or* 6, *there exists a Ribet* M*-curve* E^* *with CM by* M *completely defined over* N *for which* $\mathrm{Gal}(L/M)$ *contains*

(i) *an element* s *of order* 8 *if* $r = 4$ *and of order* 6 *if* $r = 6$;

(ii) *an element* t *such that* $tst^{-1} = t^a$ *for* $1 \leq a \leq 2r$ *such that* $a \equiv -1 \pmod{r}$.

Proof. This is obvious when $\mathrm{Gal}(L/M)$ is dihedral. For the other options allowed by Proposition 2.23, recall that

$$\mathrm{QD}_8 \simeq \langle s, t \mid s^8, t^2, tsts^5 \rangle, \qquad \mathrm{Q}_{16} \simeq \langle s, t \mid s^8, t^2 s^4, tst^{-1}s \rangle.$$ □

Remark 2.26. It is clear from the proof of Proposition 2.24 that, in the case that $N = H$ and H is not exceptional, we can choose E^* in the above corollary to be a Gross $\mathbb{Q}$-curve.

Until the end of this section, we will assume that E^* is as in the previous corollary. Let s and t be also as in the corollary, and let σ and τ be the images of s and t under the projection map

$$\mathrm{Gal}(L/M) \to \mathrm{Gal}(K/M).$$

Recall also the projection map $\bar{\cdot}: \mathrm{Gal}(K/M) \to \mathrm{Gal}(N/M)$ and note that $\bar{\sigma}$ and $\bar{\tau}$ are nontrivial elements of $\mathrm{Gal}(N/M)$.

Theorem 2.27. *Under Hypothesis 2.22, we have* $c^N_{E^*}(\bar{\tau}, \bar{\tau}) = \pm 1$.

Proof. By Lemma 2.14, we have that $c^N_{E^*}(g, g') = \pm c^N_{E^*}(g', g)$ for every $g, g' \in \mathrm{Gal}(N/M)$. Moreover, the 2-cocycle condition (2-4) asserts that

$$c^N_{E^*}(\bar{\tau}, \bar{\tau}) = c^N_{E^*}(\bar{\tau}, \bar{\tau}) c^N_{E^*}(\bar{\sigma}, 1) = c^N_{E^*}(\bar{\sigma}\bar{\tau}, \bar{\tau}) c^N_{E^*}(\bar{\sigma}, \bar{\tau}).$$

Then, we have

$$\begin{aligned} \varrho_V(t)\varrho_V(s)\varrho_V(t)^{-1} &= \varrho_V(t)\varrho_V(s)\varrho_V(t^{-1}) c^N_{E^*}(\bar{\tau}, \bar{\tau}) = \varrho_V(ts)\varrho_V(t^{-1}) c^N_{E^*}(\bar{\tau}, \bar{\sigma}) c^N_{E^*}(\bar{\tau}, \bar{\tau}) \\ &= \varrho_V(tst^{-1}) c^N_{E^*}(\bar{\tau}\bar{\sigma}, \bar{\tau}) c^N_{E^*}(\bar{\tau}, \bar{\sigma}) c^N_{E^*}(\bar{\tau}, \bar{\tau}) = \pm \varrho_V(s^a) c^N_{E^*}(\bar{\tau}, \bar{\tau})^2. \end{aligned} \tag{2-18}$$

It is easy to observe that

$$\varrho_V(s)^a = \varrho_V(s^a) c^N_{E^*}(\bar{\sigma}, \bar{\sigma})^{(a-1)/2}. \tag{2-19}$$

Letting α and β be the eigenvalues of $\varrho_V(s)$, taking traces of (2-18), and applying (2-19), we obtain

$$(\alpha + \beta) = \pm(\alpha^a + \beta^a) c^N_{E^*}(\bar{\sigma}, \bar{\sigma})^{-(a-1)/2} c^N_{E^*}(\bar{\tau}, \bar{\tau})^2.$$

But as in the proof of Theorem 2.20, we have $\beta = \zeta_r \alpha$ and $c^N_{E^*}(\bar{\sigma}, \bar{\sigma}) = \omega_r \alpha^2$, where ζ_r and ω_r are r-th roots of unity and ζ_r is primitive. This, together with the fact that $a \equiv -1 \pmod{r}$, permits to write the above equation as

$$\pm \frac{1 + \zeta_r}{\omega_r^{-(a-1)/2}(1 + \bar{\zeta}_r)} = c^N_{E^*}(\bar{\tau}, \bar{\tau})^2 \in (M^\times)^2.$$

One easily verifies that $(1 + \zeta_r)/(1 + \bar{\zeta}_r)$ is an r-th root of unity. Therefore, the left-hand side of the above equation is a root of unity in $M^\times$, and hence it must be ± 1. □

3. Restriction of scalars of Gross $\mathbb{Q}$-curves

For the convenience of the reader, in this section we review some results of [Nakamura 2004] on Gross $\mathbb{Q}$-curves, to which we refer for more details and proofs.

Let M be an imaginary quadratic field. Throughout this section, we make the following hypothesis.

Hypothesis 3.1. (i) *M is nonexceptional.*

(ii) *M has class group isomorphic to $\mathrm{C}_2 \times \mathrm{C}_2$.*

Remark 3.2. If M has class group isomorphic to $\mathrm{C}_2 \times \mathrm{C}_2$, then the discriminant D of M belongs to the set

$$\{-84, -120, -132, -168, -195, -228, -280, -312, -340, -372, -408, -435, \\ -483, -520, -532, -555, -595, -627, -708, -715, -760, -795, -1012, -1435\}.$$

This list can be easily obtained from [Watkins 2004], for example. Among them, only -340 is exceptional.

Then, by Proposition 2.10, there exists a Gross $\mathbb{Q}$-curve E with CM by M, which is thus completely defined over the Hilbert class field H of M. The aim of this section is to describe Nakamura's method for computing the endomorphism algebra of the restriction of scalars of a Gross $\mathbb{Q}$-curve, which we will then apply to all Gross $\mathbb{Q}$-curves attached to M satisfying Hypothesis 3.1. Our account of Nakamura's method will be only in the particular case where M has class group $\mathrm{C}_2 \times \mathrm{C}_2$, which is the case of interest to us.

As seen in Section 2B, one can associate a cohomology class $\gamma_E := \gamma_E^H$ in the group $H^2(\mathrm{Gal}(H/\mathbb{Q}), M^\times)$ to E. The set of cohomology classes arising from Gross $\mathbb{Q}$-curves over H has cardinality 8 (see [Nakamura 2004, Proposition 4]), and we regard the set of Gross $\mathbb{Q}$-curves over H as partitioned into 8 equivalence classes according to their cohomology class.

Let $\mathrm{Res}_{H/M}(E)$ denote Weil's restriction of scalars of E. This variety is a priori defined over M, but it can be defined over $\mathbb{Q}$, in the sense that $\mathrm{Res}_{H/M}(E) \simeq (B_E)_M$ for some variety $B_E/\mathbb{Q}$. It turns out that the endomorphism algebra $\mathcal{D}_E = \mathrm{End}(B_E)$ only depends on the cohomology class γ_E [Nakamura 2004, Proposition 6]. Nakamura devised a method for computing $\mathcal{D}_E$ in terms of the Hecke character attached to E, which he applied to compute all the endomorphism algebras arising in this way from Gross $\mathbb{Q}$-curves in the cases where $D = -84$ and $D = -195$. We extend his computation to the remaining 21 nonexceptional discriminants of Remark 3.2.

3A. ***Hecke characters of Gross $\mathbb{Q}$-curves.*** The first step is to compute a set of Hecke characters whose associated elliptic curves represent all the equivalence classes of Gross $\mathbb{Q}$-curves.

Local characters. We begin by defining certain local characters that will be used to describe the Hecke characters. Let $\mathbb{I}_M$ be the group of ideles of M. If $\mathfrak{p}$ is a prime of M, we denote by $U_\mathfrak{p} = \mathcal{O}_{M,\mathfrak{p}}^\times$ the group of local units. Also, for a rational prime p put $U_p = \prod_{\mathfrak{p}|p} U_\mathfrak{p}$.

Suppose that p is odd and inert in M. Then define η_p as the unique character $\eta_p : U_p \to \{\pm 1\}$ such that $\eta_p(-1) = (-1)^{\frac{1}{2}(p-1)}$.

Suppose now that 2 is ramified in M and write $D = 4m$. If m is odd, then

$$U_2/U_2^2 \simeq (\mathbb{Z}/2\mathbb{Z})^3 \simeq \langle \sqrt{m}, 3 - 2\sqrt{m}, 5 \rangle.$$

Define $\eta_{-4} : U_2 \to \{\pm 1\}$ to be the character with kernel $\langle 3 - 2\sqrt{m}, 5\rangle$. If m is even then

$$U_2/U_2^2 \simeq (\mathbb{Z}/2\mathbb{Z})^3 \simeq \langle 1 + \sqrt{m}, -1, 5 \rangle.$$

Define η_8 to be the character with kernel $\langle 1 + \sqrt{m}, -1\rangle$ and η_{-8} the character with kernel $\langle 1 + \sqrt{m}, -5\rangle$.

Hecke characters. Let $U_M = \prod_\mathfrak{p} U_\mathfrak{p}$ be the maximal compact subgroup of $\mathbb{I}_M$. Let S be a finite set of primes of M and put $U_S = \prod_{\mathfrak{p}\in S} U_\mathfrak{p}$. Suppose that $\eta : U_S \to \{\pm 1\}$ is a continuous homomorphism such that $\eta(-1) = -1$. Next, we explain how to construct from η a Hecke character $\phi : \mathbb{I}_M \to \mathbb{C}^\times$ (not uniquely determined) that gives rise, in certain cases, to a Gross $\mathbb{Q}$-curve.

First of all, extend η to a character that we denote by the same name $\eta : U_M \to \{\pm 1\}$ by composing with the projection $U_M \to U_S$. Now this character η can be extended to a character $\tilde{\eta} : U_M M^\times M_\infty^\times \to \mathbb{C}^\times$ by imposing that

$$\tilde{\eta}(M^\times) = 1, \quad \tilde{\eta}(z) = z^{-1} \quad \text{for } z \in M_\infty^\times. \tag{3-1}$$

Let $\phi : \mathbb{I}_M \to \mathbb{C}^\times$ be a Hecke character that extends $\tilde{\eta}$ (there are $[H : M] = 4$ such extensions; see [Shimura 1971, p. 523]). For future reference, it will be useful to have the following formula for ϕ evaluated at certain principal ideals.

Lemma 3.3. *Suppose that (α) is a principal ideal of M such that $v_\mathfrak{p}(\alpha) = 0$ for all $\mathfrak{p} \in S$, and denote by $\alpha_S \in U_S$ the natural image of α in U_S. Then*

$$\phi((\alpha)) = \eta(\alpha_S)\alpha_\infty, \tag{3-2}$$

where α_∞ denotes the image of α in $M_\infty = \mathbb{C}$.

Proof. If we write $(\alpha) = \prod_{\mathfrak{q}\in T} \mathfrak{q}^{v_\mathfrak{q}(\alpha)}$, where T denotes the support of (α), then

$$\phi((\alpha)) = \prod_{\mathfrak{q}\in T} \phi_\mathfrak{q}(\alpha_\mathfrak{q}),$$

where $\phi_\mathfrak{q}$ denotes the restriction of ϕ to $M_\mathfrak{q}$ and $\alpha_\mathfrak{q}$ the image of α in $M_\mathfrak{q}$. Observe that by hypothesis $S \cap T = \varnothing$, and that if $\mathfrak{q} \notin S \cup T$, then $\phi_\mathfrak{q}(\alpha_\mathfrak{q}) = 1$, since $\alpha_\mathfrak{q}$ belongs to $U_\mathfrak{q}$ and $\phi_{|U_\mathfrak{q}} = \tilde{\eta}_{|U_\mathfrak{q}} = 1$. Therefore, we can write

$$\phi((\alpha)) = \prod_{\mathfrak{q}\in T} \phi_\mathfrak{q}(\alpha_\mathfrak{q}) \prod_{\mathfrak{q}\notin T} \phi_\mathfrak{q}(\alpha_\mathfrak{q}) \prod_{\mathfrak{q}\in S} \phi_\mathfrak{q}^{-1}(\alpha_\mathfrak{q}) = \left(\prod_{\mathfrak{q}} \phi_\mathfrak{q}(\alpha_\mathfrak{q}) \right) \eta(\alpha_S),$$

where we have used that η has order 2. Then, by (3-1) we have that

$$\phi((\alpha)) = \left(\phi_\infty(\alpha_\infty) \prod_{\mathfrak{q}} \phi_\mathfrak{q}(\alpha_\mathfrak{q}) \right) \phi_\infty(\alpha_\infty)^{-1} \eta(\alpha_S) = \phi(\alpha)\alpha_\infty \eta(\alpha_S) = \alpha_\infty \eta(\alpha_S). \qquad \square$$

Define now a Hecke character of H by means of $\psi = \phi \circ \mathrm{N}_{H/M}$, where

$$\mathrm{N}_{H/M} : \mathbb{I}_H \to \mathbb{I}_M$$

denotes the norm on ideles. By a result of Shimura [1971, Proposition 9], the Hecke character ψ is attached to a Gross $\mathbb{Q}$-curve if and only if $\bar{\phi} = \phi$, where the bar denotes the action of complex conjugation.

For example, if D has some prime factor $q \equiv 3 \pmod 4$, put $\eta_0 = \eta_q$. If all the odd primes dividing D are congruent to 1 modulo 4, then $D = 8m$ for some odd m and we define η_0 to be η_{-8}. If we denote by $\phi_0 : \mathbb{I}_M \to \mathbb{C}^\times$ a Hecke character attached to η_0 by the above construction, then the Hecke character $\psi_0 = \phi_0 \circ \mathrm{N}_{H/M}$ is the Hecke character attached to a Gross $\mathbb{Q}$-curve over H.

Let W be the set of characters $\theta : U_M \to \{\pm 1\}$ such that $\theta(-1) = 1$ and $\bar{\theta} = \theta$. Denote also by W_0 the set of $\theta \in W$ such that $\theta = \kappa \circ \mathrm{N}_{M/\mathbb{Q}}$ for some Dirichlet character κ. By [Nakamura 2004, Proposition 3], the group W/W_0 is generated by two characters that can be described explicitly in terms of the characters η_p, η_{-4}, η_{-8}, and η_8. More precisely:

(1) If $D = -pqr$ with p, q, and r primes congruent to 3 modulo 4, then $W/W_0 = \langle \eta_p\eta_q, \eta_p\eta_r \rangle$.

(2) If $D = -pqr$ with p and q primes congruent to 1 modulo 4, and r congruent to 3 modulo 4, then $W/W_0 = \langle \eta_p, \eta_q \rangle$.

(3) If $D=-4pq$ with p and q congruent to 3 modulo 4, then $W/W_0=\langle \eta_{-4}, \eta_p\eta_q\rangle$.

(4) If $D=-8pq$ with p and q congruent to 3 modulo 4, then $W/W_0=\langle \eta_{-8}\eta_p, \eta_{-8}\eta_q\rangle$.

(5) If $D=-8pq$ with p congruent to 1 modulo 4 and q congruent to 3 modulo 4, then $W/W_0=\langle \eta_8, \eta_p\rangle$.

(6) If $D=-8pq$ with p and q congruent to 1 modulo 4, then $W/W_0=\langle \eta_p, \eta_q\rangle$.

Denote by $\widetilde{\omega}_1, \widetilde{\omega}_2$ the generators of W/W_0, and define $\omega_i=\widetilde{\omega}_i\circ \mathrm{N}_{H/M}$.

Now let k/H be a quadratic extension such that $k/\mathbb{Q}$ is Galois and k/M is nonabelian. Such quadratic extensions exist by [Nakamura 2004, Theorem 1]. Denote by $\chi : \mathbb{I}_H \to \{\pm 1\}$ the Hecke character attached to k/H.

By [Nakamura 2004, Theorem 2], the eight equivalence classes of $\mathbb{Q}$-curves over H are represented by the Hecke characters $\psi_0\cdot\omega$ with $\omega\in\langle\omega_1,\omega_2,\chi\rangle$. Observe that, in particular, this set of Hecke characters does not depend on the choice of k (any k which is Galois over $\mathbb{Q}$ and nonabelian over M will produce the same set of Hecke characters).

3B. ***Method for computing the endomorphism algebra.*** Let $\mathfrak{p}_1$ and $\mathfrak{p}_2$ be prime ideals of M that generate the class group and that are coprime to the conductors of ψ_0, ω_1, ω_2, and χ. Let L_i be the decomposition field of $\mathfrak{p}_i$ in H, and F_i the maximal totally real subfield of L_i.

Suppose that E is a Gross $\mathbb{Q}$-curve over H with Hecke character of the form $\psi=\psi_0\omega_1^a\omega_2^b$ for some $a,b\in\{0,1\}$. We can write $\psi=\phi\circ\mathrm{N}_{H/M}$, where $\phi=\phi_0\widetilde{\omega}_1^a\widetilde{\omega}_2^b$. Then $\phi(\mathfrak{p}_i)+\phi(\bar{\mathfrak{p}}_i)$ generates a quadratic number field $\mathbb{Q}(\sqrt{n_i})$, and the endomorphism algebra $\mathcal{D}_E=\mathrm{End}(B_E)$ is isomorphic to the biquadratic field $\mathbb{Q}(\sqrt{n_1},\sqrt{n_2})$; see [Nakamura 2004, Proposition 7, Theorem 3].

Remark 3.4. Observe that $\phi(\mathfrak{p}_i)+\phi(\bar{\mathfrak{p}}_i)$ can be computed if one knows the two quantities $\phi(\mathfrak{p}_i^2)$ and $\phi(\mathfrak{p}_i\bar{\mathfrak{p}}_i)$. Since $\mathfrak{p}_i^2$ and $\mathfrak{p}_i\bar{\mathfrak{p}}_i$ are principal, one can compute $\phi(\mathfrak{p}_i^2)$ and $\phi(\mathfrak{p}_i\bar{\mathfrak{p}}_i)$ by means of (3-2).

Suppose now that the Hecke character of E is of the form $\psi=\psi_0\chi\omega_1^a\omega_2^b$. Then $\mathcal{D}_E$ is a quaternion algebra over $\mathbb{Q}$, say

$$\mathcal{D}_E\simeq\left(\frac{t_1,t_2}{\mathbb{Q}}\right).$$

The t_i can be computed as follows; see [Nakamura 2004, Proposition 7]. First of all, let n_1 and n_2 be the rational numbers defined as in the previous paragraph for the character $\psi/\chi=\psi_0\omega_1^a\omega_2^b$.

(1) Suppose that $\mathrm{Gal}(k/L_i)\simeq \mathrm{C}_2\times\mathrm{C}_2$. Then:

  (a) If k/F_i is abelian then $t_i=n_i$.

  (a) If k/F_i is nonabelian, then $t_i=D/n_i$.

(2) Suppose that $\mathrm{Gal}(k/L_i)\simeq \mathrm{C}_4$. Then:

  (a) If k/F_i is abelian, then $t_i=-n_i$.

  (b) If k/F_i is nonabelian, then $t_i=-D/n_i$.

3C. ***Computations and tables.*** For each of the 23 nonexceptional imaginary quadratic fields of class group $\mathrm{C}_2 \times \mathrm{C}_2$, we have computed the 8 endomorphism algebras arising from restriction of scalars of Gross $\mathbb{Q}$-curves. The results are displayed in Table 1. The notation is as follows: for the biquadratic fields, the notation (a, b) indicates the field $\mathbb{Q}(\sqrt{a}, \sqrt{b})$; for the quaternion algebras, we write the discriminant of the algebra.

For a Gross $\mathbb{Q}$-curve E, recall that B_E denotes the abelian variety over $\mathbb{Q}$ such that $\mathrm{Res}_{H/M} E \sim (B_E)_M$. Since B_E is isogenous to its quadratic twist over M, this implies that

$$\mathrm{Res}_{H/\mathbb{Q}} E \sim (B_E)^2.$$

We observe in Table 1 that for all discriminants except -195, -312, -555, -715, and -760, at least one of the quaternion algebras is the split algebra $\mathrm{M}_2(\mathbb{Q})$ of discriminant 1. This implies that for the corresponding Gross $\mathbb{Q}$-curve E the variety B_E decomposes as

$$B_E \sim A^2,$$

with $A/\mathbb{Q}$ an abelian surface. Therefore, $\mathrm{Res}_{H/\mathbb{Q}} E$ decomposes as the fourth power of an abelian surface.

On the other hand, for the discriminants -195, -312, -555, -715, and -760 we see that B_E is always simple: its endomorphism algebra is either a biquadratic field or a quaternion division algebra over $\mathbb{Q}$. Therefore, $\mathrm{Res}_{H/\mathbb{Q}} E \sim W^2$ for some simple variety W of dimension 4. We record these findings in the following statement.

Theorem 3.5. *Let M be an imaginary quadratic field of discriminant D and Hilbert class field H. Suppose that D is nonexceptional and that* $\mathrm{Gal}(H/M) \simeq \mathrm{C}_2 \times \mathrm{C}_2$. *If $D \neq -195, -312, -555, -715, -760$, there exists a Gross $\mathbb{Q}$-curve E/H such that*

$$\mathrm{Res}_{H/\mathbb{Q}} E \sim A^4, \quad \text{for some simple abelian surface } A/\mathbb{Q}.$$

If $D = -195, -312, -555, -715, -760$, then for every Gross $\mathbb{Q}$-curve E/H we have that

$$\mathrm{Res}_{H/\mathbb{Q}} E \sim W^2, \quad \text{for some simple abelian variety } W/\mathbb{Q} \text{ of dimension } 4.$$

Remark 3.6. As mentioned above, the cases of $D = -84$ and $D = -195$ were already computed by Nakamura [2004, §6]. We note what appears to be a typo in Nakamura's table in page 647: the last biquadratic field should be $\mathbb{Q}(\sqrt{-14}, \sqrt{42})$, instead of $\mathbb{Q}(\sqrt{-14}, \sqrt{-42})$.

We have used the software [Sage] and [Magma] to perform the computations of Table 1. The interested reader can find the code we used in [Fité and Guitart 2018b].

4. Proof of the main theorems

We begin with a lemma that will be used in the proof of Theorem 1.2.

Lemma 4.1. *Let E be a Gross $\mathbb{Q}$-curve with CM by a field M of discriminant D, and suppose that* $\mathrm{Gal}(H/M)$ *is isomorphic to* $\mathrm{C}_2 \times \mathrm{C}_2$. *Denote by γ_E^H the class in $H^2(\mathrm{Gal}(H/M), M^\times)$ attached to E,*

and by c_E a cocycle representing γ_E^H. If $\sigma \in \mathrm{Gal}(H/M)$ is nontrivial, then $\pm d \cdot c_E(\sigma,\sigma) \in (M^\times)^2$ for some divisor d of D such that d is not a square in $M^\times$.

Proof. Let $\mathcal{O}_M$ denote the ring of integers of M. Denote by p_1, p_2, p_3 the primes dividing D. Observe that $p_i\mathcal{O}_M = \mathfrak{p}_i^2$, with $\mathfrak{p}_i$ a nonprincipal prime ideal of $\mathcal{O}_M$. Clearly, we can always find p_i, p_j such that $\pm p_i p_j$ is not a square in $M^\times$, and therefore $\mathfrak{p}_i\mathfrak{p}_j$ is not principal. Thus $\mathfrak{p}_i$, $\mathfrak{p}_j$ generate the class group. Therefore, we can assume that any nontrivial element of $\mathrm{Gal}(H/K)$ is of the form $\sigma_\mathfrak{q}$ for some unramified prime $\mathfrak{q}$ which is equivalent to either $\mathfrak{p}_i$, $\mathfrak{p}_j$ or $\mathfrak{p}_i\cdot\mathfrak{p}_j$. Here $\sigma_\mathfrak{q}$ stands for the Frobenius automorphism of H/K at $\mathfrak{q}$.

Now we argue (and use the same notation) as in [Nakamura 2004, Proof of Theorem 3]. Namely, denote by $u(\mathfrak{q})$ the $\mathfrak{q}$-multiplication isogenies

$$u(\mathfrak{q}) : {}^{\sigma_\mathfrak{q}}E \to E,$$

and denote by c the 2-cocycle associated to E using the system of isogenies $u(\mathfrak{q})$ (together with the identity isogeny for $1 \in \mathrm{Gal}(H/M)$). Note that c_E is any cocycle representing γ_E^H, and it may be different from c. But in any case they are cohomologous, which in particular implies that

$$c(\sigma_\mathfrak{q},\sigma_\mathfrak{q}) = b_\mathfrak{q}^2 \cdot c_E(\sigma_\mathfrak{q},\sigma_\mathfrak{q}) \quad \text{for some } b_\mathfrak{q} \in M^\times. \tag{4-1}$$

From [loc. cit., Equation (6) and the following display], since the order n of $\sigma_\mathfrak{q}$ is 2 in our case, we see that

$$c(\sigma_\mathfrak{q},\sigma_\mathfrak{q})\mathcal{O}_M = \mathfrak{q}^2.$$

The proof is finished by observing that $\mathfrak{q}^2 = \alpha\mathcal{O}_M$, where $\alpha \in M^\times$ is, up to an element of $(M^\times)^2$, equal to $\pm p_i$, $\pm p_j$, or $\pm p_i \cdot p_j$. □

Proof of Theorem 1.2. For all the quadratic imaginary fields not listed in (1-2), we have constructed in the first part of Theorem 3.5 abelian surfaces defined over $\mathbb{Q}$ satisfying the hypothesis of the theorem. To rule out the remaining 6 fields, we proceed in the following way.

Let M be one of the fields in the list (1-2) and suppose that an abelian surface A satisfying the hypothesis of the theorem exists for M. Resume the notations from Section 2D. As $\mathrm{Gal}(H/M) \simeq \mathrm{C}_2 \times \mathrm{C}_2$ and $H \subseteq K$ (by [Fité and Guitart 2018a, Theorem 2.14]), the only possibilities for $\mathrm{Gal}(K/M)$ are $\mathrm{C}_2 \times \mathrm{C}_2$, D_4, and D_6.

Suppose that $\mathrm{Gal}(K/M)$ is $\mathrm{C}_2 \times \mathrm{C}_2$. Then $K = H$ and thus E is a Gross $\mathbb{Q}$-curve. By Proposition 2.10, we have that M is not exceptional and thus we cannot have $M = \mathbb{Q}(\sqrt{-340})$. For the other possibilities for M, we have seen in the second part of Theorem 3.5 that $\mathrm{Res}_{H/\mathbb{Q}} E$ does not have any simple factor of dimension 2, but this is a contradiction with the fact that A should be a factor of $\mathrm{Res}_{H/\mathbb{Q}} E$ (indeed, the universal property of Weil's restriction of scalars implies that $\mathrm{Hom}(A, \mathrm{Res}_{H/\mathbb{Q}} E) = \mathrm{Hom}(A_H, E) \simeq M^2$, and thus $\mathrm{Hom}(A, \mathrm{Res}_{H/\mathbb{Q}} E) \neq 0$).

Suppose that $\mathrm{Gal}(K/M)$ is D_4 or D_6. Resume the notations of Section 2E. Let E^* be a Ribet M-curve completely defined over H with CM by M which we chose as in Corollary 2.25 (and which exists because of Proposition 2.10). Note that Hypothesis 2.22 is satisfied. Then, by Theorem 2.27, there is a nontrivial element $\bar{\tau} \in \mathrm{Gal}(N/M) = \mathrm{Gal}(H/N)$ such that

$$c_{E^*}^H(\bar{\tau},\bar{\tau}) = \pm 1. \tag{4-2}$$

D	Biquadratic fields	Quaternion algebras
-84	$(-14,-2),(-6,2),(-6,-42),(-14,42)$	2, 1, 2, 1
-120	$(-5,10),(5,-10),(-5,-10),(5,10)$	1, 6, 3, 1
-132	$(22,-2),(-6,-2),(6,-66),(-22,-66)$	1, 2, 1, 2
-168	$(-14,-2),(3,-21),(14,21),(-3,2)$	2, 1, 1, 1
-195	$(13,-5),(-13,-5),(-13,5),(13,5)$	13, 39, 26, 39
-228	$(-38,-2),(6,-2),(-6,-114),(38,-114)$	2, 1, 2, 1
-280	$(-10,-5),(-10,5),(10,-5),(10,5)$	2, 1, 14, 14
-312	$(13,-26),(-13,26),(-13,-26),(13,26)$	13, 39, 26, 39
-372	$(-62,31),(-6,-3),(-6,31),(-62,-3)$	2, 1, 2, 1
-408	$(-17,34),(-17,-34),(17,-34),(17,34)$	2, 1, 1, 1
-435	$(-29,-5),(-29,5),(29,-5),(29,5)$	2, 1, 1, 1
-483	$(-23,7),(23,-69),(-21,-7),(21,69)$	2, 1, 1, 1
-520	$(-13,-5),(13,-5),(-13,5),(13,5)$	1, 1, 1, 2
-532	$(-38,-19),(-14,7),(-14,-19),(-38,7)$	1, 2, 1, 2
-555	$(37,-5),(-37,-5),(-37,5),(37,5)$	37, 111, 74, 111
-595	$(-17,85),(17,-85),(-17,-85),(17,85)$	7, 1, 1, 14
-627	$(19,-11),(-19,-57),(-33,11),(33,57)$	1, 2, 1, 1
-708	$(118,-59),(-6,3),(6,-59),(-118,3)$	1, 2, 1, 2
-715	$(-13,-65),(13,-65),(-13,65),(13,65)$	5, 10, 55, 55
-760	$(-10,5),(10,-5),(-10,-5),(10,5)$	5, 95, 10, 95
-795	$(-53,-5),(53,-5),(-53,5),(53,5)$	6, 1, 1, 3
-1012	$(-46,23),(-22,-11),(-22,23),(-46,-11)$	2, 1, 2, 1
-1435	$(-41,205),(-41,-205),(41,-205),(41,205)$	2, 1, 1, 1

Table 1. Endomorphism algebras of the restriction of scalars of Gross $\mathbb{Q}$-curves. For the biquadratic fields, the notation (a,b) indicates the field $\mathbb{Q}(\sqrt{a},\sqrt{b})$; for the quaternion algebras, we write the discriminant of the algebra

If M is nonexceptional, as noted in Remark 2.26, we can suppose that E^* is in fact a Gross $\mathbb{Q}$-curve. Then (4-2) is a contradiction with Lemma 4.1.

It remains to show that (4-2) also brings a contradiction if $M = \mathbb{Q}(\sqrt{-340})$ is the exceptional field. Put $T = H^{\langle\bar{\tau}\rangle}$, the fixed field by $\bar{\tau}$. Observe that $M \subsetneq T \subsetneq H$. If $c^H_{E^*}(\bar{\tau}, \bar{\tau}) = 1$ then by Theorem 2.11 the curve E^* is isogenous to a curve defined over T, and this is a contradiction with the fact that $M(j_{E^*}) = H$.

Suppose now that $c^H_{E^*}(\bar{\tau}, \bar{\tau}) = -1$. We will see that we can apply the above argument to an appropriate quadratic twist of E^*.

Claim 4.2. *There exists a quadratic extension S/H such that S/M is Galois with* $\mathrm{Gal}(S/M) \simeq \mathrm{D}_4$ *and such that $\bar{\tau}$ lifts to an element of order* 4 *of* $\mathrm{Gal}(S/M)$.

We now show how this claim allows us to produce the appropriate twisted curve (and we will prove the claim later on). Define C to be the S/H quadratic twist of E^*. By [Fité and Guitart 2018a, Lemma 3.13], the curve C is an M-curve completely defined over H and the cohomology classes of E^* and C are related by

$$\gamma^H_C = \gamma^H_{E^*} \cdot \gamma_S,$$

where $\gamma_S \in H^2(\mathrm{Gal}(H/M), \{\pm 1\})$ is the cohomology class attached to the exact sequence

$$1 \to \mathrm{Gal}(S/H) \simeq \{\pm 1\} \to \mathrm{Gal}(S/M) \simeq \mathrm{D}_4 \to \mathrm{Gal}(H/M) \to 1. \tag{4-3}$$

If we identify $\mathrm{Gal}(S/M) \simeq \langle s, t | s^4, t^2, stst\rangle$, then $\mathrm{Gal}(S/H)$ can be identified with the subgroup generated by s^2 and we can assume that $\bar{\tau}$ lifts to s. Let c_S be a cocycle representing γ_S. The usual construction that associates a cohomology class to (4-3) gives that $c_S(\bar{\tau}, \bar{\tau}) = s \cdot s$. Since s^2 is the nontrivial element of $\mathrm{Gal}(S/H)$, it corresponds to -1 under the isomorphism $\mathrm{Gal}(S/H) \simeq \{\pm 1\}$, so that $c_S(\bar{\tau}, \bar{\tau}) = -1$.

We conclude that $c^H_C(\bar{\tau}, \bar{\tau}) = c^H_{E^*}(\bar{\tau}, \bar{\tau}) c_S(\bar{\tau}, \bar{\tau}) = 1$, and as before this implies that C can be defined over T, which is a contradiction.

Proof of Claim 4.2. The Hilbert class field of M is $H = \mathbb{Q}(i, \sqrt{5}, \sqrt{17})$. If we write $H = M(\sqrt{a}, \sqrt{b})$ and suppose that $\bar{\tau}(\sqrt{b}) = \sqrt{b}$, it is well known (see, e.g., [Ledet 2001, §0.4]) that the obstruction to the existence of S is given by the quaternion algebra

$$\left(\frac{a, ab}{M}\right)$$

being nonsplit. There are 3 possibilities for T, namely $T = M(\sqrt{5})$, $T = M(\sqrt{17})$, or $T = M(\sqrt{5 \cdot 17})$, each one giving a different obstruction. The resulting quaternion algebras giving the obstruction are

$$\left(\frac{17 \cdot 5, 5}{M}\right), \left(\frac{17 \cdot 5, 17}{M}\right), \left(\frac{17, 5}{M}\right).$$

Since they are all the split, the field S does exist in all three cases. □

Remark 4.3. As a byproduct of the above proof, we see that there do not exist abelian surfaces over $\mathbb{Q}$ such that $\mathrm{End}(A_{\overline{\mathbb{Q}}}) \simeq \mathrm{M}_2(M)$ with M a quadratic imaginary field with class group $\mathrm{C}_2 \times \mathrm{C}_2$ and $\mathrm{Gal}(K/M) \simeq \mathrm{D}_4$ or D_6. As shown by the table of [Cardona Juanals 2001, p. 112], there do exist abelian surfaces over $\mathbb{Q}$ such that $\mathrm{End}(A_{\overline{\mathbb{Q}}}) \simeq \mathrm{M}_2(M)$ with M a quadratic imaginary field with class group C_2 and $\mathrm{Gal}(K/M) \simeq \mathrm{D}_4$ (resp. D_6). If M is not exceptional, Theorem 2.20 and Lemma 4.1 imply that 2 (resp. 3) divide the discriminant of M is a necessary condition for the existence of such an A. The examples of the table of [Cardona Juanals 2001, p. 112] show that this is actually a necessary and sufficient condition.

Proof of Corollary 1.3. Suppose that A is an abelian surface defined over $\mathbb{Q}$ such that $A_{\overline{\mathbb{Q}}} \sim E \times E'$, where E and E' are elliptic curves defined over $\overline{\mathbb{Q}}$. If E and E' are not isogenous, then $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ is

$$\mathbb{Q} \times \mathbb{Q}, \quad M \times \mathbb{Q} \quad \text{or} \quad M_1 \times M_2,$$

where M, $M_1 \not\simeq M_2$ are quadratic imaginary fields, depending on whether none of E and E' has CM, only one of E and E' has CM, or both of E and E' have CM. In any case, note that by [Fité et al. 2012, Proposition 4.5], both E and E' can be defined over $\mathbb{Q}$, whereby the class number of M, M_1, and M_2 must be 1. Recalling that there are 9 quadratic imaginary fields of class number 1, this accounts for 46 distinct $\overline{\mathbb{Q}}$-endomorphism algebras.

If E and E' are isogenous, we have that $\mathrm{End}(A_{\overline{\mathbb{Q}}})$ is $\mathrm{M}_2(M)$ or $\mathrm{M}_2(\mathbb{Q})$, where M is a quadratic imaginary field, depending on whether E has CM or not. Assume that we are in the former case. By Theorem 1.1, we have that M has class group 1, C_2, or $\mathrm{C}_2 \times \mathrm{C}_2$. As explained in [Fité and Guitart 2018a, Remark 2.20], for all fields M with class group 1 (resp. C_2), abelian surfaces A over $\mathbb{Q}$ with $\mathrm{End}(A_{\overline{\mathbb{Q}}}) \simeq \mathrm{M}_2(M)$ can be easily found. Indeed, let E be an elliptic curve with CM by the maximal order of M and defined over $\mathbb{Q}$ (resp. $\mathbb{Q}(j_E)$). Then consider the square (resp. the restriction of scalars from $\mathbb{Q}(j_E)$ down to $\mathbb{Q}$) of E. If M has class group $\mathrm{C}_2 \times \mathrm{C}_2$, invoke Theorem 1.2 to obtain 18 possibilities for M. Taking into account that there are 18 quadratic imaginary fields of class group C_2 (see [Watkins 2004] for example), we obtain 46 possibilities for the endomorphism algebra of a geometrically split abelian surface over $\mathbb{Q}$ with $\overline{\mathbb{Q}}$-isogenous factors.

An open problem. We wish to conclude the article with an open question.

Question 4.4. *Which is the subset of $\mathcal{A}$ made of the $\overline{\mathbb{Q}}$-endomorphism algebras* $\mathrm{End}(\mathrm{Jac}(C)_{\overline{\mathbb{Q}}})$ *of geometrically split Jacobians of genus* 2 *curves C defined over $\mathbb{Q}$?*

Again the most intriguing case is to determine how many of the 45 possibilities for $\mathrm{M}_2(M)$, with M a quadratic imaginary field, allowed by Theorem 1.2 for geometrically split abelian surfaces defined over $\mathbb{Q}$ still occur among geometrically split Jacobians of genus 2 curves C defined over $\mathbb{Q}$. Looking at the more restrictive setting that requires $\mathrm{Jac}(C)$ to be *isomorphic* to the square of an elliptic curve with CM by the *maximal order* of M, Gélin, Howe, and Ritzenthaler [Gélin et al. 2019] have shown that there are 13 possibilities for such an M (all with class number ≤ 2).

Acknowledgements

Fité is thankful to the organizers of the workshop "Arithmetic Aspects of Explicit Moduli Problems" held at BIRS (Banff) in May 2017, where he explained Theorem 1.1 and raised the question on the existence of an abelian surface over $\mathbb{Q}$ with $\mathrm{End}(A_{\overline{\mathbb{Q}}}) \simeq \mathrm{M}_2(M)$ for an M with class group $\mathrm{C}_2 \times \mathrm{C}_2$. We thank Andrew Sutherland and John Voight for providing a positive answer to this question by pointing out the existence of an abelian surface (actually the Jacobian of a genus 2 curve) with the desired property for the field $M = \mathbb{Q}(\sqrt{-132})$. We also thank Noam Elkies for providing three additional genus 2 curves over $\mathbb{Q}$, these covering the fields $M = \mathbb{Q}(\sqrt{-408})$, $\mathbb{Q}(\sqrt{-435})$, and $\mathbb{Q}(\sqrt{-708})$. These four examples motivated the present paper.

Fité was funded by the Excellence Program María de Maeztu MDM-2014-0445. Fité was partially supported by MTM2015-63829-P. Guitart was funded by projects MTM2015-66716-P and MTM2015-63829-P. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 682152).

References

[Bruin et al. 2006] N. Bruin, E. V. Flynn, J. González, and V. Rotger, "On finiteness conjectures for endomorphism algebras of abelian surfaces", *Math. Proc. Cambridge Philos. Soc.* **141**:3 (2006), 383–408. MR Zbl

[Cardona Juanals 2001] G. Cardona Juanals, *Models racionals de corbes de gènere 2*, Ph.D. thesis, Universitat Politècnica de Catalunya, 2001, available at https://tinyurl.com/cardonaju.

[Fité and Guitart 2018a] F. Fité and X. Guitart, "Fields of definition of elliptic k-curves and the realizability of all genus 2 Sato–Tate groups over a number field", *Trans. Amer. Math. Soc.* **370**:7 (2018), 4623–4659. MR Zbl

[Fité and Guitart 2018b] F. Fité and X. Guitart, "Restriction of scalars of Q curves", 2018, available at https://github.com/xguitart/restriction_of_scalars_of_Q_curves. Sage and Magma code.

[Fité and Guitart 2019] F. Fité and X. Guitart, "Tate module tensor decompositions and the Sato–Tate conjecture for certain abelian varieties potentially of GL_2-type", preprint, 2019. arXiv

[Fité and Sutherland 2014] F. Fité and A. V. Sutherland, "Sato–Tate distributions of twists of $y^2 = x^5 - x$ and $y^2 = x^6 + 1$", *Algebra Number Theory* **8**:3 (2014), 543–585. MR Zbl

[Fité et al. 2012] F. Fité, K. S. Kedlaya, V. Rotger, and A. V. Sutherland, "Sato–Tate distributions and Galois endomorphism modules in genus 2", *Compos. Math.* **148**:5 (2012), 1390–1442. MR Zbl

[Gélin et al. 2019] A. Gélin, E. W. Howe, and C. Ritzenthaler, "Principally polarized squares of elliptic curves with field of moduli equal to $\mathbb{Q}$", pp. 257–274 in *Proceedings of the Thirteenth Algorithmic Number Theory Symposium* (Madison, WI, 2018), edited by R. Scheidler and J. Sorenson, Open Book Ser. **2**, MSP, Berkeley, 2019. MR

[González 2011] J. González, "Finiteness of endomorphism algebras of CM modular abelian varieties", *Rev. Mat. Iberoam.* **27**:3 (2011), 733–750. MR Zbl

[Gross 1980] B. H. Gross, *Arithmetic on elliptic curves with complex multiplication*, Lecture Notes in Math. **776**, Springer, 1980. MR Zbl

[Kani 2011] E. Kani, "Products of CM elliptic curves", *Collect. Math.* **62**:3 (2011), 297–339. MR Zbl

[Ledet 2001] A. Ledet, "Embedding problems and equivalence of quadratic forms", *Math. Scand.* **88**:2 (2001), 279–302. MR Zbl

[Magma] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system, I: The user language", *J. Symbolic Comput.* **24**:3-4, 235–265. MR Zbl

[Murabayashi and Umegaki 2001] N. Murabayashi and A. Umegaki, "Determination of all $\mathbb{Q}$-rational CM-points in the moduli space of principally polarized abelian surfaces", *J. Algebra* **235**:1 (2001), 267–274. MR Zbl

[Nakamura 2001] T. Nakamura, "On abelian varieties associated with elliptic curves with complex multiplication", *Acta Arith.* **97**:4 (2001), 379–385. MR Zbl

[Nakamura 2004] T. Nakamura, "A classification of $\mathbb{Q}$-curves with complex multiplication", *J. Math. Soc. Japan* **56**:2 (2004), 635–648. MR Zbl

[Orr and Skorobogatov 2018] M. Orr and A. N. Skorobogatov, "Finiteness theorems for K3 surfaces and abelian varieties of CM type", *Compos. Math.* **154**:8 (2018), 1571–1592. MR Zbl

[Quer 2000] J. Quer, "$\mathbb{Q}$-curves and abelian varieties of GL_2-type", *Proc. Lond. Math. Soc.* (3) **81**:2 (2000), 285–317. MR Zbl

[Ribet 1992] K. A. Ribet, "Abelian varieties over $\mathbb{Q}$ and modular forms", pp. 53–79 in *Algebra and topology* (Taejŏn, South Korea, 1992), edited by S. G. Hahn and D. Y. Suh, Korea Adv. Inst. Sci. Tech., Taejŏn, South Korea, 1992. MR Zbl

[Sage] W. A. Stein et al., "Sage mathematics software", available at http://www.sagemath.org. Version 6.3.

[Schütt 2007] M. Schütt, "Fields of definition of singular $K3$ surfaces", *Commun. Number Theory Phys.* **1**:2 (2007), 307–321. MR Zbl

[Shafarevich 1996] I. R. Shafarevich, "On the arithmetic of singular $K3$-surfaces", pp. 103–108 in *Algebra and analysis* (Kazan, Russia, 1994), edited by M. M. Arslanov et al., de Gruyter, Berlin, 1996. MR Zbl

[Shimura 1971] G. Shimura, "On the zeta-function of an abelian variety with complex multiplication", *Ann. of Math.* (2) **94** (1971), 504–533. MR Zbl

[Silverman 1994] J. H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Grad. Texts in Math. **151**, Springer, 1994. MR Zbl

[Watkins 2004] M. Watkins, "Class numbers of imaginary quadratic fields", *Math. Comp.* **73**:246 (2004), 907–938. MR Zbl

Communicated by Michael Rapoport
Received 2019-02-08 Revised 2019-10-31 Accepted 2020-02-26

francesc.fite@gmail.com *Massachusetts Institute of Technology, Cambridge, MA, United States*

xevi.guitart@gmail.com *Universitat de Barcelona, Barcelona, Catalonia, Spain*

msp

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1423

Uniform Yomdin–Gromov parametrizations and points of bounded height in valued fields

Raf Cluckers, Arthur Forey and François Loeser

We prove a uniform version of non-Archimedean Yomdin–Gromov parametrizations in a definable context with algebraic Skolem functions in the residue field. The parametrization result allows us to bound the number of $\mathbb{F}_q[t]$-points of bounded degrees of algebraic varieties, uniformly in the cardinality q of the finite field $\mathbb{F}_q$ and the degree, generalizing work by Sedunova for fixed q. We also deduce a uniform non-Archimedean Pila–Wilkie theorem, generalizing work by Cluckers–Comte–Loeser.

1. Introduction

Since the pioneering work [Bombieri and Pila 1989], the determinant method of Bombieri and Pila has been used in various contexts to count integer and rational points of bounded height in algebraic or analytic varieties. Parametrization results, as initiated by Yomdin and Gromov, play a prominent role in some of the most fruitful applications of this method, such as the Pila and Wilkie counting theorem [2006] for definable sets in o-minimal structures. In the non-Archimedean setting, Cluckers, Comte and Loeser prove in [Cluckers et al. 2015] an analog of the Pila–Wilkie counting theorem, but for subanalytic sets in $\mathbb{Q}_p$, the field of p-adic numbers. Their proof relies also on a Yomdin–Gromov type parametrization result. The aim of this paper is to extend their result to obtain bounds uniform in p for some counting points of bounded height problems, over $\mathbb{Q}_p$ and over $\mathbb{F}_p((t))$. Before discussing our parametrization result, we start by presenting the applications to point counting.

1.1. ***Point counting in function fields.*** For q a prime power, consider the finite field with q elements $\mathbb{F}_q$ and for each positive integer n, let $\mathbb{F}_q[t]_n$ be the set of polynomials with coefficients in $\mathbb{F}_q$ and degree (strictly) less than n. Cilleruelo and Shparlinski [2013] have raised the question of bounding the number of $\mathbb{F}_q[t]_n$-points in plane curves. That question was settled by Sedunova [2017]. A particular case of our main theorem is a uniform version of her results. We refer to Theorem 4.1.1 for a more general statement, namely for X of arbitrary dimension. For an affine variety X defined over a subring of $\mathbb{F}_q((t))$, write $X(\mathbb{F}_q[t])_n$ for the subset of $X(\mathbb{F}_q((t)))$ consisting of points whose coordinates lie in $\mathbb{F}_q[t]_n$.

Theorem A. *Fix an integer $\delta > 0$. Then there exist real numbers $C = C(\delta)$ and $N = N(\delta)$ such that for each prime $p > N$, each $q = p^\alpha$, each integer $n > 0$ and each irreducible plane curve $X \subseteq \mathbb{A}^2_{\mathbb{F}_q((t))}$ of*

MSC2010: primary 14G05; secondary 03C98, 11D88, 11G50.
Keywords: rational points, points of bounded height, parametrizations.

degree δ, one has

$$\#X(\mathbb{F}_q[t])_n \leq Cn^2 q^{\lceil n/\delta \rceil}.$$

A similar statement is proved by Sedunova [2017], for fixed q. More precisely, she proves that fixing δ, q and $\varepsilon > 0$, there exists a constant $C' = C'(\delta, q, \varepsilon)$ such that for each irreducible plane curve $X \subseteq \mathbb{A}^2_{\mathbb{F}_q[t]}$ of degree δ and positive integer n,

$$\#X(\mathbb{F}_q[t])_n \leq C' q^{n((1/\delta)+\varepsilon)}.$$

Observe that our result improves Sedunova's by replacing the ε factor by a polylogarithmic term. By the very nature of our methods, which are model-theoretic, we are however unable to establish such a result for q a power of a small prime p.

Recently, F. Vermeulen [2020] improved our Theorem A and Sedunova's results. More precisely, he obtains a variant of Theorem A for all primes p and with, moreover, a polynomial dependence of the constant C on the degree δ.

1.2. ***A uniform non-Archimedean point counting theorem.*** We state a uniform version of the Cluckers–Comte–Loeser non-Archimedean point counting theorem. A semialgebraic set is a set defined by a first-order formula in the language $\mathcal{L}_{\mathrm{div}} = \{0, 1, +, \cdot, |\}$ and parameters in $\mathbb{Z}[\![t]\!]$, where $|$ is a relation interpreted by $x \mid y$ if and only if $\mathrm{ord}(y) \leq \mathrm{ord}(x)$, with ord the valuation. As usual, we identify definable sets with the formulas that define them. Subanalytic sets are definable sets in the language obtained by adding a new symbol for each analytic function with coefficients in $\mathbb{Z}[\![t]\!]$ to the language $\mathcal{L}_{\mathrm{div}}$. For each local field L of characteristic zero, we fix a choice of uniformizer ϖ_L and view it as a $\mathbb{Z}[\![t]\!]$-ring by sending t to φ_L. Hence, we can consider the L-points of a semialgebraic or subanalytic set, for L a local field of any characteristic. The notion of semialgebraic and subanalytic sets considered in Section 5 is slightly more general than the one considered here; see also Setting 3.1.1.

The dimension of a subanalytic set X is the largest d such that there exists a coordinate projection p to a linear space of dimension d such that $p(X)$ contains an open ball. A subanalytic set is said to be of pure dimension d if for each $x \in X$ and every ball B centered at x, $X \cap B$ is of dimension d. If $X \subseteq L^n$, we denote by X^{alg} the union of all semialgebraic curves of pure dimension 1 contained in X. Observe that in general, X^{alg} is not semialgebraic (nor subanalytic).

If $X \subseteq K^m$ and $H \geq 1$, with K a field of characteristic zero, we denote by $X(\mathbb{Q}, H)$ the set of $x = (x_1, \ldots, x_m) \in X \cap \mathbb{Q}^m$ that can be written as $x_i = a_i/b_i$, with $a_i, b_i \in \mathbb{Z}$, $|a_i|, |b_i| \leq H$ (where $|\cdot|$ is the Archimedean absolute value). If $X \subseteq L^m$, where $L = \mathbb{F}_q((t))$, we denote by $X(\mathbb{F}_q(t), H)$ the set of $x = (x_1, \ldots, x_m) \in X \cap \mathbb{F}_q(t)^m$ that can be written as $x_i = a_i/b_i$, with $a_i, b_i \in \mathbb{F}_q[t]$ of degree less than or equal to $\log_q(H)$.

The following result is a particular case of Theorem 5.2.2. It provides a uniform version of Theorem 4.2.4 of [Cluckers et al. 2015].

Theorem B. *Let X be a subanalytic set of dimension m in n variables, with $m < n$. Fix $\varepsilon > 0$. Then there exists $C = C(X, \varepsilon)$, $N = N(X, \varepsilon)$, $\alpha = \alpha(n, m)$ and a semialgebraic set $W^\varepsilon \subseteq X$ such that for each*

$H \geq 1$ and each local field L, with residue field of characteristic $p_L > N$ and cardinal q_L, the following holds. We have $W^{\varepsilon}(L) \subseteq X(L)^{\mathrm{alg}}$ and if L is of characteristic zero,

$$\#(X \setminus W^{\varepsilon})(L)(\mathbb{Q}, H) \leq C(X, \varepsilon) q_L^{\alpha} H^{\varepsilon}.$$

If L is of positive characteristic, then

$$\#(X \setminus W^{\varepsilon})(L)(\mathbb{F}_{q_L}(t), H) \leq C(X, \varepsilon) q_L^{\alpha} H^{\varepsilon}.$$

An important step toward the proof of Theorem B is Proposition 5.1.4, which states that integer points of height at most H and lying in a subanalytic set X of dimension m in n variables are contained in $Cq^m \log(H)^{\alpha}$ algebraic hypersurfaces of degree $C' \log(H)^{\beta}$, where α and β are explicit constants depending only on n and m.

1.3. ***Uniform Yomdin–Gromov parametrizations.*** The proofs of Theorems A and B rely on the following parametrization result.

Fix a positive integer r. Let L be a local field, or more generally a valued field endowed with its ultrametric absolute value $|\cdot|$. A function $f : U \subseteq L^m \to L$ is said to satisfy T_r-approximation if for each $y \in U$ there is a polynomial $T_{f,y}^{<r}(x)$ of degree less than r and coefficients in L such that for each x, $y \in U$,

$$|f(x) - T_{f,y}^{<r}(x)| \leq |x - y|^r.$$

A T_r-parametrization of a set $X \subseteq L^n$ of dimension m is a finite partition of X into pieces $(X_i)_{i \in I}$ and for each $i \in I$, a subset $U_i \subseteq \mathcal{O}_L^m$ and a surjective function $f_i : U_i \to X_i$ that satisfies T_r-approximation.

The following statement is a particular case of Theorem 3.1.4.

Theorem C. *Let X be a subanalytic set included in some cartesian power of the valuation ring, and of dimension m. Then there exist integers C and N such that if L is a local field of residue characteristic $p_L \geq N$, then for each integer $r > 0$, there is a partition of $X(L)$ into Cr^m pieces such that for each piece X_i, there is a surjective function $f_i : U_i \subseteq \mathcal{O}_L^m \to X_i$ satisfying T_r-approximation on U_i.*

Observe that in the preceding theorem, we do not claim that the X_i and f_i are subanalytic, and indeed they are not in general.

Theorem C is used to deduce Theorems A and B, using an analog of the Bombieri–Pila determinant method. To be more precise, we follow closely the approach by Marmon [2010] in order to prove Theorem A.

Note also that from Theorem 3.1.3 of [Cluckers et al. 2015], we can deduce by compactness a result similar to Theorem C but for fixed r and with the number of pieces depending polynomially on the cardinality of the residue field. Such a result is however too weak to obtain a nontrivial bound in Theorem A.

The way we make Theorem C independent of the residue field is by adding algebraic Skolem functions in the residue field to the language. This enables us to work in a theory where the model-theoretic algebraic closure is equal to the definable closure. The functions involved in the parametrization are

definable in such an extension of the language. Theorem C is then deduced from a T_1-parametrization Theorem 3.4.2, where the functions are required to satisfy an extra technical condition called condition $(*)$ (see Definition 3.2.1). Such a condition implies that the function (when interpreted in any local field of large enough residue characteristic) is analytic on any box contained in its domain. This allows us to deduce the T_r-parametrization result by precomposing with power functions.

A first step toward Theorem C is Theorem 2.3.1, which states that the domain of a definable (in the above sense) function that is locally 1-Lipschitz can be partitioned into finitely many definable pieces on which the function is globally 1-Lipschitz. It is similar to Theorem 2.1.7 of [Cluckers et al. 2015], but there the domain is partitioned into infinitely many pieces parametrized (definably) by the residue field. The improvement is made possible by the fact that we work in a theory with algebraic Skolem functions in the residue field.

Let us finally observe that the number of pieces of the T_r-parametrization is Cr^m, where m is the dimension. In the Archimedean setting, a similar result has recently been proven by Cluckers, Pila and Wilkie [Cluckers et al. 2020], but there the number of pieces of the T_r-parametrization is a polynomial in r of nonexplicit degree in general; in the case of $\mathbb{R}_{\mathrm{an}}$, this degree in r has meanwhile been made explicit in Theorem 2 of [Binyamini and Novikov 2019] (see also the discussion just before Lemma 3.4.4).

The paper is organized as follows. Section 2 is devoted to the fact that one can go from local to global Lipschitz continuity. In Section 3, we prove our main parametrization result. Sections 4 and 5 are devoted to applications, the first to the counting of points of bounded degree in $\mathbb{F}_q[t]$, the second to the uniform non-Archimedean Pila–Wilkie theorem.

2. Global Lipschitz continuity

For $h : D \subseteq A \times B \to C$ any function between sets and for $a \in A$, write D_a for the set $\{b \in B \mid (a, b) \in D\}$ and $h(a, \cdot)$ or h_a for the function which sends $b \in D_a$ to $h(a, b)$. We use similar notation D_a and $h(a, \cdot)$ or h_a when D is a (subset of a) Cartesian product $\prod_{i=1}^n A_i$ and $a \in p(D)$ for some coordinate projection $p : D \to \prod_{i \in I \subseteq \{1,\dots,n\}} A_i$.

2.1. ***Tame theories.*** We consider tame structures in the sense of [Cluckers et al. 2015, Section 2.1]. We recall their definition here.

Let $\mathcal{L}_{\mathrm{Basic}}$ be the first-order language with the sorts VF, RF and VG, and symbols for addition and a constant 0 on VF; for functions $\overline{\mathrm{ac}} : \mathrm{VF} \to \mathrm{RF}$ and $|\cdot| : \mathrm{VF} \to \mathrm{VG}$; for the order, the multiplication and a constant 0 on VG; and for a constant 0 on RF. Let $\mathcal{L}$ be any expansion of $\mathcal{L}_{\mathrm{Basic}}$. By $\mathcal{L}$-definable we mean $\varnothing$-definable in the language $\mathcal{L}$, and likewise for other languages than $\mathcal{L}$. By contrast, we use the word "definable" more flexibly in this paper and it may involve parameters from a structure. Write $\mathrm{VF}^0 = \{0\}$, $\mathrm{RF}^0 = \{0\}$, and $\mathrm{VG}^0 = \{0\}$, with a slight abuse of notation. Note that $\mathcal{L}$ may have more sorts than $\mathcal{L}_{\mathrm{Basic}}$, since it is an arbitrary expansion.

We assume that all the $\mathcal{L}$-structures we consider are models of $\mathcal{T}_{\mathrm{Basic}}$, the $\mathcal{L}_{\mathrm{Basic}}$-theory stating that VF is an abelian group, that $\mathrm{VG} = \mathrm{VG}^\times \cup \{0\}$ with $\mathrm{VG}^\times$ a (multiplicatively written) ordered abelian group,

that $|\cdot| : \mathrm{VF} \to \mathrm{VG}$ is a surjective ultrametric absolute value (for groups), and that $\overline{\mathrm{ac}} : \mathrm{VF} \to \mathrm{RF}$ is surjective with $\overline{\mathrm{ac}}^{-1}(0) = \{0\}$.

Consider an $\mathcal{L}$-structure with K for the universe of the sort VF, k for RF and Γ for VG. We usually denote this structure by $(K, \mathcal{L})$.

Remark 2.1.1. Most often, K will be a valued field, k its residue field and Γ its value group (hence the sort names VF, RF and VG), although here we just require K to be a (valued) abelian group.

We define an open ball as a subset $B \subseteq K$ of the form $\{x \in K \mid |x - a| < \alpha\}$, for some $a \in K$ and $\alpha \in \Gamma^\times$, and similarly a closed ball as $\{x \in K \mid |x - a| \leq \alpha\}$.

We define $k^\times$ as $k \backslash \{0\}$. For $\xi \in k$ and $\alpha \in \Gamma$, we introduce the notation

$$A_{\xi,\gamma} = \{x \in K \mid \overline{\mathrm{ac}}(x) = \xi, |x| = \alpha\}.$$

Observe that if $\xi \in k^\times$ and $\alpha \in \Gamma^\times$, then $A_{\xi,\gamma}$ is an open ball.

We put on K the valuation topology, that is, the topology with the collection of open balls as base and the product topology on Cartesian powers of K.

For a tuple $x = (x_1, \ldots, x_n) \in K^n$, set $|x| = \max_{1 \leq i \leq n}\{|x_i|\}$.

Definition 2.1.2. Let $f : X \subseteq K^m \to K$ be a function. The function f is called 1-Lipschitz continuous (globally on X) or, in a short form, 1-Lipschitz if for all x and y in X,

$$|f(x) - f(y)| \leq |x - y|.$$

The function f is called locally 1-Lipschitz if, locally around each point of X, the function f is 1-Lipschitz continuous.

For $\gamma \in \Gamma^\times$, a function $f : X \subseteq K^n \to K$ is called γ-Lipschitz if for all x and y in X,

$$|f(x) - f(y)| \leq \gamma \cdot |x - y|.$$

Definition 2.1.3 (s-continuity). Let $F : A \to K$ be a function for some set $A \subseteq K$. We say that F is s-continuous if for each open ball $B \subseteq A$ the set $F(B)$ is either a singleton or an open ball, and there exists $\gamma = \gamma(B) \in \Gamma$ such that

$$|F(x) - F(y)| = \gamma |x - y| \quad \text{for all } x, y \in B. \tag{2.1.1}$$

If a function $g : U \subseteq K^n \to K$ on an open U is s-continuous in, say, the variable x_n, by which we mean that $g(a, \cdot)$ is s-continuous for each choice of $a = (x_1, \ldots, x_{n-1})$, then we write $|\partial g / \partial x_n(a, x_n)|$ for the element $\gamma \in \Gamma$ witnessing the s-continuity of $g(a, \cdot)$ locally at x_n, namely, γ is as in (2.1.1) for the function $F(\cdot) = g(a, \cdot)$, where x, y run over some ball B containing x_n and with $\{a\} \times B \subseteq U$.

Definition 2.1.4 (tame configurations). Fix integers $a \geq 0$, $b \geq 0$, a set

$$T \subseteq K \times k^a \times \Gamma^b,$$

and some $c \in K$. We say that T is in c-config if there is $\xi \in k$ such that T equals the union over $\gamma \in \Gamma$ of sets

$$(c + A_{\xi,\gamma}) \times U_\gamma$$

for some $U_\gamma \subseteq k^a \times \Gamma^b$. If moreover $\xi \neq 0$ we speak of an open c-config, and if $\xi = 0$ we speak of a graph c-config. If T is nonempty and in c-config, then ξ and the sets U_γ with $A_{\xi,\gamma}$ nonempty are uniquely determined by T and c.

We say that $T \subseteq K \times k^a \times \Gamma^b$ is in $\mathcal{L}$-tame config if there exist $s \geq 0$ and $\mathcal{L}$-definable functions

$$g : K \to k^s \quad \text{and} \quad c : k^s \to K$$

such that the range of c is finite, and, for each $\eta \in k^s$, the set

$$T \cap (g^{-1}(\eta) \times k^a \times \Gamma^b)$$

is in $c(\eta)$-config. We call c the center of T (despite not being in T in the case of open c-config).

For any $\mathcal{L}$-structure M elementarily equivalent to $(K, \mathcal{L})$ and for any language L obtained from $\mathcal{L}$ by adding some elements of M (of any sort) as constant symbols, call (M, L) a test pair for $(K, \mathcal{L})$.

Definition 2.1.5 (tameness). We say that $(K, \mathcal{L})$ is weakly tame if the following conditions hold.

(1) Each $\mathcal{L}$-definable set $T \subseteq K \times k^a \times \Gamma^b$ with $a \geq 0, b \geq 0$ is in $\mathcal{L}$-tame config.

(2) For any $\mathcal{L}$-definable function $F : X \subseteq K \to K$ there exist $s \geq 0$ and an $\mathcal{L}$-definable function $g : X \to k^s$ such that, for each $\eta \in k^s$, the restriction of F to $g^{-1}(\eta)$ is s-continuous.

We say that $(K, \mathcal{L})$ is tame when each test pair (M, L) for $(K, \mathcal{L})$ is weakly tame. Call an $\mathcal{L}$-theory $\mathcal{T}$ tame if for each model $\mathcal{M}$ of $\mathcal{T}$, the pair $(\mathcal{M}, \mathcal{L})$ is tame.

Recall [Cluckers et al. 2015, Corollary 2.1.11], which states that a tame theory, restricted in the sorts VF, RF, VG, is b-minimal, in the sense of [Cluckers and Loeser 2007]. In particular, one can make use of dimension theory for b-minimal structures.

2.2. ***Skolem functions.*** Recall that an $\mathcal{L}$-structure M has algebraic Skolem functions if for any $A \subseteq M$ every finite A-definable set $X \subseteq M^n$ admits an A-definable point. Observe that this condition is equivalent to the fact that the model-theoretic algebraic closure is equal to the definable closure. More generally, for a multisorted language, we say that a structure M has algebraic Skolem functions in the sort S if for any $A \subseteq M$ and every finite A-definable set $X \subseteq S_M^n$ there is an A-definable point, with S_M the universe for the sort S in the structure M.

We say that a theory T has algebraic Skolem functions (in the sort S), if each model has. In any case, one can algebraically Skolemize in the usual sense, that is, given a theory T in a language $\mathcal{L}$, the algebraic Skolemization of T in the sort S is the theory T^s in an expansion $\mathcal{L}^s$ of $\mathcal{L}$ obtained by adding function symbols, such that T^s has algebraic Skolem functions in the sort S and such that $(\mathcal{L}^s, T^s)$ is minimal with this property (where minimality is seen after identifying pairs with exactly the same models and definable sets); see also [Nübling 2004].

Lemma 2.2.1. *Let $\mathcal{L}$ a countable language extending $\mathcal{L}_{\mathrm{Basic}}$ and T a tame $\mathcal{L}$-theory. If T has algebraic Skolem functions in the sort* RF, *then it also has algebraic Skolem functions in the sort* VF. *In any case, there is a countable extension $\mathcal{L}'$ of $\mathcal{L}$ by function symbols on the sort* RF *and an $\mathcal{L}'$-theory T' extending T such that T' has algebraic Skolem functions in the sort* RF *and hence also in the sort* VF. *Moreover, every model of T can be extended to an $\mathcal{L}'$-structure that is a model of T', and, T' is tame.*

Proof. Since T is tame, every finite definable (with parameters) set in the VF sort is in definable bijection with a definable set in the RF sort. The first statement follows: If T has algebraic Skolem functions in the sort RF, then also in the sort VF. In general, let us algebraically Skolemize the theory T in the sort RF. Denote by $\mathcal{L}'$ and T' the obtained language and theory. Clearly one may take $\mathcal{L}'$ to be countable. It remains to prove that T' is tame. One needs to check condition (1) and (2) of Definition 2.1.5. Assume that $(K, \mathcal{L}')$ is a model of T' and let $T \subseteq K \times k^a \times \Gamma^b$ be some $\mathcal{L}'$-definable set. Then there is an $\mathcal{L}$-definable set T_0 such that $T \subseteq T_0$ and for each $(x, \xi_0, \alpha) \in T_0$, there is ξ such that $(x, \xi, \alpha) \in T$ and $(x, \xi_0, \alpha) \in \mathrm{acl}_{\mathcal{L}}(x, \xi, \alpha)$. Indeed, an $\mathcal{L}$-formula for T_0 is made from one for T by replacing each occurrence of a new function symbol by a formula for the definable set it lands in. The fact that T_0 is in $\mathcal{L}$-tame config then implies that T is in $\mathcal{L}'$-tame config. The reasoning for (2) is similar. □

Remark 2.2.2. Let $\mathcal{L}$ be an extension of $\mathcal{L}_{\mathrm{Basic}}$ such that any local field can be endowed with an $\mathcal{L}$-structure. Let T be an $\mathcal{L}$-theory such that any ultraproduct of local fields which is of residue characteristic zero is a model of T. Consider the algebraic Skolemization $\mathcal{L}'$, T' in the sort RF from Lemma 2.2.1. Then one can endow every local field with an $\mathcal{L}'$-structure such that moreover any ultraproduct of such structures that is of residue characteristic zero is a model of T'. Indeed, for each new function symbol in $\mathcal{L}' \setminus \mathcal{L}$ set the function output to be 0 if the corresponding set is empty, and to be any point in the set if nonempty. Such a choice of $\mathcal{L}'$-structure is often highly noncanonical and is not required to be compatible among field extensions.

Remark 2.2.3. Usually the Skolemization process breaks most of the model-theoretic properties of the theory. However, since we apply it only to the residue field many results such as cell decomposition are preserved. Moreover, since we add only algebraic Skolem functions in the sort RF, the situation is somehow controlled. For example, if the theory of the residue field is simple in the sense of model theory, then adding algebraic Skolem functions in the residue field preserves simplicity; see [Nübling 2004].

It is also worth noting that we will apply our results in the case where the residue field is pseudofinite, and that such fields almost always have algebraic Skolem functions; see [Beyarslan and Hrushovski 2012]. See also [Beyarslan and Chatzidakis 2017] for a more concrete characterization.

2.3. ***Lipschitz continuity.*** We can now state our first main result on Lipschitz continuity, going from local to piecewise global (with finitely many pieces).

Theorem 2.3.1. *Suppose that $(K, \mathcal{L})$ is tame with algebraic Skolem functions in the sort* RF. *Let $f : X \subseteq K^n \to K$ be an $\mathcal{L}$-definable function which is locally* 1*-Lipschitz. Then there exists a finite definable partition of X such that the restriction of f on each of the parts is* 1*-Lipschitz.*

As in [Cluckers et al. 2015], Theorem 2.3.1 is complemented by Theorem 2.3.2 about simultaneous partitions of domain and range into parts with 1-Lipschitz centers. They are proved by a joint induction on n.

Theorem 2.3.2 (Lipschitz continuous centers in domain and range). *Suppose that $(K, \mathcal{L})$ is tame with algebraic Skolem functions in the sort* RF. *Let $f : A \subseteq K^n \to K$ be an $\mathcal{L}$-definable function which is locally 1-Lipschitz. Then for a finite partition of A into definable parts, the following holds for each part X. There exist $s \geq 0$, a coordinate projection $p : K^n \to K^{n-1}$ and $\mathcal{L}$-definable functions*

$$g : X \to k^s, \quad c : p(X) \subseteq K^{n-1} \to K \quad \text{and} \quad d : p(X) \subseteq K^{n-1} \to K$$

such that c and d are 1-Lipschitz, and for each $\eta \in k^s$ and w in $p(X)$, the set $g^{-1}(\eta)_w$ is in $c(w)$-config and the image of $g^{-1}(\eta)_w$ under f_w is in $d(w)$-config.

Before proving Theorems 2.3.1 and 2.3.2, we obtain in Lemma 2.3.5 a weaker version of Theorem 2.3.2, where the centers are only required to be locally 1-Lipschitz. It will itself rely on [Cluckers et al. 2015, Theorem 2.1.8], which looks similar, but there the centers depend on auxiliary parameters.

Lemma 2.3.3. *Suppose that $(K, \mathcal{L})$ is tame with algebraic Skolem functions in the sort* RF. *Let $Y \subseteq K^n \times k^s$ be a definable set, $p : Y \to K^n$ be the canonical projection, $X = p(Y)$, and $f : X \to K$ be a definable function such that for each $\eta \in k^s$, the restriction of f to Y_η is locally 1-Lipschitz. Then there is a finite definable partition of X such that the restriction of f on each of the pieces is locally 1-Lipschitz.*

The proof of Lemma 2.3.3 is a joint induction with the following lemma.

Lemma 2.3.4. *Suppose that $(K, \mathcal{L})$ is tame with algebraic Skolem functions in the sort* RF. *Let $A \subseteq K^m$ be a definable set of dimension n. Then there is a finite definable partition of A such that for each part X, there is an injective projection $X \subseteq K^m \to K^n$ and its inverse is locally 1-Lipschitz.*

Proof of Lemma 2.3.4. Assume Lemma 2.3.3 holds for integers up to n. We use dimension theory for b-minimal structures. We get a finite definable partition of A such that on each piece X, there is a projection $p : X \to K^n$ which is finite-to-one. For each $w \in p(X)$, the fiber X_w is finite. By the existence of algebraic Skolem functions in the sort RF and hence also in VF by Lemma 2.2.1, each of the points of X_w is definable. By compactness, we can find a finite definable partition of X such that p is injective on each of the pieces.

By [Cluckers et al. 2015, Corollary 2.1.14], up to changing the coordinate projection we see that the inverse of p is locally 1-Lipschitz when restricted to fibers of some definable function $g : p(X) \to k^r$. By Lemma 2.3.3, we can find a finite partition of $p(X)$ such that the inverse of p is locally 1-Lipschitz on each of the parts. □

Proof of Lemma 2.3.3. We work by induction on n. If $n = 0$ there is nothing to prove. Assume now $n \geq 1$ and that Lemmas 2.3.3 and 2.3.4 hold for integers up to $n - 1$. Assume first that X is of dimension n. By dimension theory, there is at least one η such that Y_η is of dimension n. Define X' to be the union of the interior of Y_η for all such $\eta \in k^s$. The function f is locally 1-Lipschitz on X'. It remains to deal with $X'' = X \setminus X'$. By dimension theory, X'' is of dimension less than n. Assume $X'' = X$ for simplicity. By

Lemma 2.3.4, up to considering a finite definable partition of X we can assume that there is an injective coordinate projection $p : X \to K^{n-1}$ with inverse locally 1-Lipschitz. Then f is locally 1-Lipschitz if and only if $f \circ p^{-1}$ is. Now $p(X)$ with the function $f \circ p^{-1}$ satisfies the hypothesis of Lemma 2.3.3. By induction hypothesis, we have the result. □

Lemma 2.3.5. *Suppose that $(K, \mathcal{L})$ is tame with algebraic Skolem functions in the sort* RF. *Let $f : A \subseteq K^n \to K$ be an $\mathcal{L}$-definable function which is locally* 1*-Lipschitz. Then for a finite partition of A into definable parts, the following holds for each part X. There exist $s \geq 0$, a coordinate projection $p : K^n \to K^{n-1}$ and $\mathcal{L}$-definable functions*

$$g : X \to k^s, \quad c : p(X) \subseteq K^{n-1} \to K \quad \text{and} \quad d : K^{n-1} \to K$$

such that the functions c and d are locally 1*-Lipschitz, and for each w in $p(K^n)$, the set $g^{-1}(\eta)_w$ is in $c(w)$-config and the image of $g^{-1}(\eta)_w$ under f_w is in $d(w)$-config.*

The proof uses [Cluckers et al. 2015, Theorem 2.1.8], but only a weaker version is actually needed: we only need to require the centers to be locally 1-Lipschitz.

Proof. Apply [Cluckers et al. 2015, Theorem 2.1.8] to f. Work on one of the definable pieces X of A and use notations from the application of [Cluckers et al. 2015, Theorem 2.1.8], which is similar to Theorem 2.3.2 except that the input of c and d may additionally depend on some k-variables. We now show that these additional k-variables are not needed as input for c and d. We first show (after possibly taking a finite definable partition of X) that $c(\cdot, w)$ and $d(\cdot, w)$ are constant.

Fix some $w \in p(X)$. Since the range of the w-definable function $c_w : \eta \in k^s \mapsto c(\eta, w) \in K$ does not contain an open ball, it must be finite. By tameness, there is a w-definable bijection h_w between the range of c_w and a subset of $B_w \subseteq k^{s'}$, for some $s' \in \mathbb{N}$. By the existence of algebraic Skolem functions in the sort RF, and hence also in VF by Lemma 2.2.1, each of the points of B_w is w-definable. Taking the preimage of those points by $h_w \circ c_w$ leads to a w-definable finite partition of k^s. After taking preimages by g, this itself leads to a w-definable finite partition of X_w. By compactness, we find a finite partition of X such that on each piece, the function $c(g(x), p(x))$ is independent of $g(x) \in k^s$ and can be (abusively) written $c(p(x))$. The argument for d is similar.

By Lemma 2.3.3, we can refine the partition such that the functions $c, d : p(X) \to K$ are locally 1-Lipschitz. □

Proof of Theorem 2.3.2. We proceed by induction on n. Theorem 2.3.2 for $n = 1$ is exactly Lemma 2.3.5 for $n = 1$ since the Lipschitz condition is empty in this case. Assume now that Theorems 2.3.1 and 2.3.2 hold for integers up to $n - 1$. Apply Lemma 2.3.5. On each of the definable pieces X obtained, one has a coordinate projection p and definable functions $c, d : p(X) \to K$ that are locally 1-Lipschitz. By Theorem 2.3.1 for $n - 1$, we have a finite definable partition of $p(X)$ such that c and d are 1-Lipschitz on each of the pieces. This induces a finite definable partition of X satisfying the required properties. □

Proof of Theorem 2.3.1. We work by induction on n, assuming that Theorem 2.3.2 holds for integers up to n and Theorem 2.3.1 holds for integers up to $n - 1$. For $n = 0$ there is nothing to show, and hence

we assume that $n \geq 1$. Write $p : X \to K^{n-1}$ for the coordinate projection sending $x = (x_1, \ldots, x_n)$ to $\hat{x} = (x_1, \ldots, x_{n-1})$, and define Y as the image of X under the function $h : X \to K^n$ sending x to $(\hat{x}, f(x))$.

Up to taking a finite definable partition of X, switching the variables, by induction on the number of variables on which f depends, by Lemma 2.3.4 and Theorem 2.3.2, tameness and compactness, we may assume that the following holds:

- X is open in K^n,
- there is a definable function $g : X \to k^s$, and definable functions $c, d : p(X) \to K$,
- for each $\hat{x} \in p(X)$ and $\eta \in k^s$, $g^{-1}(\eta)_{\hat{x}}$ is in open $c(\hat{x})$-config and $h(g^{-1}(\eta))_{\hat{x}}$ is in $d(\hat{x})$-config,
- the restriction of $f(\hat{x}, \cdot)$ to $g^{-1}(\eta)_{\hat{x}}$ is s-continuous for each $\hat{x} \in p(X)$ and $\eta \in k^s$,
- the functions c and d are 1-Lipschitz,
- the function $f(\cdot, x_n)$ is 1-Lipschitz for each x_n.

We show that under these assumptions, f is 1-Lipschitz. Since d is 1-Lipschitz, we can replace f by $x \mapsto f(\hat{x}, x_n) - d(\hat{x})$ (and translate Y accordingly) in order to assume $d = 0$.

Let $x, y \in X$ and assume first that both x_n and y_n lie in an open ball $B \subseteq X_{\hat{x}}$. Then $g(x) = g(\hat{x}, y_n)$; indeed, otherwise $c(\hat{x}) \in B$, which would contradict that $g^{-1}(\eta)_{\hat{x}}$ is in open $c(\hat{x})$-config for every $\eta \in k^s$. It follows that $f(\hat{x}, \cdot)$ is s-continuous on B. Since f is locally 1-Lipschitz, the constant γ involved in the definition of s-continuity on B satisfies $\gamma \leq 1$.

Thus, using the ultrametric inequality and the assumption about $f(\cdot, y_n)$, we have

$$\begin{aligned} |f(x) - f(y)| &= |f(x) - f(\hat{x}, y_n) + f(\hat{x}, y_n) - f(y)| \\ &\leq \max\big(|f(x) - f(\hat{x}, y_n)|, |f(\hat{x}, y_n) - f(y)|\big) \\ &\leq \max\big(|x_n - y_n|, |\hat{x} - \hat{y}|\big) \\ &= |x - y|, \end{aligned}$$

which settles this case.

Suppose now that x_n and y_n do not lie in an open ball included in $X_{\hat{x}}$, and by symmetry nor in an open ball included in $X_{\hat{y}}$. This implies that

$$|x_n - c(\hat{x})| \leq |x_n - y_n| \quad \text{and} \quad |y_n - c(\hat{y})| \leq |x_n - y_n|. \tag{2.3.1}$$

By s-continuity and the fact that f is locally 1-Lipschitz, the image of a small enough open ball in $X_{\hat{x}}$ of radius α is either a point or an open ball of radius less than or equal to α. This implies that

$$|f(x) - d(\hat{x})| \leq |x_n - c(\hat{x})| \quad \text{and} \quad |f(y) - d(\hat{y})| \leq |y_n - c(\hat{y})|. \tag{2.3.2}$$

Recall that $d = 0$. Combining (2.3.1) and (2.3.2), we have by the ultrametric inequality

$$|f(x) - f(y)| \leq \max\big(|x_n - c(\hat{x})|, |y_n - c(\hat{y})|\big) \leq |x_n - y_n| \leq |x - y|,$$

which finishes the proof. □

Remark 2.3.6. Let us recall that [Cluckers et al. 2010] and [Cluckers and Halupczok 2012], with related results on Lipschitz continuity on p-adic fields, are amended in Remark 2.1.16 of [Cluckers et al. 2015]. When making $d = 0$ it is important to keep c possibly nonzero in the proof of [Cluckers et al. 2015, Theorem 2.1.7] and in the above proof of Theorem 2.3.1; this was forgotten in the proofs of the corresponding results [Cluckers et al. 2010, Theorems 2.3] and [Cluckers and Halupczok 2012, Theorem 3.5], where c should also have been kept.

3. Analytic parametrizations

The goal of this section is to prove a uniform version of non-Archimedean Yomdin–Gromov parametrizations.

3.1. *T_r-approximation.*

Setting 3.1.1. We fix for the whole section one of the two following settings, of $\mathcal{T}_{\mathrm{DP}}$ or $\mathcal{T}_{\mathrm{DP}}^{\mathrm{an}}$, both of which we now introduce. Let $\mathcal{O}$ be the ring of integers of a number field. Recall that the Denef–Pas language is a three sorted language, with one sort VF for the valued field with the ring language, one sort RF for the residue field with the ring language, one sort VG for the value group with the Presburger language with an extra symbol for ∞, and function symbols $\mathrm{ord} : \mathrm{VF} \mapsto \mathrm{VG}$ for the valuation (sometimes denoted multiplicatively $|\cdot|$) and $\overline{\mathrm{ac}} : \mathrm{VF} \to \mathrm{RF}$ for an angular component map (namely a multiplicative map sending 0 to 0 and sending a unit of the valuation ring to its reduction modulo the maximal ideal). Consider the theory of henselian discretely valued fields of residue field characteristic zero in the Denef–Pas language, with constants symbols from $\mathcal{O}[\![t]\!]$ and with t as a uniformizer of the valuation ring. This theory is tame by Theorem 6.3.7 of [Cluckers and Lipshitz 2011]. Applying Lemma 2.2.1, one obtains a new language and a new theory which we denote by $\mathcal{L}_{\mathrm{DP}}$ and $\mathcal{T}_{\mathrm{DP}}$, which thus has algebraic Skolem functions in each of the sorts.

We can also work in an analytic setting corresponding to Example 4.4(1) of [Cluckers and Lipshitz 2011], as follows. Consider the expansion of the Denef–Pas language $\mathcal{L}_{\mathrm{DP}}$ by adding function symbols for elements of

$$\mathcal{O}[\![t]\!]\{x_1, \ldots, x_n\} = \left\{ f = \sum_{I \in \mathbb{N}^n} a_I x^I \;\middle|\; a_I \in \mathcal{O}[\![t]\!], \mathrm{ord}_t(a_I) \underset{|I| \to +\infty}{\longrightarrow} +\infty \right\}.$$

Any complete discretely valued field over $\mathcal{O}$ (namely, with a unital ring homomorphism from $\mathcal{O}$ into the valued field) can be endowed with a structure for this expansion, by interpreting the new function symbols as the corresponding power series evaluated on the unit box and put equal to zero outside the unit box. Let $\mathcal{L}_{\mathrm{DP}}^{\mathrm{an}}$ and $\mathcal{T}_{\mathrm{DP}}^{\mathrm{an}}$ be the resulting language and the theory of these models, respectively. (For a shorter and explicit axiomatization for the analytic case, see the axioms of Definition 4.3.6(i) of [Cluckers and Lipshitz 2011].)

From now on, we work in a language $\mathcal{L}$ that is either $\mathcal{L}_{\mathrm{DP}}$ or $\mathcal{L}_{\mathrm{DP}}^{\mathrm{an}}$ and in the theory $\mathcal{T}$ that is correspondingly $\mathcal{T}_{\mathrm{DP}}$ or $\mathcal{T}_{\mathrm{DP}}^{\mathrm{an}}$.

Let us summarize our theory once more: $\mathcal{T}$ is the $\mathcal{L}$-theory which is the algebraic Skolemization in the residue field sort of the theory of complete discrete valued fields, residue field of characteristic zero, with constants symbols from $\mathcal{O}[\![t]\!]$ (as a subring) and where t has valuation 1, and (in the subanalytic case), with the restricted analytic function symbols as the corresponding power series evaluated on the unit box and put equal to zero outside the unit box.

In any case, the theory $\mathcal{T}$ is tame by Theorem 6.3.7 of [Cluckers and Lipshitz 2011], and, it has algebraic Skolem functions in each sort by Lemma 2.2.1 and by Example 4.4(1) with the homothecy with factor t on the valuation ring to make the system strict instead of separated. Note that there is no need to algebraically Skolemize again when going from $\mathcal{T}_{\mathrm{DP}}$ to the larger theory $\mathcal{T}_{\mathrm{DP}}^{\mathrm{an}}$ by the elimination of valued field quantifiers from Theorem 6.3.7 of [Cluckers and Lipshitz 2011]. Definable means definable without parameters in the theory $\mathcal{T}$.

Definition 3.1.2 (T_r-approximation). Let L be any valued field. Consider a set $P \subseteq L^m$, a function $f = (f_1, \ldots, f_n) : P \to \mathcal{O}_L^n$ and an integer $r > 0$. We say that f satisfies *T_r-approximation* if P is open in L^m, and, for each $y \in P$, there is an n-tuple $T_{f,y}^{<r}$ of polynomials with coefficients in $\mathcal{O}_L$ and of degree less than r that satisfies

$$|f(x) - T_{f,y}^{<r}(x)| \leq |x - y|^r \quad \text{for all } x \in P.$$

We say that a family $(g_i)_{i\in I}$ of functions $g_i : P_i \to X_i \subseteq \mathcal{O}_L^n$ is a T_r-parametrization of $X = \bigcup_{i\in I} X_i$ if each g_i is surjective and satisfies T_r-approximation.

Observe that if f satisfies T_r-approximation, then the polynomials $T_{f,y}^{<r}$ are uniquely determined.

Observe also that if K is a complete valued field of characteristic zero, if f is of class $\mathscr{C}^r$ (i.e., f is r times differentiable and the r-th differential is continuous) and satisfies T_r-approximation, then $T_{f,y}^{<r}$ is just the tuple of Taylor polynomials of f at y of order r.

Notation 3.1.3. Let $\mathcal{O}$ be the ring of integers of a number field. We denote by $\mathcal{A}_{\mathcal{O}}$ the collection of all local fields of characteristic zero over $\mathcal{O}$ and by $\mathcal{B}_{\mathcal{O}}$ those of positive characteristic, and set $\mathscr{C}_{\mathcal{O}} = \mathcal{A}_{\mathcal{O}} \cup \mathcal{B}_{\mathcal{O}}$. (By a local field L over $\mathcal{O}$ we mean a non-Archimedean locally compact field, i.e., a finite field extension of $\mathbb{Q}_p$ or of $\mathbb{F}_p((t))$ for a prime p, allowing a unital homomorphism $\mathcal{O} \to L$.) If $L \in \mathscr{C}_{\mathcal{O}}$, we denote by ord its valuation (normalized such that $\mathrm{ord}(L^\times) = \mathbb{Z}$), $\mathcal{O}_L$ its valuation ring, $\mathcal{M}_L$ its maximal ideal, $\varpi_L \in \mathcal{M}_L$ a fixed choice of uniformizer, k_L its residue field, q_L the cardinality of k_L and p_L the characteristic of k_L. If $N \in \mathbb{N}$, we define $\mathcal{A}_{\mathcal{O},N}$ (resp. $\mathcal{B}_{\mathcal{O},N}$, $\mathscr{C}_{\mathcal{O},N}$) to be the set of $L \in \mathcal{A}_{\mathcal{O}}$ (resp. $L \in \mathcal{B}_{\mathcal{O}}$, $L \in \mathscr{C}_{\mathcal{O}}$) such that $p_L \geq N$. By Remark 2.2.2, we can consider $L \in \mathscr{C}_{\mathcal{O}}$ as an $\mathcal{L}$-structure, and any nonprincipal ultraproduct of residue characteristic zero of such local fields is a model of $\mathcal{T}$.

A family $(X_y)_{y\in Y}$ of sets X_y indexed by $y \in Y$ is called a definable family if the total set $\mathcal{X} := \{(x, y) \mid x \in X_y,\ y \in Y\}$ (and hence also Y) is a definable set. Likewise, a family of functions is called a definable family if the family of graphs is a definable family. We use notations like $\mathcal{O}_{\mathrm{VF}}$ for the definable set which in any model K is the valuation ring $\mathcal{O}_K$, and similarly $\mathcal{M}_{\mathrm{VF}}$ for the maximal ideal, and so

on. For a definable set X and a structure L, we write $X(L)$ for the L-points on X, and for a definable function $f : X \to Y$, we write f_L for the corresponding function $X(L) \to Y(L)$.[1]

The main goal of this section is to prove the next two theorems on the existence of T_r-parametrizations with rather few maps, in terms of r. Even the mere finiteness of the parametrizing maps is new, as compared to [Cluckers et al. 2015] where "residue many" maps were allowed, but we even get an upper bound which is polynomial in r. This finiteness is crucial for Theorem A, and, useful for Theorem B, where it makes the exponent α of q_L independent of X. Recall from Setting 3.1.1 that we work in a theory with algebraic Skolem functions.

Theorem 3.1.4 (uniform T_r-approximation in local fields). *Let $n \geq 0$, $m \geq 0$ be integers and $X = (X_y)_{y \in Y}$ a definable family of subsets $X_y \subseteq \mathcal{O}_{\mathrm{VF}}^n$, for y running over a definable set Y. Suppose that X_y has dimension m for each $y \in Y$ (and in each model of $\mathcal{T}$). Then there exist integers $c > 0$ and $M > 0$ such that for each $L \in \mathscr{C}_{\mathcal{O},M}$ and for each integer $r > 0$, there are a finite set $I_{r,q}$ of cardinality cr^m and an R_r-definable family $g = (g_{y,i})_{(y,i) \in Y(L) \times I_r}$ of (R_r, y)-definable functions*

$$g_{y,i} : P_{y,i} \to X_y(L)$$

with $P_{y,i} \subseteq \mathcal{O}_L^m$ such that for each $y \in Y(L)$, the family $(g_{y,i})_{i \in I_{r,q}}$ forms a T_r-parametrization of $X_y(L)$ and $R_r \subset \mathcal{O}_L^\times$ is a set of lifts of representatives for the r-th powers in $\mathbb{F}_{q_L}^\times$.

Note that Theorem C in the introduction is a particular case and a less precise version of Theorem 3.1.4.

The following result is uniform in all models K of $\mathcal{T}$. Note that $\mathcal{T}$ requires in particular the residue field to have characteristic zero, and the value group to be elementarily equivalent to $\mathbb{Z}$.

Theorem 3.1.5 (uniform T_r-approximation for models of $\mathcal{T}$). *Let $n \geq 0$, $m \geq 0$ be integers and let $X = (X_y)_{y \in Y}$ be a definable family of subsets $X_y \subseteq \mathcal{O}_{\mathrm{VF}}^n$, for y running over a definable set Y. Suppose that X_y has dimension m for each $y \in Y$ and each model of $\mathcal{T}$. Then there exists an integer $c > 0$ such that for each model K of $\mathcal{T}$ and for each integer $r > 0$ such that the r-th powers in the residue field have a finite number $b_r = b_r(K)$ of cosets, there are a finite set I_r of cardinality $c(b_r r)^m$ and an R_r-definable family $g = (g_{y,i})_{(y,i) \in Y(K) \times I_r}$ of (R_r, y)-definable functions*

$$g_{y,i} : P_{y,i} \to X_y(K)$$

with $P_{y,i} \subseteq \mathcal{O}_K^m$ such that for each $y \in Y(K)$, the family $(g_{y,i})_{i \in I_r}$ forms a T_r-parametrization of $X_y(K)$ and $R_r \subset \mathcal{O}_K^\times$ is a set of lifts of representatives for the r-th powers in $k^\times$.

Remark 3.1.6. Observe that even if Theorems 3.1.4 and 3.1.5 are very similar, one cannot deduce the first from the second by compactness. The reason is the quantification over r in the statement. They will, however, both be deduced from the upcoming Theorem 3.4.2, which is a T_1-parametrization theorem

[1] When we interpret definable sets or functions into local fields L (or, more generally, $\mathcal{L}$-structures that are not models of our theory $\mathcal{T}$), we implicitly assume that we have chosen some formula φ that defines the set and consider $\varphi(L)$. This set $\varphi(L)$ may of course change with a different choice of formula φ for small values of the residue field characteristic of L, but this is not a problem by Remark 2.2.2, and since we are interested only in the case of large residue field characteristic.

with an extra technical condition. It will allow us to define a T_r-parametrization by precomposing by power functions. Furthermore, note that in Theorem 3.1.4, the factor b_r for the index of r-th powers in the residue field is not needed; this is because of an additional trick using a property true in finite fields.

Remark 3.1.7. For most of the section, we could in fact work in a slightly more general setting (up to imposing some additional requirements for Theorem 3.1.4). Using resplendent relative quantifier elimination as in [Rideau 2017], we can add arbitrary constant symbols and allow an arbitrary residual extension (and an arbitrary extension on the value group) of the language and the theory before applying the algebraic Skolemization in the residue field sort. In particular, Theorem 3.1.5 holds in this more general setting. If the extended language and theory still have the property that any local field can be equipped with a structure for the extended language such that, moreover, any ultraproduct of such equipped local fields which is of residue characteristic zero is a model of the extended theory, then also Theorem 3.1.4 would go through.

Remark 3.1.8. The condition that the value group be a Presburger group can probably be relaxed to any value group in which the index v_r of the subgroup of r-multiples is finite, by replacing $c(b_r r)^m$ by $c(b_r v_r)^m$ for the cardinality of I_r and taking $R_r \cup V_r$ instead of R_r, V_r a set of lifts of representatives for the r-multiples in the value group.

Note that extending Theorem 3.1.5 and its proof to mixed characteristic henselian valued fields may be possible too, with the adequate adaptations. For example, when going from local to piecewise Lipschitz continuity, the Lipschitz constant should be allowed to grow. (Indeed, look at the function $x \mapsto x^p$ on the valuation ring of $\mathbb{C}_p$.)

Before starting the proofs of Theorems 3.1.4 and 3.1.5, we need a few more definitions.

Definition 3.1.9 (cell with center). Consider an integer $n \geq 0$. For nonempty definable sets Y and $X \subseteq Y \times \mathrm{VF}^n$, the set X is called a cell over Y with center $(c_i)_{i=1,\dots,n}$ if it is of the form

$$\left\{(y, x) \in Y \times \mathrm{VF}^n \mid \left(y, \left(\overline{\mathrm{ac}}(x_i - c_i(x_{<i})), |x_i - c_i(x_{<i})|\right)_{i=1}^n\right) \in G\right\},$$

for some definable set $G \subseteq Y \times \mathrm{RF}^n \times \mathrm{VG}^n$ and some definable functions and $c_i : Y \times \mathrm{VF}^{i-1} \to \mathrm{VF}$, where $x_{<i} = (y, x_1, \dots, x_{i-1})$. If moreover G is a subset of $Y \times (\mathrm{RF}^\times)^n \times (\mathrm{VG}^\times)^n$, where $(\mathrm{VG}^\times)^0 = \{0\}$, then X is called an open cell over Y (with center $(c_i)_{i=1,\dots,n}$).

We next give a special name to cells over Y whose center equals 0.

Definition 3.1.10 (cell around zero). We say that a nonempty set $X \subseteq Y \times \mathrm{VF}^n$ is a cell around zero (over Y) if it is of the form

$$X = \left\{(y, x) = (y, x_1, \dots, x_n) \in Y \times \mathrm{VF}^n \mid \left(y, (\overline{\mathrm{ac}}(x_i), |x_i|)_{i=1}^n\right) \in G\right\}$$

for some definable set $G \subseteq Y \times \mathrm{RF}^n \times \mathrm{VG}^n$. Similarly, one can call a set X a cell around zero (over Y) for $X \subset L^n$ for some valued field L with an angular component map, if it is of the corresponding form.

Definition 3.1.11 (associated cell around zero). Let X be a cell over Y with center, with notation from Definition 3.1.9. The cell around zero associated to X is by definition the cell $X^{(0)}$ obtained by forgetting the centers, namely

$$X^{(0)} = \{(y, x) \in Y \times \mathrm{VF}^n \mid y \in Y,\ \overline{\mathrm{ac}}(x_i) = \xi_i(y),\ (y, (|x_i|)_i) \in G\}$$

with associated bijection $\theta_X : X \to X^{(0)}$ sending (y, x) to $(y, (x_i - c_i(x_{<i}))_i)$. For a definable map $f : X \to Z$ there is the natural corresponding function $f^{(0)} = f \circ \theta_X^{-1}$ from $X^{(0)}$ to Z.

Definition 3.1.12 (associated box). Let K be a valued field. By a box $B \subset K^n$ we mean a product of open balls in K. Let $B = \prod_{1 \le i \le n} B(a_i, r_i) \subseteq K^n$ be a box, with open balls

$$B(a_i, r_i) = \{x \in K \mid |x - a_i| < r_i\},$$

with $a_i \in K$ and nonzero $r_i \in \Gamma_K$. The box associated to B is the box $B_{\mathrm{as}} \subseteq K^{\mathrm{alg}}$ defined by

$$B_{\mathrm{as}} = \{x \in (K^{\mathrm{alg}})^n \mid |x - a_i| < r_i\},$$

where K^{alg} is an algebraic closure of K, endowed with the canonical extension of the valuation of K.

We now define the term language. This is an expansion $\mathcal{L}^*$ of $\mathcal{L}$, by joining division and witnesses for henselian zeros and roots.

Definition 3.1.13. Let $\mathcal{L}^*$ be the expansion of $\mathcal{L} \cup \{^{-1}\}$ obtained by joining to $\mathcal{L} \cup \{^{-1}\}$ function symbols h_m and root_m for integers $m > 1$. The h_m are interpreted on a henselian valued field K of equicharacteristic zero and residue field k as the functions

$$h_m : K^{m+1} \times k \to K$$

sending $(a_0, \ldots, a_m, \xi)$ to the unique y satisfying $\mathrm{ord}(y) = 0$, $\overline{\mathrm{ac}}(y) \equiv \xi \mod \mathcal{M}_K$, and $\sum_{i=0}^m a_i y^i = 0$, whenever ξ is a unit, $\mathrm{ord}(a_i) \ge 0$, $\sum_{i=0}^m a_i \xi^i \equiv 0 \mod \mathcal{M}_K$, and

$$f'(\xi) \not\equiv 0 \mod \mathcal{M}_K$$

with f' the derivative of f, and to 0 otherwise. Likewise, root_m is the function $K \times k \to K$ sending (x, ξ) to the unique y with $y^m = x$ and $\overline{\mathrm{ac}}(y) = \xi$ if there is such y, and to 0 otherwise.

Proposition 3.1.14 (term structure of definable functions). *Every* VF*-valued definable function is piecewise given by a term. More precisely, given a definable set X and a definable function $f : X \to \mathrm{VF}$, there exists a finite partition of X into definable parts and for each part A an $\mathcal{L}^*$-term t such that*

$$t(x) = f(x) \quad \textit{for all } x \in A.$$

Proof. By Theorem 7.5 of [Cluckers et al. 2006] there exists a definable function $g : X \to \mathrm{RF}^m$ for some $m \ge 0$ and an $\mathcal{L}^*$-term t_0 such that

$$t_0(x, g(x)) = f(x).$$

Since the terms h_n (the henselian witnesses) and root_n (the root functions) involve at most a finite choice in the residue field, one can reduce to the case that g has finite image. The fibers of g can then be taken as part of the partition to end the proof. □

3.2. *Condition* (∗). We now introduce a technical condition, named (∗), that will be used in Section 3.3 to show a strong form of analyticity of definable functions, named global analyticity in Definition 3.3.1.

Definition 3.2.1 (condition (∗)). We first define condition (∗) for $\mathcal{L}^*$-terms, inductively on the complexity of terms. Consider a definable set $X \subseteq \mathrm{VF}^m$ and let x run over X.

We say that a VF-valued $\mathcal{L}^*$-term $t(x)$ satisfies condition (∗) on X if the following holds.

If $t(x)$ is a term of complexity 0 (i.e., a constant or a variable), then it satisfies condition (∗) on X.

Suppose now that the term t is of the form $t_1 + t_2$, $t_1 \cdot t_2$, t_0^{-1}, $h_n(t_0, \ldots, t_n; t_{-1})$, $\mathrm{root}_n(t_0; t_{-1})$ for some $n > 0$, or $\underline{f}(t_1, \ldots, t_n)$, with $\underline{f}$ one of the analytic functions of the language. In the first two cases, we just require that t_1 and t_2 satisfy condition (∗) on X. In the remaining four cases, we require that $t_0, \ldots, t_n$ satisfy condition (∗) on X and moreover that for any box $B \subseteq X$, the functions t_{-1} and $\overline{\mathrm{ac}}(t_0), \ldots, \overline{\mathrm{ac}}(t_n), \mathrm{ord}(t_0), \ldots, \mathrm{ord}(t_n)$ are constant on B.

We finally say that an $\mathcal{L}$-definable function $f : X \subseteq \mathrm{VF}^m \to \mathrm{VF}^{m'}$ for $m' > 0$ satisfies condition (∗) on X if there is a tuple t of $\mathcal{L}^*$-terms $t_i(x)$ satisfying condition (∗) on X and such that $f(x) = t(x)$ for $x \in X$.

The following lemma ensures existence of functions satisfying condition (∗).

Lemma 3.2.2. *Let $f : X \subseteq Y \times \mathrm{VF}^m \to \mathrm{VF}^{m'}$ be a definable function for some m and m'. Then there is a finite partition of X into some open cells A over Y with center $(c_i)_{i=1,\ldots,m}$ and a set B such that B_y is of dimension less than m for each $y \in Y$, such that the function*

$$(A^{(0)})_y \to \mathrm{VF}^{m'} : x \mapsto f^{(0)}(y, x)$$

satisfies condition (∗) on $(A^{(0)})_y$ for each y, with notation from Definition 3.1.11.

Proof. We proceed by induction on m. By Proposition 3.1.14 for f we may suppose that f is given by a tuple $t(x)$ of $\mathcal{L}^*$-terms. Let $h : X \to \mathrm{RF}^s \times \Gamma^{s'}$ be the definable function created from t such that h has a component function of the form t' for each RF-valued subterm t' of t and also of the forms $\mathrm{ord}(t'')$ and $\overline{\mathrm{ac}}(t'')$ for each VF-valued subterm t'' of t. The proposition requires us to find a finite partition of X into cells over Y such that for each open cell A over Y, the map $(f_{|A})^{(0)}(y, \cdot)$ has condition (∗) on $A_y^{(0)}$, with notation from Definition 3.1.11. Now apply the cell decomposition theorem adapted to h and work on one of the open pieces A. Thus, A is an open cell over Y with some center $(c_i)_{i=1,\ldots,m}$ adapted to h, namely, there are definable functions $c_i : A^i \subseteq \mathrm{VF}^i \to \mathrm{VF}$ for $i = 0, \ldots, m-1$ such that $h^{(0)}$ is constant on each box contained in $c^{-1}(A)$, which is moreover an open cell around zero, where

$$c : x \in \mathrm{VF}^m \mapsto (x_1 + c_0, x_2 + c_1(x), \ldots, x_m + c_{m-1}(x)),$$

with notation from Definition 3.1.11. Note that $c = \theta_A^{-1}$ and $c^{-1}(A) = A^{(0)}$ in that notation. □

3.3. ***Global analyticity.*** To more easily speak of analyticity in this section, we work with complete discretely valued fields (a meaning of analyticity exists for all models of $\mathcal{T}$ by [Cluckers and Lipshitz 2011]).

Definition 3.3.1 (globally analytic map). Let K be a complete discretely valued field. Let $X \subseteq K^m$ be a set and $f : X \to K^n$ a function. We say that f is globally analytic on X if for each box $B \subseteq X$, the restriction of f to B is given by a tuple of power series with coefficients in K (say, taken around some $a \in B$), which converges on the associated box B_{as}.[2]

The following proposition is the reason why we introduced condition $(*)$. Observe that it applies also to local fields, and thus not only to models of our theory $\mathcal{T}$.

Proposition 3.3.2 (analyticity, [Cluckers and Lipshitz 2011, Lemma 6.3.15]). *Let f be a definable function satisfying condition $(*)$ on some definable set X. Then there is some $M > 0$ such that for L which is either a local field with residue field cardinality at least M, or a model of $\mathcal{T}$ which is moreover a complete discretely valued field, the following holds. For any box $B \subseteq X(L)$ and $b \in B$, there is a power series g centered at b and converging on B_{as} such that f is equal to g on B. Moreover, M can be taken uniformly in definable families of definable functions.*

Proof. We recall the strategy of the proof of [Cluckers and Lipshitz 2011, Lemma 6.3.15]. One works by induction on the complexity of the $\mathcal{L}^*$-term corresponding to the definition of condition $(*)$, using compositions of power series as in Remark 4.5.2 of [Cluckers and Lipshitz 2011]. The only nontrivial cases are t_0^{-1}, $h_n(t_0, \ldots, t_n; t_{-1})$, $\mathrm{root}_n(t_0; t_{-1})$, and $\underline{f}(t_1, \ldots, t_n)$ for some restricted analytic function $\underline{f}$ from the language. If L is a model of $\mathcal{T}$, we may assume by the definition of condition $(*)$ that the terms t_i satisfy condition $(*)$ on X and that t_{-1} and $\overline{\mathrm{ac}}(t_0), \ldots, \overline{\mathrm{ac}}(t_n), \mathrm{ord}(t_0), \ldots, \mathrm{ord}(t_n)$ are constant on B. In the local field case, by compactness there is some $M > 0$ such that if the residue field of L is of cardinality at least M, the functions t_{-1} and $\overline{\mathrm{ac}}(t_0), \ldots, \overline{\mathrm{ac}}(t_n), \mathrm{ord}(t_0), \ldots, \mathrm{ord}(t_n)$ are constant on any box B contained in $X(L)$. One finishes exactly as in the proof of [Cluckers and Lipshitz 2011, Lemma 6.3.11], where for the case $\underline{f}(t_0, \ldots, t_n)$, with $\underline{f}$ one of the analytic functions of the language, condition $(*)$ ensures that either the function f is interpreted as the zero function on a box B or the image of the box B by $(t_1, \ldots, t_n)$ is strictly contained in the unit box, whence so is the image of B_{as}, ensuring convergence of f on it, and giving analyticity of $\underline{f}(t_0, \ldots, t_n)$ on B_{as}. □

3.4. ***Strong T_r-approximation.*** We can now state a stronger notion of T_r-approximation, for definable functions. The strong T_1-approximation will be key for the proofs of Theorems 3.1.4 and 3.1.5. Strong T_r-approximation for $r > 1$ is not needed in this paper, but we include its definition for the sake of completeness.

Definition 3.4.1 (strong T_r-approximation). Let $P \subseteq \mathrm{VF}^m$ be definable, $f = (f_1, \ldots, f_n) : P \to \mathrm{VF}^n$ a definable function, and $r > 0$ an integer.

[2] Here, converging on B_{as} means that the partial sums obtained by evaluating at any element of B_{as} form a Cauchy sequence (the limits actually lie inside K^{alg} by [Cluckers and Lipshitz 2011]).

(1) We say that f satisfies strong T_r-approximation if P is an open cell around zero, f satisfies condition $(*)$ on P and, for each model L of $\mathcal{T}$, the function f_L satisfies T_r-approximation and moreover for each box $B \subseteq P(L)$, the $\mathcal{L}^*$-term associated to f satisfies T_r-approximation on B_{as}.

(2) A family $f_i : P_i \to X$ for $i \in I$ of definable functions is called a (strong) T_r-parametrization of $X \subseteq \mathrm{VF}^n$ if each f_i is a (strong) T_r-approximation and

$$\bigcup_{i \in I} f_i(P_i) = X.$$

The fact that P is an open cell around zero in Definition 3.4.1 is particularly handy since it enables an easy description of the maximal boxes contained in P, which combines well with condition $(*)$ and for composing with power maps. Global analyticity in complete models as given in Section 3.3, together with a calculation on the coefficients of the occurring power series, will then complete the proofs of the parametrization Theorems 3.1.4 and 3.1.5.

Theorem 3.4.2 (strong T_1-parametrization). *Let $n \geq 0$, $m \geq 0$ be integers and let $X = (X_y)_{y \in Y}$ be a definable family of subsets $X_y \subseteq \mathcal{O}_{\mathrm{VF}}^n$ for y running over a definable set Y. Suppose that X_y has dimension m for each $y \in Y$. Then there exist a finite set I and a definable family $g = (g_{y,i})_{(y,i) \in Y \times I}$ of definable functions*

$$g_{y,i} : P_{y,i} \to X_y$$

such that $P_{y,i} \subseteq \mathcal{O}_{\mathrm{VF}}^m$ and for each y, $(g_{y,i})_{i \in I}$ forms a strong T_1-parametrization of X_y.

Proof. We work by induction on m. We repeatedly throw away pieces of lower dimension and treat them by induction, working uniformly in y. We also successively consider finite definable partitions of X without renaming. By Lemma 2.3.4, up to taking a finite definable partition of X, we can find a locally 1-Lipschitz surjective function $f_y : P_y \subseteq \mathrm{VF}^m \to X_y$, with P_y open for each $y \in Y$. By Theorem 2.3.1, we can further assume that f_y is globally 1-Lipschitz on P_y, or equivalently, that f_y satisfies T_1-approximation on P_y. By Proposition 3.1.14 we may moreover suppose that the component functions of f are given by $\mathcal{L}^*$-terms. We still need to improve f and P in order for the f_y to satisfy strong T_1-approximation, in particular, condition $(*)$, T_1-approximation on associated boxes of boxes in its domain, and that P_y is an open cell around zero.

First we ensure, as an auxiliary step, that the first partial derivatives of the f_y are bounded by 1 on the associated box of any box in its domain P_y, by passing to an algebraic closure $\mathrm{VF}^{\mathrm{alg}}$ of VF with the natural $\mathcal{L}$ and $\mathcal{L}^*$ structures. This passage to $\mathrm{VF}^{\mathrm{alg}}$ preserves well properties of quantifier-free formulas and of terms by results from [Cluckers and Lipshitz 2011; 2017] for the involved analytic structures on VF and on $\mathrm{VF}^{\mathrm{alg}}$. This step is done by switching again the order of coordinates as in the proof of Lemma 2.3.4 where necessary. Since it is completely similar to the corresponding part of the proof of [Cluckers et al. 2015, Theorem 3.1.3], we skip the details.

Finally we show that we can ensure all remaining properties, using induction. Apply Lemma 3.2.2, uniformly in y, to obtain a partition of $P = (P_y)_y$ into open cells $A = (A_y)_y$ over Y with center $(c_i)_{i=1}^m$

and an associated bijection θ_A in the notation of Definition 3.1.11, while neglecting a definable subset B of P where B_y is of dimension less than m. By induction on m, we may apply Theorem 3.4.2 (for the value $m-1$) to the graph of $(c_i)_{i=1}^m$ to find a strong T_1-parametrization for this graph. One obtains the required parametrization of X by composing the parametrization of the graph of $(c_i)_{i=1}^m$ with θ_A^{-1} and f. Indeed, one first concludes as in the proof of Lemma 3.2.2 that property $(*)$ is satisfied for this composition and that the domain is an open cell around zero. Secondly, the composition of 1-Lipschitz functions is 1-Lipschitz, and the first-order partial derivatives are bounded by 1 on associated boxes of its domain. Finally, the condition of T_1-approximation on each associated box follows from Proposition 3.3.2 and [Cluckers et al. 2015, Corollary 3.2.12], since the derivative is bounded by 1 on associated boxes of its domain. □

The whole purpose of requiring the domains of strong T_1-parametrizations to be cells around zero is to deduce existence of T_r-parametrizations from strong T_1-parametrizations by precomposing with power functions. This is enabled by the next two lemmas.

Lemma 3.4.3. *Let f be a definable function on $X \subset \mathrm{VF}$ satisfying strong T_1-approximation. Then there is some $M > 0$ such that for L either a model of $\mathcal{T}$ which is a complete discretely valued field, or a local field with residue field cardinality at least M, the following holds for any integer $r > 0$ and with p_r being the r-power map sending x in L to x^r. For any open ball $B = b(1+\mathcal{M}_L) \subseteq L$ with $B \subset X$, and for any ball $D \subseteq L$ satisfying $p_r(D) \subseteq B$, the function*

$$f_r := f_L \circ p_r$$

satisfies T_r-approximation on D. Moreover, f_r can be developed around any point $b' \in D$ as a power series which is converging on D_{as} and whose coefficients c_i satisfy

$$|c_i| \leq |b'|^{r-i} \quad \text{for all } i > 0.$$

Proof. Observe first that since the choice of $b \in B$ is arbitrary, it suffices to show the lemma for $b' \in D$ with $b'^r = b$. Since f satisfies condition $(*)$, there is a converging power series $\sum_{i\in\mathbb{N}} a_i(x-b)^i$ as given by Proposition 3.3.2. Since $x \mapsto \sum_{i\in\mathbb{N}} a_i(x-b)^i$ satisfies T_1-approximation on B_{as}, we have

$$\left|\sum_{i\geq 1} a_i(x-b)^i\right| < |b|$$

for all $x \in B_{\mathrm{as}}$. By the relation between the Gauss norm and the supremum norm on B_{as}, we then have

$$|a_i| \leq |b|^{1-i} \tag{3.4.1}$$

for all $i \geq 1$. Fix $b' \in D$ with $b'^r = b$. Since f is given by a power series on B, by composition we can develop $f_r = \sum_{k\geq 0} c_k(x-b')^k$ as a power series around b'. Using multinomial development, we find that for $k \geq 1$,

$$|c_k| \leq \max_{i\geq 1}\{|a_i| \cdot |b'|^{ri-k}\}.$$

Note that we could also get an explicit expression for c_k using the chain rule for Hasse derivatives.

Combining with equation (3.4.1) yields

$$|c_k| \leq |b'|^{r-k}.$$

In particular, we have $|c_k| \leq 1$ for $k \leq r$ and for any $x \in D$,

$$|f_r(x) - T^{<r}_{f_r,b'}(x)| = \left|\sum_{k\geq r} c_k (x-b')^k\right| \leq |x-b'|^r,$$

which concludes the proof. □

We now formulate a multidimensional version of Lemma 3.4.3. To do so we introduce the following notations. For a tuple $i = (i_1, \ldots, i_m) \in \mathbb{N}^m$ and $x = (x_1, \ldots, x_m) \in L^m$, recall that x^i is $\prod_{1\leq k\leq m} x_k^{i_k}$ and $|i| = i_1 + \cdots + i_m$. Also define $|x|_{\min,i}$ to be

$$\min_{1\leq j\leq m,\ i_j>0} \{|x_j|\}.$$

The idea is also to precompose with the r-th power to achieve the T_r-property on boxes. A naive approach to estimate the coefficients of the composite function, using the maximum modulus principle on the associated box, would lead to a bound for the $i \in \mathbb{N}^m$ coefficient of $|b|^r |b^i|^{-1}$. This however is not optimal and not enough for our needs. We improve it, working one variable at a time. The same idea (of composing with r-th power maps while controlling how many pieces are needed) is used in the real case in [Cluckers et al. 2020], but in our situation we get sharper control on the number of pieces in terms of r, resembling the sharper control of [Binyamini and Novikov 2019]. The difficulty for the corresponding control in [Cluckers et al. 2020] is that the cells in the o-minimal case have cell walls which also need to get small derivatives, and, composing with powers maps changes these cell walls. In our situation, there are no cell walls which can be considered as an advantage. On the other side, the absence of cell walls, and more generally of convexity arguments, has been a challenge in the non-Archimedean case that we have overcome by working with T_r-maps here and in [Cluckers et al. 2015].

Lemma 3.4.4. *Let f be a definable function on $X \subset \mathrm{VF}^m$ satisfying strong T_1-approximation. Then there is some $M > 0$ such that for L either a model of $\mathcal{T}$ which is a complete discretely valued field, or a local field with residue field cardinality at least M, the following holds for any integer $r > 0$.*

Let $b = (b_1, \ldots, b_m)$ be in L^m and suppose that $B = \prod_i b_i(1 + \mathcal{M}_L) \subseteq L^m$ is a subset of $X(L)$. For any $d = (d_1, \ldots, d_m)$ in L^m, write $p_{r,d}$ for the function $(x_1, \ldots, x_m) \mapsto (d_1 x_1^r, \ldots, d_m x_m^r)$. Then for any box $D \subseteq L^m$ such that $p_{r,d}(D) \subseteq B$, the function

$$f_{r,d} := f_L \circ p_{r,d}$$

satisfies T_r-approximation on D. Moreover, $f_{r,d}$ can be developed around any point $b' \in D$ as a power series converging on D_{as} with coefficients c_k satisfying

$$|c_k| \leq |b'|^r_{\min,k} |b'^k|^{-1} \quad \text{for all } k \in \mathbb{N}^m \backslash \{0\}.$$

Proof. Up to rescaling, we can assume $d_1 = \cdots = d_m = 1$. As in the proof of Lemma 3.4.3, we can fix $b \in B$, $b' \in D$ such that $b'^r = b$ and develop f as a power series $\sum_{i\in\mathbb{N}^m} a_i(x-b)^i$ that converges on B_{as}. Fix $\hat{x}_1 \in \hat{b}(1+\mathcal{M}_L)^{m-1}_{\mathrm{as}}$ and consider the function

$$f_{\hat{x}_1} : b_1(1+\mathcal{M}_L)_{\mathrm{as}} \to L, \quad x_1 \mapsto f(x_1, \hat{x}_1).$$

It is given by a power series $\sum_{i_1\in\mathbb{N}} a_{i_1}(\hat{x}_1)(x_1-b_1)^{i_1}$ around b_1 that converges on $b_1(1+\mathcal{M}_L)_{\mathrm{as}}$.

By the T_1-property for f on B_{as}, we have that for any $x_1 \in b_1(1+\mathcal{M}_L)_{\mathrm{as}}$,

$$|f_{\hat{x}_1}(x_1) - f_{\hat{x}_1}(b_1)| = |f(x_1,\hat{x}_1) - f(b_1,\hat{x}_1)| \le |x_1 - b_1| \le |b_1|.$$

Hence by the relation between the Gauss norm and the supremum norm on $b_1(1+\mathcal{M}_L)_{\mathrm{as}}$, for each $i_1 > 0$ we have

$$|a_{i_1}(\hat{x}_1)| \le |b_1|^{1-i_1}.$$

Now view $a_{i_1}(\hat{x}_1)$ as a function of $\hat{x}_1 \in \hat{b}(1+\mathcal{M}_L)^{m-1}_{\mathrm{as}}$, and by using again the relation between Gauss norm and sup norm, we find that for each $i \in \mathbb{N}$ such that $i_1 > 0$,

$$|a_i| \le |b_1|^{1-i_1} \cdot \left|\hat{b}^{(i_2,\ldots,i_m)}\right|^{-1} = |b_1||b^i|^{-1}.$$

By switching the numbering of the coordinates, we get that for each $i \in \mathbb{N}^m\backslash\{0\}$,

$$|a_i| \le |b|_{\min,i}|b^i|^{-1}.$$

The end of the proof is now similar to that of Lemma 3.4.3. Indeed, we develop $f_{r,d} = f \circ p_{r,d}$ into a power series around b', denoted by $\sum_{c_k\in\mathbb{N}^m} c_k(x-b')$. Then by multinomial development and using the bound for a_i we find that for $k \in \mathbb{N}^m\backslash\{0\}$,

$$|c_k| \le |b'|^r_{\min,k}|b'^k|^{-1}.$$

It is now a direct consequence of this bound that $|c_k| \le 1$ for $k \in \mathbb{N}^m\backslash\{0\}$ with $|k| < r$.

Now fix $x \in D$ and $k \in \mathbb{N}^m\backslash\{0\}$ with $|k| \ge r$. Choose some $\underline{r} \in \mathbb{N}^m$ such that $|\underline{r}| = r$ and $\underline{r}_j \le k_j$ for $j = 1,\ldots,m$. We have

$$\begin{aligned}
|c_k(x-b')^k| &\le |b'|^r_{\min,k}|b'^k|^{-1}|(x-b')^k| \\
&\le |b'|^r_{\min,k}|b'^k|^{-1}|(x-b')^{k-\underline{r}}||x-b'|^r \\
&\le |b'^{\underline{r}}||b'^k|^{-1}|(x-b')^{k-\underline{r}}||x-b'|^r \\
&\le |x-b'|^r.
\end{aligned}$$

Hence $f_{r,d}$ satisfies T_r-approximation on D. □

Proof of Theorem 3.1.4. First apply Theorem 3.4.2 to X to get a finite set I and a family $g = (g_{y,i})_{(y,i)\in Y\times I}$ of definable functions

$$g_{y,i} : P_{y,i} \to X_y$$

such that $P_{y,i} \subseteq \mathcal{O}^m_{\mathrm{VF}}$ and for each y, $(g_{y,i})_{i\in I}$ forms a strong T_1-parametrization of X_y.

By Proposition 3.3.2, we find $M \in \mathbb{N}$ such that for any $L \in \mathscr{C}_{\mathcal{O},M}$, any $y \in Y(L)$, any box $B \subseteq P_{y,i}(L)$ and any $b \in B$, there is a power series centered at b, converging on B_{as} and equal on B_{as} to g_{as}. Fix such an L and write q for q_L.

Observe that it is enough to prove the theorem for r prime to q. Indeed, a T_{r+1}-parametrization is also a T_r-parametrization. Hence, up to enlarging the constant, if r is not prime to q one can apply the theorem with $r+1$ to obtain a T_r-parametrization.

We fix an integer r prime to q and we partition $\mathbb{F}_q^\times$ into $\ell = \gcd(r, q-1)$ sets $A_1, \dots, A_\ell$ such that $x \mapsto x^r$ is a bijection from each A_i to $(\mathbb{F}_q^\times)^r$, the set of r-th powers in $\mathbb{F}_q^\times$. We choose representatives $\bar{d}_1, \dots, \bar{d}_\ell$ for cosets of $(\mathbb{F}_q^\times)^r$ and we fix lifts of them, denoted by $d_1, \dots, d_\ell \in \mathcal{O}_L$. For $x \in \mathcal{O}_L \backslash \{0\}$, we set $\xi(x) = d_i$ for i such that $\overline{\mathrm{ac}}(x) \in A_i$.

Now define for $j = (j_1, \dots, j_m) \in \{0, \dots, r-1\}^m$ the function

$$p_{r,j} : (\mathcal{O}_L \backslash \{0\})^m \to (\mathcal{O}_L \backslash \{0\})^m, \quad x = (x_1, \dots, x_m) \mapsto (t^{j_1} \xi(x_1) x_1^r, \dots, t^{j_m} \xi(x_m) x_m^r),$$

where t is our constant symbol for a uniformizer of $\mathcal{O}_L$.

Let $D_{y,i,j} = p_{r,j}^{-1}(P_{y,i}(L))$. By compactness and up to making M larger if necessary, we have that $P_{y,i}(L)$ is a cell around zero. By Hensel's lemma, the union over $j \in \{0, \dots, r-1\}^m$ of the sets $p_{r,j}(D_{y,i,j})$ is equal to $P_{y,i}(L)$. We claim that the family $(\bar{g}_{y,i,j} = g_{y,i} \circ p_{r,j})_{(y,i,j) \in Y(L) \times I \times \{0,\dots,r-1\}^m}$ is the desired T_r-parametrization of $X(L)$. Note that since we used in its definition the lifts d_i, it is an R_r-definable family, where R_r is a set of lifts of representatives of cosets of $(\mathbb{F}_q^\times)^r$. To lighten notations, let us skip for the rest of the proof the subscript (y, i, j). By Lemma 3.4.4 and up to making M larger if necessary, $\bar{g}$ satisfies T_r-approximation on each box contained in D. We show using T_1-approximation for g and ultrametric computations that $\bar{g}$ satisfies T_r-approximation on the whole D.

Fix $x, y \in D$. If x and y are in the same box contained in D, then we are done. Assume then that they are not.

Choose $v \in D$ such that $\overline{\mathrm{ac}}(v_i) = \overline{\mathrm{ac}}(y_i)$ and $|v_i| = |x_i|$, and in the case we moreover have $\overline{\mathrm{ac}}(x_i) = \overline{\mathrm{ac}}(y_i)$, set $v_i = x_i$. Such a v exists by Hensel's lemma and the fact that D is a cell around zero. Define $w \in D$ such that $w_i = v_i$ if $|v_i| = |y_i|$ and $w_i = y_i$ if $|v_i| \neq |y_i|$. We have that w and y lie in the same box contained in D. There are also $d, d', d'' \in \mathcal{O}_L^m$ as prescribed by $p_{r,j}$ such that $\bar{g}(x) = g(dx^r)$, $\bar{g}(w) = g(d'w^r)$ and $\bar{g}(y) = g(d''y^r)$.

We then have

$$\begin{aligned}
|\bar{g}(x) - T_{\bar{g},y}^{<r}(x)| &\leq \max\left\{|\bar{g}(x) - \bar{g}(w)|, |\bar{g}(w) - T_{\bar{g},y}^{<r}(w)|, |T_{\bar{g},y}^{<r}(w) - T_{\bar{g},y}^{<r}(x)|\right\} \\
&= \max\left\{|g(dx^r) - g(d'w^r)|, |\bar{g}(w) - T_{\bar{g},y}^{<r}(w)|, |T_{\bar{g},y}^{<r}(w) - T_{\bar{g},y}^{<r}(x)|\right\} \\
&\leq \max\left\{|dx^r - d'w^r|, |w - y|^r, |T_{\bar{g},y}^{<r}(w) - T_{\bar{g},y}^{<r}(x)|\right\} \\
&\leq \max\left\{|x - y|^r, |w - y|^r, |T_{\bar{g},y}^{<r}(w) - T_{\bar{g},y}^{<r}(x)|\right\} \\
&\leq \max\left\{|x - y|^r, |T_{\bar{g},y}^{<r}(w) - T_{\bar{g},y}^{<r}(x)|\right\} \\
&\leq |x - y|^r.
\end{aligned}$$

The first inequality is by the ultrametric triangular inequality, the second is by the global T_1-property for g and the T_r-property on boxes for $\bar{g}$. The third one is because for each i, we have $|d_i x_i^r - d_i' w^r| \leq |x_i - y_i|^r$. Indeed, there are three cases to consider. In one case, we have $x_i = w_i$ and $d_i = d_i'$, and then $d_i x_i^r - d_i' w^r = 0$. Or we have $|x_i| \neq |y_i|$. In that case, $|w_i| = |y_i|$ and $|d_i| = |d_1'| \leq 1$. Then by the ultrametric property we have $|x_i - y_i| = \max\{|x_i|, |y_i|\}$ and

$$|d_i x_i^r - d_i' w^r| = \max\{|d_i x_i^r|, |d_i' w_i^r|\} \leq \max\{|x_i|, |w_i|\}^r = \max\{|x_i|, |y_i|\}^r.$$

The last case is when $|x_i| = |y_i|$ and $\overline{\mathrm{ac}}(x_i) \neq \overline{\mathrm{ac}}(y_i)$. In that case,

$$|w_i| = |x_i|, \qquad \overline{\mathrm{ac}}(w_i) = \overline{\mathrm{ac}}(y_i), \qquad |d_i| = |d_i'| \leq 1.$$

We then have $|x_i - y_i| = |x_i|$ and by the choice made in the definition of $p_{r,j}$,

$$\overline{\mathrm{ac}}(d_i x^r) \neq \overline{\mathrm{ac}}(d_i' w^r),$$

whence $|d_i x_i^r - d_i' w^r| = |d_i x^r| \leq |x_i|^r = |x_i - y_i|^r$.

The fourth inequality holds because by definition of w, either $w_i = y_i$, or $w_i = x_i$, or $|w_i| = |x_i| = |y_i|$ and $\overline{\mathrm{ac}}(x_i) \neq \overline{\mathrm{ac}}(w_i) = \overline{\mathrm{ac}}(y_i)$. In those three cases, we have $|w_i - y_i| \leq |x_i - y_i|$.

To conclude the proof, it remains to prove the last inequality

$$|T_{\bar{g},y}^{<r}(w) - T_{\bar{g},y}^{<r}(x)| \leq |x - y|^r.$$

Suppose $T_{\bar{g},y}^{<r}(x) = \sum_{k \in \mathbb{N}^m, |k| < r} c_k (x - y)^k$. For $A \subseteq \mathbb{N}^m$, introduce the notation

$$T_{\bar{g},y}^{<r,A}(x) = \sum_{k \in \mathbb{N}^m,\ |k| < r, k \in A} c_k (x - y)^k.$$

Then set

$$A' = \{k = (k_1, \ldots, k_m) \in \mathbb{N}^m \mid k_i = 0 \text{ if } |y_i| \leq |x_i - y_i|\},$$

and let A be its complement. The condition can be rephrased by writing that $k_i = 0$ if $w_i \neq x_i$. In particular, for $k \in A'$ we have $(x - y)^k = (w - y)^k$, and hence $T_{\bar{g},y}^{<r,A'}(x) = T_{\bar{g},y}^{<r,A'}(w)$.

Thus it remains to show that

$$|T_{\bar{g},y}^{<r,A}(w) - T_{\bar{g},y}^{<r,A}(x)| \leq |x - y|^r.$$

We claim that

$$|T_{\bar{g},y}^{<r,A}(x)| \leq |x - y|^r \quad \text{and} \quad |T_{\bar{g},y}^{<r,A}(w)| \leq |x - y|^r,$$

which implies the preceding inequality.

Since for each i, $|w_i - y_i| \leq |x_i - y_i|$, it is enough to prove that for each $k \in A$ such that $0 < |k| < r$,

$$|c_k (x - y)^k| \leq |x - y|^r.$$

From the definition of A, there is some i_0 such that $k_{i_0} > 0$ and $|y_{i_0}| \leq |x_{i_0} - y_{i_0}|$. Suppose to lighten the notation that $i_0 = 1$. Set $\underline{r} = (\underline{r}_1, \ldots, \underline{r}_m)$ with $\underline{r}_i = k_i$ for $i > 1$ and $\underline{r}_i = r - |k| + k_1 \geq 1$.

Recall the bound for $|c_k|$ obtained from Lemma 3.4.4. We now compute, using this bound and the definition of $\underline{r}$,

$$\begin{aligned}|c_k(x-y)^k| &\leq |y|^r_{\min,k}\,|y^k|^{-1}\,|(x-y)^k| \\ &\leq |y^r||y^k|^{-1}\,|(x-y)|^{|k|} \\ &\leq |y^{r-k}||(x-y)|^{|k|} \\ &= |y_1|^{r-|k|}\,|(x-y)|^{|k|} \\ &\leq |x_1-y_1|^{r-|k|}\,|(x-y)|^{|k|} \\ &\leq |(x-y)|^{|r|}.\end{aligned}$$

This finishes the proof of the theorem. □

Proof of Theorem 3.1.5. The proof is similar to that of Theorem 3.1.4 above, using Theorem 3.4.2 and then precomposition by power functions. One just needs to delete the application of compactness, and, instead of using the map ξ which chooses and exploits the lifts of cosets of r-th powers in the residue field, one uses parameters from R_r to paste pieces together. (A factor b_r comes in because in this general case the pasting is rougher, since in the residue field, the number of cosets of the r-th powers fails to equal the number of solutions of $x^r = 1$ in general.) The rest of the proof is completely similar. □

4. Points of bounded degree in $\mathbb{F}_q[t]$

4.1. ***A counting theorem.*** The goal of this section is to prove the following theorem, of which Theorem A is a particular case. Recall from the introduction that, for q a prime power and n a positive integer, $\mathbb{F}_q[t]_n$ is the set of polynomials with coefficients in $\mathbb{F}_q$ and degree (strictly) less than n, and, for an affine variety X defined over a subring of $\mathbb{F}_q((t))$, $X(\mathbb{F}_q[t]_n)$ denotes the subset of $X(\mathbb{F}_q((t)))$ consisting of points whose coordinates lie in $\mathbb{F}_q[t]_n$. Also, for a subset A of $\mathbb{F}_q((t))^m$, write A_n for the subset of A consisting of points whose coordinates lie in $\mathbb{F}_q[t]_n$.

For an affine (reduced) variety $X \subset \mathbb{A}^m_R$ with R an integral domain contained in an algebraically closed field K, we define the degree of X as the degree of the closure of X_K in $\mathbb{P}^m_K$. For example, if X is a hypersurface given by one (reduced) equation f, then the degree of X equals the (total) degree of f.

Theorem 4.1.1. *Let d, m and δ be positive integers. Then there exist real numbers $C = C(d, m, \delta)$ and $N = N(d, m, \delta)$ such that for each prime $p > N$, each power $q = p^\alpha$ with $\alpha > 0$ an integer, each integer $n > 0$ and each irreducible variety $X \subseteq \mathbb{A}^m_{\mathbb{F}_q((t))}$ of degree δ and dimension d, one has*

$$\#X(\mathbb{F}_q[t])_n \leq Cn^2 q^{n(d-1)+\lceil n/\delta \rceil}.$$

We first give a bound for a so-called naive degree. Define the naive degree of a variety $X \subset \mathbb{A}^m_R$ with R an integral domain as the minimum, taken over all tuples of (nonzero) polynomials $f = (f_1, \ldots, f_s)$ over R with $X(K) = \{x \in K^m \mid f(x) = 0\}$, of the product of the degrees of the f_i.

Lemma 4.1.2. *Let d, m, and δ be positive integers. Then there exist numbers $C = C(d, m, \delta)$ and $N = N(d, m, \delta)$ such that for each prime $p > N$, each power $q = p^\alpha$ with $\alpha > 0$ an integer, and each*

geometrically irreducible variety $X \subseteq \mathbb{A}^m_{\mathbb{F}_q((t))}$ of degree δ and dimension d, one has that the naive degree of X is bounded by C.

Proof. From the theory of Chow forms (see [Samuel 1955] or [Catanese 1992]), a variety $X \subseteq \mathbb{A}^m_{\mathbb{F}_q((t))}$ of degree δ and dimension d is determined set-theoretically by a hypersurface of degree δ in the Grasmanniann of $G(m-d-1, m)$ of $(m-d-1)$-dimensional vector subspaces of the m-dimensional space. As explained for example in [Catanese 1992], one can construct from such a hypersurface a system of $m(d+1)$ equations of degrees at most δ such that their zero sets coincide set-theoretically with X. Hence the naive degree of X is bounded by $\delta m(d+1)$. □

The following trivial bound for points of bounded height is typical.

Lemma 4.1.3. *Let d, m and δ be positive integers. Then there exist real numbers $C = C(d, m, \delta)$ and $N = N(d, m, \delta)$ such that for each prime $p > N$, each power $q = p^\alpha$ with $\alpha > 0$ an integer, each integer $n > 0$ and each irreducible variety $X \subseteq \mathbb{A}^m_{\mathbb{F}_q((t))}$ of degree δ and dimension d, one has*

$$\#X(\mathbb{F}_q[t])_n \leq Cq^{nd}.$$

Proof. The lemma follows easily from Noether's normalization lemma and Lemma 4.1.2. □

Let us first reduce the statement of Theorem 4.1.1 to the case of planar curves, similarly to [Pila 1995]. In this section, definable means definable in the language $\mathcal{L}_{\mathrm{DP}}$ of Setting 3.1.1 and with $\mathcal{O} = \mathbb{Z}$.

Reduction of Theorem 4.1.1 to the case $m = 2$ and $d = 1$.. Fix positive integers d, m, δ. By Lemma 4.1.2, irreducible varieties in $\mathbb{A}^m$ of dimension d and of degree δ form a definable family of sets, say, with parameter z in a definable (and Zariski-constructible) set Z; write X_z for the variety in $\mathbb{A}^m$ corresponding to the parameter $z \in Z$. Assume first that $m > 2$ and $d = 1$. Consider the family of linear projections $p_{a,b} : \mathbb{A}^m \to \mathbb{A}^2$ written in coordinates $x = \sum a_i x_i$ and $y = \sum b_i y_i$ and with parameters $(a, b) \in \mathbb{A}^{2m}$. Then, for each $z \in Z$, there is a nonempty Zariski open subset of parameters $O_z \subseteq \mathbb{A}^{2m}$ such that $p_{a,b}$ is surjective and the varieties X_z and $p_{a,b}(X_z)$ have the same degree δ (and are both irreducible of dimension 1) for all $(a, b) \in O_Z$. Clearly the open sets O_z form a definable family of sets with parameter $z \in Z$.

Now suppose that the prime p is large enough and that $q = p^\alpha$ for some α. Since the complement of O_z is of dimension less than $2m$ by Lemma 4.1.3, and since the O_z form a definable family, we can find for each $z \in Z(\mathbb{F}_q((t)))$ a point (a^0, b^0) in $O_z(\mathbb{F}_q[t])_1$ (hence, so to say, a tuple of polynomials in t over $\mathbb{F}_q$ and of degree 0). Hence, p_{a^0,b^0} maps points in $\mathbb{F}_q[t]^m_n$ to points in $\mathbb{F}_q[t]^2_n$. Furthermore, the fibers of p_{a^0,b^0} on X_z are finite, uniformly in z, say, bounded by C. We thus have that for each large enough p, each z in $Z(\mathbb{F}_q((t)))$, and each $n > 0$, that

$$\#X_z(\mathbb{F}_q[t])_n \leq C\#p(X_z)(\mathbb{F}_q[t])_n.$$

Hence the result for $d = 1$ and general $m > 1$ follows from the case $d = 1$ and $m = 2$.

Assume now that $m \geq 2$ and $d > 1$. By a projection argument as above, we can assume that $d = m - 1$. Consider the family of hyperplanes $H = H_{\alpha,b}$ with equation $\sum \alpha_i x_i = b$ and parameters α and b. Then

for each $z \in Z$ there is a nonempty Zariski open subset O_z of $\mathbb{A}^{m+1}$ such that if (α, b) lies in O_z, then $X_z \cap H_{\alpha,b}$ is irreducible, of degree δ and dimension d. Hence, similarly as above, for large enough primes p and with $q = p^\alpha$, we can find for each z in $Z(\mathbb{F}_q((t)))$ a point (α^0, b^0) in $O_z(\mathbb{F}_q[t])_1$. Now consider the family of hyperplanes H_b of equations $\sum \alpha_i^0 x_i = b$ with parameter b running over $\mathbb{F}_q((t))$. Since (α^0, b^0) belongs to $O_z(\mathbb{F}_q[t])_1$, and by construction, there are at most finitely many values for b such that $(\alpha^0, b) \notin O_z(\mathbb{F}_q((t)))$, say, $b_1, \ldots, b_k$. In any case we can assume that $X_z \cap H_{b_j}$ is of dimension at most $m-1$ for each j, and hence that

$$\#(X_z \cap H_{b_j}) \leq C q^{m-1}$$

for some C which is independent of q and n, by Lemma 4.1.3. To treat the remaining part, we apply the induction hypothesis to $X'_z = (X_z \cap H_b)$ for b outside $\{b_1, \ldots, b_k\}$, and we take the sum of the bounds over all values of b in $\mathbb{F}_q[t]_n$. □

4.2. ***Determinant lemma.*** We fix the following notation for the rest of the paper. For $\alpha = (\alpha_1, \ldots, \alpha_m)$ in $\mathbb{N}^m$, set $|\alpha| = \alpha_1 + \cdots + \alpha_m$. Set also

$$\Lambda_m(k) := \{\alpha \in \mathbb{N}^m \mid |\alpha| = k\}, \qquad \Delta_m(k) := \{\alpha \in \mathbb{N}^m \mid |\alpha| \leq k\},$$
$$L_m(k) := \#\Lambda_m(k), \qquad D_m(k) = \#\Delta_m(k).$$

Lemma 4.2.1 [Cluckers et al. 2015, Lemma 3.3.1]. *Let K be a discretely valued henselian field. Fix $\mu, r \in \mathbb{N}$, and U an open subset of K^m contained in a box that is a product of m closed balls of valuative radius ρ. Fix $x_1, \ldots, x_\mu \in U$, and functions $\psi_1, \ldots, \psi_\mu : U \to K$. Assume that*

- *the integer r satisfies*
$$D_m(r-1) \leq \mu < D_m(r);$$
- *the functions $\psi_1, \ldots, \psi_\mu$ satisfy T_r on U.*

Then

$$\mathrm{ord}_t(\det(\psi_i(x_j))) \geq \rho e,$$

where $e = \sum_{i=0}^{r-1} i L_m(i) + r(\mu - D_m(r-1))$.

4.3. ***Hilbert functions.*** Fix a field K. For s a positive integer, denote $K[x_0, \ldots, x_n]_s$ the space of homogenous polynomials of degree s. Let I be a homogenous ideal of $K[x_0, \ldots, x_n]$, associated to an irreducible variety of dimension d and degree δ of $\mathbb{P}^n_K$. Let $I_s = I \cap K[x_0, \ldots, x_n]_s$ and let $\mathrm{HF}_I(s) = \dim_K K[x_0, \ldots, x_n]_s / I_s$ be the (projective) Hilbert function of I. The Hilbert polynomial HP_I of I is a polynomial such that for s large enough, $\mathrm{HP}_I(s) = \mathrm{HF}_I(s)$. It is a polynomial of degree d and leading coefficient $\delta/d!$.

Fix some monomial ordering in the sense of [Cox et al. 2015]. Denote by $\mathrm{LT}(I)$ the ideal generated by leading terms of elements of I. By [Cox et al. 2015], the Hilbert functions of I and $\mathrm{LT}(I)$ are equal. It follows that

$$\mathrm{HF}_I(s) = \#\{\alpha \in \Lambda_{n+1}(s) \mid x^\alpha \notin \mathrm{LT}(I)\}.$$

Define also for $i = 0, \dots, n$,

$$\sigma_{I,i}(s) = \sum_{\alpha \in \Lambda_{n+1}(s),\, x^\alpha \notin \mathrm{LT}(I)} \alpha_i. \tag{4.3.1}$$

Hence, we have $s\,\mathrm{HF}_I(s) = \sum_{i=0}^n \sigma_{I,i}(s)$. The function $\sigma_{I,i}$ is also equal to a polynomial function of degree at most $d+1$, for s large enough. It follows that there exist nonnegative real numbers $a_{I,i}$ such that

$$\frac{\sigma_{I,i}(s)}{s\,\mathrm{HP}_I(s)} = a_{I,i} + O(1/s) \tag{4.3.2}$$

when s goes to $+\infty$.

Remark 4.3.1. The s chosen large enough so that $\mathrm{HF}_I(s)$ is a polynomial and the implicit constant in (4.3.2) depend on I. However, since $\mathrm{HF}_I(s) = \mathrm{HF}_{\mathrm{LT}(I)}$, they in fact only depend on $\mathrm{LT}(I)$. Since they are obtained in a pure combinatorial way, they do not depend on the field K. If we let I vary among ideals generated by a polynomial of degree at most d, then only finitely many different $\mathrm{LT}(I)$ appear. So the previous constants can be chosen uniformly over the whole family of such ideals I.

We will also use the following lemma of Salberger [2007], which is the reason why we will use a projective embedding in the proof of Theorem 4.1.1.

Lemma 4.3.2 [Salberger 2007]. *Let X be a closed equidimensional subscheme of dimension d of $\mathbb{P}^m_K$. Assume that no irreducible component of X is contained in the hyperplane at infinity defined by $x_0 = 0$. Let $<$ be the monomial ordering defined by $\alpha \le \beta$ if $|\alpha| < |\beta|$ or $|\alpha| = |\beta|$ and for some i, $\alpha_i > \beta_i$ and $\alpha_j = \beta_j$ for $j < i$. Then*

$$a_{I,1} + \cdots + a_{I,m} \le \frac{d}{d+1}.$$

4.4. ***Proof of Theorem 4.1.1 for $m = 2$ and $d = 1$.*** Fix a positive integer δ. Clearly all irreducible curves in $\mathbb{A}^2$ of degree δ form a definable family, say, with parameter z in a definable (and Zariski-constructible) set Z; write X_z for the curve in $\mathbb{A}^2$ corresponding to the parameter $z \in Z$.

Apply Theorem 3.1.4 to the definable family of the definable sets X_z. It gives some constant C and, for some M, all local fields K in $\mathcal{B}_{\mathbb{Z},M}$ and all integers $r > 0$ prime to q_K, a T_r-parametrization of $X_z(\mathcal{O}_K)$ with Cr many pieces. Fix such a K and a parameter $z \in Z(K)$ corresponding to an irreducible curve $X_z \subset \mathbb{A}^2_K$ of degree δ.

Consider the map

$$\iota : \mathbb{A}^2_K \to \mathbb{A}^3_K, \quad (x, y) \mapsto (1, x, y)$$

and the corresponding embedding

$$\underline{\iota} : \mathbb{A}^2_K \hookrightarrow \mathbb{P}^2_K, \quad (x, y) \mapsto [1 : x : y].$$

Denote by I_z the homogenous ideal associated to the closure of $\underline{\iota}(X_z)$.

Fix some positive integer s, set

$$M_z(s) = \{\alpha \in \Lambda_3(s) \mid x^\alpha \notin \mathrm{LT}(I_z)\}, \quad \mu = \#M_z(s) \quad \text{and} \quad e = \tfrac{1}{2}\mu(\mu-1).$$

Now consider the given T_r-parametrization of $X_z(\mathcal{O}_K)$ with $r = \mu$ and work on one of the $C\mu$ pieces $U_z \subseteq \mathcal{O}_K$ with function $g_z : U_z \to X(\mathcal{O}_K)$ satisfying T_μ on U_z.

Fix a closed ball $B_\beta \subseteq \mathcal{O}_K$ of valuative radius β. Fix some points $y_1, \dots, y_\mu$ in $(g(B_\beta \cap U))_n$ and consider the determinant

$$\Delta = \det(\iota(y_i)^\alpha)_{1\le i\le\mu,\alpha\in M_z(s)}.$$

Since the composition of functions satisfying T_μ also satisfies T_μ, we can apply Lemma 4.2.1 with $m = 1$, $r = \mu$ to get that

$$\operatorname{ord}_t \Delta \ge \beta e.$$

On the other hand, since the points y_i are of degree less than n as polynomials in t over $\mathbb{F}_{q_K}$, we also have

$$\deg \Delta \le (n-1)(\sigma_1 + \sigma_2),$$

where σ_1, σ_2 are defined by equation (4.3.1). Hence, if Δ is not zero, then

$$\operatorname{ord}_t \Delta \le (n-1)(\sigma_1 + \sigma_2).$$

It follows that $\Delta = 0$ whenever

$$\beta e > (n-1)(\sigma_1 + \sigma_2). \tag{4.4.1}$$

When such an inequality holds, the matrix $A = (y_i^\alpha)$ is of rank less than μ. Fix a minor of maximal rank $B = (y_i^\alpha)_{i\in I,\alpha\in J}$ and some $\alpha_0 \in M_z(s)\backslash J$. Then the polynomial

$$f(x, y) = \det\begin{pmatrix} y_i^\alpha \\ (1, x, y)^\alpha \end{pmatrix}_{i\in I,\,\alpha\in J\cup\{\alpha_0\}}$$

is of total degree at most s and nonzero, since the coefficient of $(1, x, y)^{\alpha_0}$ is $\det(B)$. Moreover, it vanishes at all points in $g(B_\beta \cap U)_n$ but does not vanish on the whole X_z, since its exponents lie in $M_z(s)$ and X_z is irreducible. Hence, by Bézout's theorem, there are at most $s\delta$ points in $(g(B_\beta \cap U))_n$.

We now show how to choose s and β in terms of n such that inequality (4.4.1) holds. Recall that $\mu = \#M_z(s) = \mathrm{HF}_{I_z}(s)$. By properties of Hilbert polynomials and equation (4.3.2), we have

$$\mu = \delta s + O(1) \tag{4.4.2}$$

and

$$\frac{\sigma_i}{\mu} = a_i s + O(1).$$

Here and below, the notation $O(1)$ refers to $s \to +\infty$, and by Remark 4.3.1, the implicit constant is independent of z and q_K. Combining those two equations, we get

$$\sigma_i = a_i \delta s^2 + O(s) \quad \text{and} \quad e = \tfrac{1}{2}\delta^2 s^2 + O(s),$$

and finally, by applying Lemma 4.3.2,

$$\frac{\sigma_1 + \sigma_2}{e} \le \frac{1}{\delta} + O(s^{-1}).$$

Hence there is some s_0 and $C_0 > 0$ such that for every $s \geq s_0$,

$$\frac{\sigma_1 + \sigma_2}{e} \leq \frac{1}{\delta} + C_0 s^{-1}.$$

Recall that the coefficients of Hilbert polynomials can be bounded in terms of the degree of the curve and that the characteristic is assumed to be large. Hence s_0 and C_0 depend only on the degree δ of the curve X_z.

If follows that for

$$s = \lceil \max\{s_0, 2C_0(n-1)\} \rceil, \tag{4.4.3}$$

we have

$$(n-1)\frac{\sigma_1 + \sigma_2}{e} \leq \left\lceil \frac{n}{\delta} \right\rceil.$$

We can thus set $\beta = \lceil n/\delta \rceil$ to satisfy inequality (4.4.1). It follows from the preceding discussion that there are at most $s\delta$ points in $g(B_\beta \cap U)_n$. From equation (4.4.2), we have $\mu \leq \delta s + C_1$, for some constant C_1, and from (4.4.3), that $s \leq C_2 n$ for some constant C_2, with C_i independent of n. Since we need q^β closed balls of valuative radius β to cover $\mathbb{F}_q[\![t]\!] = \mathcal{O}_K$, and since we have a T_μ-parametrization of $X(\mathbb{F}_q[\![t]\!])$ involving $C\mu$ pieces, we find that (after enlarging C) there are at most

$$Cn^2 q^{\lceil n/\delta \rceil}$$

points in $X(\mathbb{F}_q[t])_n$. □

Remark 4.4.1. In the preprint [Bhargava et al. 2017], Sedunova's result [2017] is used to bound the 2-torsion of class groups of function fields over finite fields; see their Theorem 7.1. One can use instead our Theorem 4.1.1 in the special case of Theorem A to obtain a uniform version of their result. We thank Paul Nelson for directing us to the reference [Bhargava et al. 2017].

5. Uniform non-Archimedean Pila–Wilkie counting theorem

In this section we provide uniform versions in the p-adic fields for large p and also in the fields $\mathbb{F}_q((t))$ of large characteristic of several of the main counting results of [Cluckers et al. 2015] (on rational points on p-adic subanalytic sets). To achieve this we use the uniform parametrization result of Theorem 3.1.4. Furthermore, Proposition 5.1.4 is new in all senses, and is a (uniform) non-Archimedean variant of recent results of [Cluckers et al. 2020; Binyamini and Novikov 2019]; it should be put in contrast with Proposition 4.1.3 of [Cluckers et al. 2015].

5.1. *Hypersurface coverings.*

We begin by fixing some terminology.

Consider the language $\mathcal{L} = \mathcal{L}_{\mathrm{DP}}^{\mathrm{an}}$ as described in Setting 3.1.1. From now on we only consider definable sets which are subsets of the Cartesian powers of the valued field sort (sometimes in a concrete $\mathcal{L}$-structure, and sometimes for the theory $\mathcal{T}$).

Definition 5.1.1. Let K be an $\mathcal{L}$-structure. An $\mathcal{L}(K)$-definable set $X \subset K^n$ is said to be of dimension d at $x \in X$ if for every small enough box containing x, $X \cap B$ is of dimension d. An $\mathcal{L}(K)$-definable set $X \subset K^n$ is said to be of pure dimension d if it is of dimension d at all points x in $X(K)$.

For an $\mathcal{L}(K)$-definable set $X \subset K^n$, define the algebraic part X^{alg} of X to be the union of all quantifier-free $\mathcal{L}_{\mathrm{DP}}(K)$-definable sets of pure positive dimension and contained in X. Note that the set X^{alg} is in general neither semialgebraic nor subanalytic.

By subanalytic we mean from now on $\mathcal{L}$-definable, or $\mathcal{L}(K)$-definable if we are in a fixed $\mathcal{L}$-structure, and we speak about definable families in the sense explained just below Notation 3.1.3. Likewise, by semialgebraic we mean definable in the language $\mathcal{L}_{\mathrm{DP}}$, or $\mathcal{L}_{\mathrm{DP}}(K)$-definable if we are in a fixed structure (see Setting 3.1.1). Write $\mathcal{T}$ for $\mathcal{T}_{\mathrm{DP}}^{\mathrm{an}}$.

Remark 5.1.2. Observe that the definition of the algebraic part is insensitive to having or not having algebraic Skolem functions on the residue field. Indeed, its definition is local and allows parameters from the structure.

If $x \in \mathbb{Z}$, set $H(x) = |x|$, the absolute value of x. If $x = (x_1, \ldots, x_n) \in \mathbb{Z}^n$, set $H(x) = \max_i\{H(x_i)\}$. If L is a local field of characteristic zero, $B \geq 1$ and $X \subseteq L^n$, we set

$$X(\mathbb{Z}, B) = \{x \in X \cap \mathbb{Z} \mid H(x) \leq B\}.$$

If $x \in \mathbb{F}_q[t]$, we set

$$H(x) = q^{\deg_t(x)},$$

where $\deg_t(x)$ is the degree in t of the polynomial x over $\mathbb{F}_q$. For $x = (x_1, \ldots, x_n) \in (\mathbb{F}_q[t])^n$, put $H(x) = \max_i\{H(x_i)\}$. We now set for $X \subseteq \mathbb{F}_q[\![t]\!]$ and $B \geq 1$

$$X(\mathbb{F}_q[t], B) = \{x \in X \cap \mathbb{F}_q[t] \mid H(x) \leq B\}.$$

Recall the notation at the beginning of Section 4.1. For all integers d, n, m, set $\mu = D_n(d)$ and let r be the smallest integer such that $D_m(r-1) \leq \mu < D_m(r)$. Then set $V = \sum_{k=0}^{d} kL_n(k)$ and $e = \sum_{k=1}^{r-1} kL_m(k) + r(\mu - D_m(r-1))$.

The following result refines Lemma 4.1.2 of [Cluckers et al. 2015] and has a similar proof.

Lemma 5.1.3. *For all integers d, n, m with $m < n$, consider the integers r, V, e as defined above. Fix a local field L, a subset $U \subseteq \mathcal{O}_L^m$, an integer H and maps $\psi = (\psi_1, \ldots, \psi_n) : U \to \mathcal{O}_L^n$ that satisfy T_r-approximation. Then if L is of characteristic zero, the set $\psi(U)(\mathbb{Z}, H)$ is contained in at most*

$$q^m(\mu!)^{m/e}H^{mV/e}$$

hypersurfaces of degree at most d. If L is of positive characteristic, the set $\psi(U)(\mathbb{F}_q[t], H)$ is contained in at most

$$q^m H^{mV/e}$$

hypersurfaces of degree at most d. Moreover, when d goes to infinity, mV/e goes to 0.

Proof. We use the notation introduced at the beginning of Section 4.1. Under the hypothesis of the lemma, fix a closed box $B \subseteq \mathcal{O}_L^m$ of valuative radius α. Then fix points $P_1, \ldots, P_\mu \in \psi(B \cap U)(\mathbb{Z}, H)$ (or $\psi(B \cap U)(\mathbb{F}_q[t], H)$) and consider $x_i \in B \cap U$ such that $\psi(x_i) = P_i$. Consider the determinant $\Delta = \det((\psi(x_i)^j)_{1 \leq i \leq \mu,\, j \in \Delta_n(d)}$. Since ψ satisfies T_r-approximation, Lemma 4.2.1 gives $\mathrm{ord}(\Delta) \geq \alpha e$.

In the positive characteristic case, since the P_i are in $\mathbb{F}_q[t]$ of degree less than or equal to $\log_q(H)$, if $\Delta \neq 0$, then $\mathrm{ord}(\Delta) \leq \log_q(H)V$. Hence if $\alpha > \log_q(H)V/e$, then $\Delta = 0$.

In the characteristic zero case, since the P_i are in $\mathbb{Z}$ of height at most H, it follows that $\Delta \in \mathbb{Z}$ is of (Archimedean) absolute value at most $\mu! H^V$. If $\Delta \neq 0$, this implies that $\mathrm{ord}(\Delta) \leq \log_q(\mu! H^V)$. Hence if $\alpha > \log_q(\mu! H^V)/e$, then $\Delta = 0$.

We now assume that α is chosen such that $\Delta = 0$. As in the Bombieri–Pila case, by considering minors of maximal rank, we can produce a hypersurface D of degree d such that all the P_i are contained in D. See the proof of Theorem 4.1.1 for details.

Since we need $q^{m\alpha}$ boxes of radius α to cover $\mathcal{O}_L^m$, in the characteristic zero case, we find that we can cover $\psi(U)(\mathbb{Z}, H)$ by $q^m \mu!^{m/e} H^{mV/e}$ hypersurfaces of degree d. In the positive characteristic case, we can cover $\psi(U)(\mathbb{F}_q[t], H)$ by $q^m H^{mV/e}$ hypersurfaces of degree at most d.

By an explicit computation (see [Pila 2004, p. 212]), we get $e \sim_d C_1(m,n) d^{n+n/m}$ and $V \sim_d C_2(m,n) d^{n+1}$, the equivalences being for $d \to +\infty$. Since $m < n$, mV/e goes to zero as $d \to +\infty$. □

Proposition 5.1.4. *Let integers $m \geq 0$ and $n > m$ be given. Let $X = (X_y)_{y \in Y} \subseteq (\mathrm{VF}^n)_{y \in Y}$ be an $\mathcal{L}$-definable family of subanalytic sets with X_y of dimension m in each model K of $\mathcal{T}$ and each y in $Y(K)$. Then there are a constant $C(X)$ depending only on X, a constant $C'(n,m)$ depending only on n and m, and an integer $N = N(X)$ such that for each $H \geq 2$ and each local field $L \in \mathscr{C}_{\mathcal{O},N}$, the following holds.*

For $y \in Y(L)$ and $H \geq 2$, the set $X_y(L)(\mathbb{Z}, H)$ (or $X_y(L)(\mathbb{F}_{q_L}[t], H)$ for the positive characteristic case) is covered by at most

$$C(X) q_L^m \log(H)^\alpha$$

hypersurfaces of degree at most $C'(n,m)\log(H)^{m/(n-m)}$.

Moreover, we have $\alpha = nm/((m-1)(n-m))$ if $m > 1$ and $\alpha = n/(n-1)$ if $m = 1$.

Proof. We work inductively on m. The case $m = 0$ is clear, as the cardinality of the fibers is then uniformly bounded in y. Assume now $1 \leq m$. Apply the parametrization Theorem 3.1.4 to the definable family X.

We keep the notation from the proof of Lemma 5.1.3. Choose d as a function of H such that $H^{mV/e}$ is bounded (say by 2). From the computations at the end of the proof of Lemma 5.1.3, we can choose $d \sim_H C'(m,n)\log(H)^{m/(n-m)}$.

We have $\mu \sim_H C_3(n,m) d^n$, and since r is the smallest integer such that $D_m(r-1) \leq \mu < D_m(r)$, we have that if $m > 1$, then $r = O_H(\mu^{1/(m-1)})$ and if $m = 1$, then $r = \mu$. From Theorem 3.1.4, we find a T_r-parametrization of X involving $C(X) r^m$ pieces. From Lemma 5.1.3, the points of height at most H on one of the pieces are included in at most $q_L^m (\mu!)^{m/e} H^{mV/e}$ (if $L \in \mathcal{A}_\mathcal{O}$) or $q_L^m H^{mV/e}$ (if $L \in \mathcal{B}_\mathcal{O}$) hypersurfaces of degree at most d. From the Stirling formula, we see that $(\mu!)^{m/e}$ is bounded. Hence overall, up to enlarging $C(X)$, we find that $X_y(L)(\mathbb{Z}, H)$ or $X_y(L)(\mathbb{F}_{q_L}[t], H)$ is contained in

$$C(X) q_L^m \log(H)^\alpha$$

hypersurfaces of degree at most $C'(n,m)\log(H)^{m/(n-m)}$, with $\alpha = nm/((m-1)(n-m))$ if $m > 1$ and $\alpha = n/(n-1)$ if $m = 1$. □

5.2. ***Blocks.*** In this final section, we give uniform versions of results of [Cluckers et al. 2015, Section 4.2] for local fields of large residue characteristic, in particular of Theorems 4.2.3 and 4.2.4 of [Cluckers et al. 2015]. We thus obtain analogs of Pila–Wilkie counting results, uniformly for local fields of large enough positive characteristic. We leave proofs, which are analogous to the ones for Theorems 4.2.3 and 4.2.4 of [Cluckers et al. 2015], to the reader.

Definition 5.2.1. A subset $W \subset K^m$, with K an $\mathcal{L}$-structure, is called a block if it is either a singleton or a smooth subanalytic set of pure dimension $d > 0$ contained in a smooth semialgebraic set of pure dimension d.

A family of blocks $W \subseteq \mathrm{VF}^{m+s}$, with parameters running over VF^s, is a subanalytic set W such that there exists an integer $s' \geq 0$ and a semialgebraic set $W' \subseteq \mathrm{VF}^{m+s'}$ such that for each model K of $\mathcal{T}$, for each $y \in K^s$ there is a $y' \in K^{s'}$ such that both $W_y(K)$ and $W'_{y'}(K)$ are smooth of the same pure dimension and such that $W_y(K) \subseteq W'_{y'}(K)$.

Note that if W is a block of positive dimension, then $W = W^{\mathrm{alg}}$.

Note that our notion of family of blocks, which corresponds to the one in [Chambert-Loir and Loeser 2017], is a strengthening of the one in [Cluckers et al. 2015] which solely ask that a family of blocks W is such that W_y is a block for each $y \in Y$. However, all the results in Section 4.2 of [Cluckers et al. 2015] hold with this strengthened definition.

Let L be in $\mathcal{A}_{\mathcal{O}}$ and let $k > 0$ be an integer. We define the k-height of $x \in L$ as

$$H_k(x) = \min_a \Big\{ H(a) \;\Big|\; a = (a_1, \ldots, a_k) \in \mathbb{Z}^k, \sum_{i=0}^{k} a_i x^i = 0, a \neq 0 \Big\}$$

and for $x = (x_1, \ldots, x_n) \in L^n$, $H_k(x) = \max_i \{H(x_i)\}$.

Let $L \in \mathcal{B}_{\mathcal{O}}$ and $k > 0$ be an integer. We define the k-height of $x \in L$ as

$$H_k(x) = \min_a \Big\{ H(a) \;\Big|\; a = (a_1, \ldots, a_k) \in \mathbb{F}_{q_L}[t]^k, \sum_{i=0}^{k} a_i x^i = 0, a \neq 0 \Big\}$$

and for $x = (x_1, \ldots, x_n) \in L^n$, $H_k(x) = \max_i \{H(x_i)\}$.

If $X \subseteq L^n$, we set

$$X(k, H) = \{x \in X \mid H_k(x) \leq H\}.$$

The following result is a generalized and uniform version of Theorems 4.2.3 and 4.2.4 of [Cluckers et al. 2015].

Theorem 5.2.2. *Let $X = (X_y)_{y \in Y} \subseteq (K^n)_{y \in Y}$ be a subanalytic family of subanalytic sets of dimension $m < n$ in each model of $\mathcal{T}$. Fix $\varepsilon > 0$. Then there are a positive constant $C(X, k, \varepsilon)$, integers $l = l(X, k, \varepsilon)$, $N = N(X, k, \varepsilon)$, $\alpha = \alpha(m, n, k)$, and a family of blocks $W = (W_{y,s})_{(y,s) \in Y \times K^l} \subseteq K^n \times Y \times K^l$ such that the following holds.*

For each $L \in \mathscr{C}_{\mathcal{O},N}$, $H \geq 1$ and $y \in Y(L)$, there is a subset $S = S(X, k, L, H, y) \subseteq K^s$ of cardinality at most $C(X, \varepsilon)q^{\alpha}H^{\varepsilon}$ such that

$$X_y(L)(k, H) \subseteq \bigcup_{s \in S} W_{y,s}.$$

In particular, if we denote by W_y^{ε} the union over $s \in S$ of the $W_{y,s}(L)$ of positive dimension, we have $W_y^{\varepsilon} \subseteq X_y(L)^{\mathrm{alg}}$ and

$$\#(X_y(L) \backslash W_y^{\varepsilon})(k, H) \leq C(X, \varepsilon)q^{\alpha}H^{\varepsilon}.$$

The proof of Theorem 5.2.2 is completely similar to those of [Cluckers et al. 2015, Section 4.2] (namely to the proofs of Proposition 4.2.2 and Theorems 4.2.3 and 4.2.4), where instead of using [Cluckers et al. 2015, Proposition 4.2], one uses Proposition 5.1.4. We skip the proofs and refer to [Cluckers et al. 2015] for details. Theorem B in the introduction is the particular case of Theorem 5.2.2 when $k = 2$.

Remark 5.2.3. Note also that the bound in Proposition 5.1.4 is polylogarithmic, whereas the bound of [Cluckers et al. 2015, Proposition 4.2] is subpolynomial. However, this improvement does not guarantee a polylogarithmic bound in the counting theorems. As in the o-minimal case, such a bound is not expected to hold in general, but might be true in some specific situations, similar to the context of Wilkie's conjecture for $\mathbb{R}^{\exp}$-definable sets.

Acknowledgements

We would like to thank I. Halupczok for sharing inspiring ideas towards the piecewise Lipschitz continuity results of this paper. We thank also Z. Chatzidakis and M. Hils for useful discussions and comments, and the referee for valuable remarks. Cluckers was partially supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) with ERC Grant Agreement nr. 615722 MOTMELSUM, by the Labex CEMPI (ANR-11-LABX-0007-01), and by KU Leuven IF C14/17/083. Forey was partially supported by ANR-15-CE40-0008 (Défigéo) and by DFG-SNF lead agency program grant number 200020L_175755. Loeser was partially supported by ANR-15-CE40-0008 (Défigéo) and by the Institut Universitaire de France.

References

[Beyarslan and Chatzidakis 2017] Ö. Beyarslan and Z. Chatzidakis, "Geometric representation in the theory of pseudo-finite fields", *J. Symb. Log.* **82**:3 (2017), 1132–1139. MR Zbl

[Beyarslan and Hrushovski 2012] Ö. Beyarslan and E. Hrushovski, "On algebraic closure in pseudofinite fields", *J. Symb. Log.* **77**:4 (2012), 1057–1066. MR Zbl

[Bhargava et al. 2017] M. Bhargava, A. Shankar, T. Taniguchi, F. Thorne, J. Tsimerman, and Y. Zhao, "Bounds on 2-torsion in class groups of number fields and integral points on elliptic curves", preprint, 2017. To appear in J. Amer. Math. Soc. arXiv

[Binyamini and Novikov 2019] G. Binyamini and D. Novikov, "Complex cellular structures", *Ann. of Math.* (2) **190**:1 (2019), 145–248. MR Zbl

[Bombieri and Pila 1989] E. Bombieri and J. Pila, "The number of integral points on arcs and ovals", *Duke Math. J.* **59**:2 (1989), 337–357. MR Zbl

[Catanese 1992] F. Catanese, "Chow varieties, Hilbert schemes and moduli spaces of surfaces of general type", *J. Algebraic Geom.* **1**:4 (1992), 561–595. MR Zbl

[Chambert-Loir and Loeser 2017] A. Chambert-Loir and F. Loeser, "A nonarchimedean Ax–Lindemann theorem", *Algebra Number Theory* **11**:9 (2017), 1967–1999. MR Zbl

[Cilleruelo and Shparlinski 2013] J. Cilleruelo and I. Shparlinski, "Concentration of points on curves in finite fields", *Monatsh. Math.* **171**:3-4 (2013), 315–327. MR Zbl

[Cluckers and Halupczok 2012] R. Cluckers and I. Halupczok, "Approximations and Lipschitz continuity in p-adic semi-algebraic and subanalytic geometry", *Selecta Math.* (*N.S.*) **18**:4 (2012), 825–837. MR Zbl

[Cluckers and Lipshitz 2011] R. Cluckers and L. Lipshitz, "Fields with analytic structure", *J. Eur. Math. Soc.* **13**:4 (2011), 1147–1223. MR Zbl

[Cluckers and Lipshitz 2017] R. Cluckers and L. Lipshitz, "Strictly convergent analytic structures", *J. Eur. Math. Soc.* **19**:1 (2017), 107–149. MR Zbl

[Cluckers and Loeser 2007] R. Cluckers and F. Loeser, "b-minimality", *J. Math. Log.* **7**:2 (2007), 195–227. MR Zbl

[Cluckers et al. 2006] R. Cluckers, L. Lipshitz, and Z. Robinson, "Analytic cell decomposition and analytic motivic integration", *Ann. Sci. École Norm. Sup.* (4) **39**:4 (2006), 535–568. MR Zbl

[Cluckers et al. 2010] R. Cluckers, G. Comte, and F. Loeser, "Lipschitz continuity properties for p-adic semi-algebraic and subanalytic functions", *Geom. Funct. Anal.* **20**:1 (2010), 68–87. MR Zbl

[Cluckers et al. 2015] R. Cluckers, G. Comte, and F. Loeser, "Non-archimedean Yomdin–Gromov parametrizations and points of bounded height", *Forum Math. Pi* **3** (2015), art. id. e5. MR Zbl

[Cluckers et al. 2020] R. Cluckers, J. Pila, and A. Wilkie, "Uniform parameterization of subanalytic sets and Diophantine applications", *Ann. Sci. École Norm. Sup.* (4) **53**:1 (2020), 1–42.

[Cox et al. 2015] D. A. Cox, J. Little, and D. O'Shea, *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*, 4th ed., Springer, 2015. MR Zbl

[Marmon 2010] O. Marmon, "A generalization of the Bombieri–Pila determinant method", *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* **377**:10 (2010), 63–77. MR Zbl

[Nübling 2004] H. Nübling, "Adding Skolem functions to simple theories", *Arch. Math. Logic* **43**:3 (2004), 359–370. MR Zbl

[Pila 1995] J. Pila, "Density of integral and rational points on varieties", pp. 183–187 in *Columbia University Number Theory Seminar* (New York, 1992), Astérisque **228**, Soc. Math. France, Paris, 1995. MR Zbl

[Pila 2004] J. Pila, "Integer points on the dilation of a subanalytic surface", *Q. J. Math.* **55**:2 (2004), 207–223. MR Zbl

[Pila and Wilkie 2006] J. Pila and A. J. Wilkie, "The rational points of a definable set", *Duke Math. J.* **133**:3 (2006), 591–616. MR Zbl

[Rideau 2017] S. Rideau, "Some properties of analytic difference valued fields", *J. Inst. Math. Jussieu* **16**:3 (2017), 447–499. MR Zbl

[Salberger 2007] P. Salberger, "On the density of rational and integral points on algebraic varieties", *J. Reine Angew. Math.* **606** (2007), 123–147. MR Zbl

[Samuel 1955] P. Samuel, *Méthodes d'algèbre abstraite en géométrie algébrique*, Ergebnisse der Mathematik (2) **4**, Springer, 1955. MR Zbl

[Sedunova 2017] A. Sedunova, "On the Bombieri–Pila method over function fields", *Acta Arith.* **181**:4 (2017), 321–331. MR Zbl

[Vermeulen 2020] F. Vermeulen, "Points of bounded height on curves and the dimension growth conjecture over $\mathbb{F}_q[t]$", preprint, 2020. arXiv

Communicated by Philippe Michel
Received 2019-02-18 Revised 2020-01-09 Accepted 2020-03-03

raf.cluckers@univ-lille.fr *University of Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, F-59000 Lille, France*

KU Leuven, Department of Mathematics, Leuven, Belgium

arthur.forey@math.ethz.ch *D-Math, ETH Zürich, Zürich, Switzerland*

francois.loeser@imj-prg.fr *Institut Universitaire de France, Sorbonne Université, UMR 7586 CNRS, Institut Mathématique de Jussieu, Paris, France*

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1457

Gowers norms control diophantine inequalities

Aled Walker

A central tool in the study of systems of linear equations with integer coefficients is the generalised von Neumann theorem of Green and Tao. This theorem reduces the task of counting the weighted solutions of these equations to that of counting the weighted solutions for a particular family of forms, the Gowers norms $\|f\|_{U^{s+1}[N]}$ of the weight f. In this paper we consider systems of linear inequalities with real coefficients, and show that the number of solutions to such weighted diophantine inequalities may also be bounded by Gowers norms. Furthermore, we provide a necessary and sufficient condition for a system of real linear forms to be governed by Gowers norms in this way. We present applications to cancellation of the Möbius function over certain sequences.

The machinery developed in this paper can be adapted to the case in which the weights are unbounded but suitably pseudorandom, with applications to counting the number of solutions to diophantine inequalities over the primes. Substantial extra difficulties occur in this setting, however, and we have prepared a separate paper on these issues.

MSC2010: primary 11D75; secondary 11B30, 11J25.
Keywords: Gowers norms, diophantine inequalities, Möbius orthogonality, generalised von Neumann theorem.

1. Introduction

The field of *diophantine inequalities* is a large, and somewhat loosely defined, collection of problems which lie at the intersection of traditional number theory and diophantine approximation. As far as this paper is concerned, we will restrict our attention to the following class of questions. Let A be a set of integers, let ε be a positive parameter, and let L be an m-by-d real matrix. One might then ask whether there are infinitely many solutions to

$$\|L\boldsymbol{a}\|_\infty \leqslant \varepsilon \tag{1-1}$$

in which all of the coordinates of $\boldsymbol{a}$ lie in A. Further, letting N be an integer parameter tending to infinity, one might seek an asymptotic formula for the number of such solutions $\boldsymbol{a}$ which lie in the box given by the condition $\|\boldsymbol{a}\|_\infty \leqslant N$. One might even try to count solutions in certain cases in which L depends on N.

Much is known about these problems for certain special sets A (see [Baker 1967; 1986; Davenport and Heilbronn 1946; Margulis 1989; Müller 2005; Parsell 2002a; 2002b]), in particular for the image sets of polynomials. This work is discussed in Section 1A below. However, as far we are aware, the inequality (1-1) has not been considered before in such generality. Naturally there are some advantages and some disadvantages in pursuing such a general formulation, the main disadvantage being that the statements of our main results must perforce include some complicated technical hypotheses on the matrix L.

It will take us the rest of Sections 1 and 2 to properly motivate these hypotheses, culminating in the statement of Theorem 2.12 (our main theorem). Section 1 will focus on qualitative results and applications, whereas Section 2 goes on to explore the issues of diophantine approximation and nondegeneracy which are required for a quantitative treatment when L depends on N. At the end of Section 2 we will give a detailed sketch of our entire proof strategy. For now, we present the reader with a certain corollary of our main theorem, which we hope will encourage further reading through this long introduction.

Corollary 1.1 (example of Möbius orthogonality). *Let $\theta_1, \dots, \theta_s \in \mathbb{R}$ be distinct irrational numbers, let N be an integer parameter, and let $f_1, f_2, \dots, f_{s+1} : \{1, \dots, N\} \to [-1, 1]$ be arbitrary 1-bounded functions. Then*

$$\frac{1}{N^2} \sum_{\substack{x,d \in \mathbb{Z} \\ 1 \leqslant x \leqslant N}} \mu(x) f_1(x+d) \Bigg(\prod_{i=2}^{s+1} f_i([x+\theta_{i-1}d]) \Bigg) = o(1) \tag{1-2}$$

as $N \to \infty$, where μ denotes the Möbius function and $[x] := \lfloor x + \frac{1}{2} \rfloor$ is the nearest integer to x. The $o(1)$ error term may depend on the numbers $\theta_1, \dots, \theta_s$ but is independent of the choice of functions $f_1, \dots, f_{s+1}$.

1A. ***Classical results.*** As we said above, much is known about the inequality (1-1) for certain special sets A, particularly when $m = 1$. For example, if A is the set of squares, it was shown by Davenport and Heilbronn [1946] that there are infinitely many solutions to (1-1) for $m = 1$ and $d = 5$, i.e., infinitely

many solutions to

$$|\lambda_1 n_1^2 + \lambda_2 n_2^2 + \lambda_3 n_3^2 + \lambda_4 n_4^2 + \lambda_5 n_5^2| \leqslant \varepsilon, \tag{1-3}$$

provided the coefficients λ_i are nonzero, not all of the same sign, and not all in pairwise rational ratio. Their work also proves the same result for k-th powers, provided that the number of variables is at least $2^k + 1$. Some 55 years after Davenport and Heilbronn, Freeman [2002] refined the analysis from [Davenport and Heilbronn 1946] to obtain asymptotic formulas for the number of solutions to (1-3) in which $n_i \leqslant N$ for every i, and he also reduced the number of variables required in the case of k-th powers, to $k(\log k + \log\log k + O(1))$. Wooley [2003] further reduced this number, particularly for small k.

The Davenport–Heilbronn method is Fourier-analytic. One begins by replacing the interval $[-\varepsilon, \varepsilon]$ with a Lipschitz cutoff function, and then one expresses the solution count via the Fourier inversion formula (see [Davenport 1963, Chapter 20] or [Vaughan 1981, Chapter 11]). The device of replacing $[-\varepsilon, \varepsilon]$ with a friendlier cutoff plays an important role in our argument too, and we discuss it at length in Section 2E.

There are also some results on the inequality (1-1) when $m \geqslant 2$, although this setting has been studied less intensively. For example, Parsell [2002b] considered the setting of k-th powers, with Müller [2005] developing a refined result in the case of inequalities for general real quadratics. Parsell's result is rather technical to state, and we defer the interested reader to the original paper. Later on in Section 2, however, we will state Müller's result precisely, as one of his hypotheses is closely related to a hypothesis in our main theorem.

One of our main goals, for this paper and for our follow-up [Walker 2019], is to find a method of proving asymptotic formulae for the number of solutions to diophantine inequalities which goes beyond what can be done using the Davenport–Heilbronn method. Of particular interest to us is the case of inequality (1-1) when A is the set of prime numbers. A result first claimed by A. Baker [1967][1] states that for any fixed positive ε there exist infinitely many triples of primes (p_1, p_2, p_3) satisfying

$$|\lambda_1 p_1 + \lambda_2 p_2 + \lambda_3 p_3| \leqslant \varepsilon, \tag{1-4}$$

assuming again that the coefficients λ_i are nonzero, not all of the same sign, and not all in pairwise rational ratio. Parsell [2002a] then used a similar refinement to that of Freeman to prove a lower bound on the number of solutions to (1-4) in the box $p_1, p_2, p_3 \leqslant N$. For m simultaneous inequalities, and for a generic matrix L, Parsell's method is powerful enough[2] to prove an asymptotic formula for the number of solutions to (1-1) in prime variables $p_1, p_2, \ldots, p_d \leqslant N$, provided that $d \geqslant 2m + 1$. In [Walker 2019], building on the work of the present paper, we manage to reach the same conclusion under the weaker hypothesis that $d \geqslant m + 2$, provided that L has algebraic coefficients.

[1] In fact Baker [1967] proved a slightly different result, writing that the result we quote here followed easily from the then-existing methods.

[2] This does not seem to be present in the literature except in an appendix of our paper [Walker 2019].

A discussion of the literature on diophantine inequalities would not be complete without at least making reference to Margulis's famous resolution [1989] of the Oppenheim conjecture. With this work Margulis reduced the number of variables required to show the existence of infinitely many solutions to the inequality (1-3) from 5 to 3. Margulis's approach used dynamical methods, and is rather different in flavour to anything in this paper. In particular this method does not provide an asymptotic formula for the number of solutions in which the variables are bounded in a box.

1B. ***Notation.*** Before continuing with the rest of our introduction, we feel that, given the technical nature of some of the statements to follow, it is prudent to fix all our notation at the outset.

We will use standard asymptotic notation O, o, and Ω. We do not, as is sometimes the convention, for a function f and a positive function g choose to write $f = O(g)$ if there exists a constant C such that $|f(N)| \leqslant Cg(N)$ *for N sufficiently large*. Rather we require the inequality to hold for all N in some prespecified range. If N is a natural number, the range is always assumed to be $\mathbb{N}$ unless otherwise specified. For us, $0 \notin \mathbb{N}$.

It will be a convenient shorthand to use these symbols in conjunction with minus signs. So, by convention, we determine that expressions such as $-O(1)$, $-o(1)$, $-\Omega(1)$ are negative, e.g., $N^{-\Omega(1)}$ refers to a term N^{-c}, where c is some positive quantity bounded away from 0 as the asymptotic parameter tends to infinity. It will also be convenient to use the Vinogradov symbol $\ll$, where for a function f and a positive function g we write $f \ll g$ if and only if $f = O(g)$. We write $f \asymp g$ if $f \ll g$ and $g \ll f$. We also adopt the κ notation from [Green and Tao 2010a]: $\kappa(x)$ denotes any quantity that tends to zero as x tends to zero, with the exact value being permitted to change from line to line.

All the implied constants may depend on the dimensions of the underlying spaces. These will be obvious in context, and will always be denoted by m, d, h, or s (or, in the case of Proposition 4.8, by n). If an implied constant depends on other parameters, we will denote these by subscripts, e.g., $O_{c,C,\varepsilon}(1)$, or $f \asymp_\varepsilon g$. By notation such as $o_\rho(1)$ we mean a term which tends to zero as the asymptotic parameter tends to infinity with ρ fixed.

If N is a natural number, we use $[N]$ to denote $\{n \in \mathbb{N} : n \leqslant N\}$, whereas $[1, N]$ will be reserved for the closed real interval. For $x \in \mathbb{R}$, we write $[x] := \lfloor x + \frac{1}{2} \rfloor$ for the nearest integer to x, and $\|x\|$ for $|x - [x]|$. This means that there is slight overloading of the notation $[N]$, but the sense will always be obvious in context. When other norms are present, we may write $\|x\|_{\mathbb{R}/\mathbb{Z}}$ for $\|x\|$ to avoid confusion. For $\boldsymbol{x} \in \mathbb{R}^m$, we let $\|\boldsymbol{x}\|_{\mathbb{R}^m/\mathbb{Z}^m}$ denote $\sup_i |x_i - [x_i]|$.

If $X, Y \subset \mathbb{R}^d$ for some d, we define

$$\operatorname{dist}(X, Y) := \inf_{x \in X, y \in Y} \|x - y\|_\infty.$$

If X is the singleton $\{x\}$, we write $\operatorname{dist}(x, Y)$ for $\operatorname{dist}(\{x\}, Y)$. By identifying sets of m-by-d matrices with subsets of $\mathbb{R}^{md}$ (by identifying the coefficients of the matrices with coordinates in $\mathbb{R}^{md}$), we may also define $\operatorname{dist}(X, Y)$ when X and Y are sets of matrices of the same dimensions. We will consider a linear map $L : \mathbb{R}^d \to \mathbb{R}^m$ to be synonymous with the m-by-d matrix that represents L with respect to

the standard bases. The norm $\|L\|_\infty$ will refer to the maximum absolute value of the coefficients of this matrix. We use the notation $L^*:(\mathbb{R}^m)^*\to(\mathbb{R}^d)^*$ for the dual linear map between the dual spaces. For a set $U\subset\mathbb{R}^d$ we use U^0 to denote the annihilator of U, i.e., the set of all f in the dual space $(\mathbb{R}^d)^*$ for which $f|_U\equiv 0$.

We let $\partial(X)$ denote the topological boundary of a set $X\subset\mathbb{R}^d$. Given $S\subset\mathbb{R}$ and $\lambda\in\mathbb{R}$, we let $\lambda S:=\{x\in\mathbb{R}:\exists s\in S \text{ for which } \lambda s=x\}$. If A and B are two sets with $A\subseteq B$, we let $1_A:B\to\{0,1\}$ denote the indicator function of A. The relevant set B will usually be obvious from context. The notation for logarithms, log, will always denote the natural log. For $\theta\in\mathbb{R}$ we also adopt the standard shorthand $e(\theta)$ to mean $e^{2\pi i\theta}$.

In Section 8, if $\boldsymbol{x}\in\mathbb{R}^d$ and if a and b are two subscripts with $1\leqslant a\leqslant b\leqslant d$, we use the notation $\boldsymbol{x}_a^b$ to denote the vector $(x_a,x_{a+1},\dots,x_b)^T\in\mathbb{R}^{b-a+1}$.

1C. ***The main corollary.*** We will now begin the process of developing our first main result, namely Corollary 1.4. This result is the first to link diophantine inequalities, such as (1-1), to *Gowers norms*.

Given natural numbers N and d, and a function $f:[N]\to\mathbb{C}$, the *Gowers* U^d *norm* $\|f\|_{U^d[N]}$ was introduced into the literature around twenty years ago, as part of Gowers' [2001] proof of Szemerédi's theorem.[3] The U^d norms are genuine norms for $d\geqslant 2$, with $\|f\|_{U^d[N]}$ measuring the density of certain linear patterns weighted by f. Their presence in analytic number theory is by now well established (see for instance [Green and Tao 2008a; 2010a; Tao and Teräväinen 2018; 2019]), but, to help any readers who are unfamiliar with these norms, in Appendix A we have given a summary of the necessary definitions and basic notions.

Our present endeavour is motivated by one particular application of Gowers norms, namely the so-called "generalised von Neumann theorem" developed by Green and Tao [2008a; 2010a] to study linear equations with rational coefficients.

Theorem 1.2 (generalised von Neumann theorem for rational forms (nonquantitative)). *Let m,d be natural numbers, satisfying $d\geqslant m+2$. Let L be an m-by-d real matrix with integer coefficients, with rank m. Suppose that there does not exist any nonzero row-vector in the row-space of L that has two or fewer nonzero coordinates. Then there is some natural number s at most $d-2$ that satisfies the following. Let N be an integer parameter, let $f_1,\dots,f_d:[N]\to[-1,1]$ be arbitrary functions, and suppose that*

$$\min_j\|f_j\|_{U^{s+1}[N]}\leqslant\rho$$

for some parameter ρ in the range $0<\rho\leqslant 1$. Then

$$\frac{1}{N^{d-m}}\sum_{\substack{\boldsymbol{n}\in[N]^d\\ L\boldsymbol{n}=\boldsymbol{0}}}\prod_{j=1}^{d}f_j(n_j)\ll_L\rho^{\Omega(1)}+o_\rho(1)$$

as $N\to\infty$.

[3] Gowers worked over the cyclic group $\mathbb{Z}/N\mathbb{Z}$ rather than $[N]$, but this is a very minor difference.

Results similar to Theorem 1.2 are central to Green and Tao's approach [2010a] to counting solutions to linear equations in primes. It seems reasonable to hope that, if one could combine Gowers norms and diophantine inequalities in a suitable way, then one might be able to develop a strong understanding of linear inequalities in primes. As we have already intimated in Section 1A, when describing our improvements to Parsell's results, this can be done. However, many additional technical difficulties occur for the primes, as the von Mangoldt function is unbounded; we have chosen to present a separate work on these issues [Walker 2019].

We should briefly discuss the nondegeneracy condition on L in the statement of Theorem 1.2, namely that "there does not exist any nonzero row-vector in the row-space of L that has two or fewer nonzero coordinates", as it may seem a little unnatural at first sight.[4] Suppose that such a row-vector existed. Suppose also that it is the coordinates at index i and index j which are nonzero. Then, by a short linear algebra argument (see Proposition 4.5), for any linear parametrisation $(\psi_1, \dots, \psi_d) = \Psi : \mathbb{R}^{d-m} \to \ker L$, ψ_i is a multiple of ψ_j. Such a coupling between the coordinates has dire consequences for any averaging approach built upon the independence of the different coordinates, such as the averaging in Gowers norms, and so this coupling must be precluded. We will present a rigorous version of this principle in the context of linear inequalities, in Theorem 2.14 below.

Regarding the condition $d \geqslant m+2$, if L has rank m and $d \leqslant m+1$ then in fact, as follows from Gaussian elimination, there must always exist a nonzero vector in the row space of L with two or fewer nonzero coordinates. Thus, the condition $d \geqslant m+2$ is a necessary one if the coordinates of $\ker L$ are to be suitably independent.

Remark 1.3. Theorem 1.2 is implicit in [Green and Tao 2010a], but there is no explicit such statement presented there, as those authors were focussed on results over the primes. We will describe how to extract Theorem 1.2 from the arguments of [loc. cit.], but we postpone this task until Section 5, as at that point we will also introduce a quantitative version (this will be Theorem 5.2).

Our first main result is a version of Theorem 1.2 for diophantine inequalities.

Corollary 1.4 (Gowers norms control diophantine inequalities (nonquantitative)). *Let m, d be natural numbers, satisfying $d \geqslant m+2$, and let ε be a positive parameter. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be an m-by-d real matrix, with rank m. Suppose that there does not exist any nonzero row-vector in the row-space of L that has two or fewer nonzero coordinates. Then there is some natural number s at most $d-2$, independent of ε, such that the following is true. Let N be an integer parameter and let $f_1, \dots, f_d : [N] \to [-1,1]$ be arbitrary functions, and suppose that*

$$\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho,$$

[4] For readers who are already familiar with the notion of Cauchy–Schwarz complexity, imposing this nondegeneracy condition on L is equivalent to insisting that $\ker L$ may be parametrised by a system of linear forms with finite Cauchy–Schwarz complexity.

for some parameter ρ in the range $0 < \rho \leqslant 1$. Then

$$\left| \frac{1}{N^{d-m}} \sum_{\substack{\boldsymbol{n} \in [N]^d \\ \|L\boldsymbol{n}\|_\infty \leqslant \varepsilon}} \left(\prod_{j=1}^{d} f_j(n_j) \right) \right| \ll_{L,\varepsilon} \rho^{\Omega(1)} + o_{\rho,L}(1)$$

as $N \to \infty$.

We can provide detailed information about how the implied constant and the $o_{\rho,L}(1)$ term depend on L, but we leave that to the next section and to Theorem 2.12.

Before giving some examples, let us make a few remarks about Corollary 1.4. Firstly, note that if L has integer coefficients then, by picking ε small enough, Corollary 1.4 immediately implies Theorem 1.2, since the inequality $\|L\boldsymbol{n}\|_\infty \leqslant \varepsilon$ is only satisfied if $L\boldsymbol{n} = \boldsymbol{0}$.

Next, due to the nested property of Gowers norms (see Appendix A) one sees that Corollary 1.4 may be fruitfully applied under the hypothesis $\min_j \|f_j\|_{U^{d-1}[N]} \leqslant \rho$.

Finally, we note that it might be tempting to think that Corollary 1.4 would follow easily from taking rational approximations of the coefficients of L and then using Theorem 1.2 as a black box. Though of course we cannot completely rule out an alternative approach to that of this paper, when one investigates the quantitative details it seems that such an argument will only quickly succeed if the coefficients of L are all extremely well-approximable by rationals, else the height of the rational approximations becomes too great to apply results like Theorem 1.2. We will need to employ a different strategy, and we discuss this at length in Section 2E.

1D. ***Fourier uniform sets and other examples.*** Let us illustrate the applications of Corollary 1.4 with certain explicit examples.

Example 1.5 (three-term irrational AP). The first example could have been analysed by Davenport and Heilbronn using the methods they developed [1946], but we include it here to demonstrate the simplest case in which Corollary 1.4 applies.

Let

$$L := (1 \quad -\sqrt{2} \quad -1+\sqrt{2}).$$

Then $m = 1$ and $d = 3$, and manifestly there does not exist any nonzero row-vector in the row-space of L that has two or fewer nonzero coordinates. Therefore Corollary 1.4 applies, and so, if for each N we have three functions $f_1, f_2, f_3 : [N] \to [-1, 1]$ satisfying $\min_j \|f_j\|_{U^2[N]} \leqslant \rho$ for some ρ in the range $0 < \rho \leqslant 1$, then we have

$$\left| \frac{1}{N^2} \sum_{\substack{n_1,n_2,n_3 \leqslant N \\ |n_1 - \sqrt{2}n_2 + (-1+\sqrt{2})n_3| \leqslant \varepsilon}} f_1(n_1) f_2(n_2) f_3(n_3) \right| \ll_\varepsilon \rho^{\Omega(1)} + o_\rho(1) \tag{1-5}$$

as $N \to \infty$.

The statement (1-5) admits a different interpretation, which some readers may find more natural, that of counting the number of occurrences of a certain irrational pattern: a "three-term irrational arithmetic progression". Indeed, recall that for $\theta \in \mathbb{R}$ we let $[\theta]$ denote $\lfloor \theta + \frac{1}{2} \rfloor$, i.e., the nearest integer to θ. Then for any three functions $f_1, f_2, f_3 : [N] \to [-1, 1]$, we make the definition

$$T(f_1, f_2, f_3) := \frac{1}{N^2} \sum_{x,d \in \mathbb{Z}} f_3(x) f_2(x+d) f_1([x + \sqrt{2}d]). \tag{1-6}$$

Informally speaking, T counts the number of near-occurrences of the pattern $(x, x+d, x+\sqrt{2}d)$, weighted by the functions f_j. By performing the change of variables $n_1 = [x + \sqrt{2}d]$, $n_2 = x + d$, $n_3 = x$, and noting that $x + \sqrt{2}d \notin \frac{1}{2}\mathbb{Z}$, we see that

$$T(f_1, f_2, f_3) = \frac{1}{N^2} \sum_{\substack{n_1,n_2,n_3 \leqslant N \\ |n_1 - \sqrt{2}n_2 + (-1+\sqrt{2})n_3| \leqslant \frac{1}{2}}} f_1(n_1) f_2(n_2) f_3(n_3). \tag{1-7}$$

By (1-5), this means that if $\min_j \|f_j\|_{U^2[N]} \leqslant \rho$ then

$$|T(f_1, f_2, f_3)| \ll \rho^{\Omega(1)} + o_\rho(1) \tag{1-8}$$

as $N \to \infty$.

Now suppose that A_N is a subset of $[N]$ with $|A_N| = \alpha_N N$. Let

$$f_{A_N} := 1_{A_N} - \alpha_N 1_{[N]}$$

be its so-called "balanced function". By the usual telescoping trick, $T(1_{A_N}, 1_{A_N}, 1_{A_N})$ is equal to

$$T(\alpha_N 1_{[N]}, \alpha_N 1_{[N]}, \alpha_N 1_{[N]}) + T(f_{A_N}, \alpha_N 1_{[N]}, \alpha_N 1_{[N]}) + T(1_{A_N}, f_{A_N}, \alpha_N 1_{[N]}) + T(1_{A_N}, 1_{A_N}, f_{A_N}).$$

One may then bound the final three terms using $\|f_{A_N}\|_{U^2[N]}$ and, from the relation (1-7), one has then established that, provided $\|f_{A_N}\|_{U^2[N]} \leqslant \rho$,

$$\frac{1}{N^2} \sum_{x,d \in \mathbb{Z}} 1_{A_N}(x) 1_{A_N}(x+d) 1_{A_N}([x + \sqrt{2}d])$$

is equal to

$$\frac{\alpha_N^3}{N^2} \sum_{x,d \in \mathbb{Z}} 1_{[N]}(x) 1_{[N]}(x+d) 1_{[N]}([x + \sqrt{2}d]) + O(\rho^{\Omega(1)}) + o_\rho(1) \tag{1-9}$$

as $N \to \infty$. If $\|f_{A_N}\|_{U^2[N]} = o(1)$ as $N \to \infty$ then, by picking $\rho = \rho(N)$ to be a quantity that tends to zero suitably slowly, one can ensure that the error term in (1-9) is $o(1)$ as $N \to \infty$.

As is familiar from [Gowers 2001], for bounded functions the U^2-norm is closely related to the Fourier transform. Indeed, we say that the family of sets A_N is Fourier-uniform if the balanced functions f_{A_N} satisfy

$$\sup_{\theta \in [0,1]} \left| \frac{1}{N} \sum_{n \leqslant N} f_{A_N}(n) e(n\theta) \right| = o(1)$$

as $N \to \infty$, and it is a standard result (see [Tao 2012, Exercise 1.3.18]) that A_N is Fourier uniform if and only if $\|f_{A_N}\|_{U^2[N]} = o(1)$ as $N \to \infty$. Therefore expression (1-9), and the remarks following it, imply the following corollary.

Corollary 1.6 (Fourier-uniform sets). *Let $\beta \in \mathbb{R} \setminus \mathbb{Q}$, and for each natural number N let A_N be a subset of $[N]$ with $|A_N| = \alpha_N N$. Suppose that A_N is a Fourier-uniform family of sets. Then*

$$\frac{1}{N^2} \sum_{x,d \in \mathbb{Z}} 1_{A_N}(x) 1_{A_N}(x+d) 1_{A_N}([x+\beta d])$$

is equal to

$$\frac{\alpha_N^3}{N^2} \sum_{x,d \in \mathbb{Z}} 1_{[N]}(x) 1_{[N]}(x+d) 1_{[N]}([x+\beta d]) + o_\beta(1)$$

as $N \to \infty$, where the $o_\beta(1)$ term also depends on the $o(1)$ term in the Fourier-uniformity expression for the family A_N.

Example 1.7. Let

$$L := \begin{pmatrix} 1 & 0 & -\sqrt{2} & -1+\sqrt{2} \\ 0 & 1 & -\sqrt{3} & -1+\sqrt{3} \end{pmatrix}. \tag{1-10}$$

Since $\sqrt{2}$ and $\sqrt{3}$ are distinct irrationals it is not hard to see that all elements of the row-space of L must have three or four nonzero coordinates, and so Corollary 1.4 applies. Letting $f_1, f_2, f_3, f_4 : [N] \to [-1, 1]$ be arbitrary functions, the reparametrisation $n_1 = [x + \sqrt{2}d]$, $n_2 = [x + \sqrt{3}d]$, $n_3 = x + d$, $n_4 = x$, shows that

$$\frac{1}{N^2} \sum_{\substack{\boldsymbol{n} \in [N]^4 \\ \|L\boldsymbol{n}\|_\infty \leqslant \frac{1}{2}}} \left(\prod_{j=1}^4 f_j(n_j) \right) = \frac{1}{N^2} \sum_{x,d \in \mathbb{Z}} f_4(x) f_3(x+d) f_1([x+\sqrt{2}d]) f_2([x+\sqrt{3}d]).$$

Corollary 1.4 controls the left-hand side of this expression in terms of the Gowers norms of the functions f_j, and so the right-hand side is controlled as well.

We can generalise the previous two examples as follows.

Corollary 1.8. *Let $\theta_1, \dots, \theta_s \in \mathbb{R}$ be distinct irrational numbers. For each natural number N let A_N be a subset of $[N]$, with $|A_N| = \alpha_N N$ and with balanced function f_{A_N}. Suppose that $\|f_{A_N}\|_{U^{s+1}[N]} = o(1)$ as $N \to \infty$. Then*

$$\frac{1}{N^2} \sum_{x,d \in \mathbb{Z}} 1_{A_N}(x) 1_{A_N}(x+d) \left(\prod_{i=1}^s 1_{A_N}([x+\theta_i d]) \right) \tag{1-11}$$

is equal to

$$\frac{\alpha_N^{s+2}}{N^2} \sum_{x,d \in \mathbb{Z}} 1_{[N]}(x) 1_{[N]}(x+d) \left(\prod_{i=1}^s 1_{[N]}([x+\theta_i d]) \right) + o(1)$$

as $N \to \infty$, where the $o(1)$ error term may depend on $\theta_1, \dots, \theta_s$ and on the rate of decay of $\|f_{A_N}\|_{U^{s+1}[N]}$.

Proof. Apply Corollary 1.4 to the s-by-$s+2$ matrix

$$L = (I \quad -\boldsymbol{\theta} \quad -1+\boldsymbol{\theta}), \tag{1-12}$$

where I denotes the identity matrix and $\boldsymbol{\theta}$ denotes the vector $(\theta_1, \dots, \theta_s)^T \in \mathbb{R}^s$. □

The question remains as to whether one can use Corollary 1.8, perhaps in conjunction with a density increment argument such as is used in [Green and Tao 2010b], to deduce that there are infinitely many $s+2$-tuples of the form $(x, x+d, [x+\theta_1 d], \dots, [x+\theta_s d])$ inside any set of natural numbers with positive upper Banach density. More generally, one might wish to find tuples in which all the coordinates are of the form $x + p(d)$ where p is a generalised polynomial of degree 1 without a constant term. This result is already known, in fact, but as it stands the only proof uses ergodic theory methods (see [McCutcheon 2005, Theorem B]). We view Corollary 1.8 as a promising first step towards a purely combinatorial proof of this result, with a chance to prove explicit bounds.

Corollary 1.4 has immediate consequences for counting solutions to diophantine inequalities weighted by explicit bounded pseudorandom functions. In particular there is the following natural analogue of [Green and Tao 2010a, Proposition 9.1] concerning the cancellation of the Möbius function, which we mentioned earlier in Corollary 1.1.

Corollary 1.9 (Möbius orthogonality). *Let m, d be natural numbers satisfying $d \geqslant m+2$, and let ε be a positive parameter. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be an m-by-d real matrix, with rank m. Suppose that there does not exist any nonzero row-vector in the row-space of L that has two or fewer nonzero coordinates. Let μ denote the Möbius function and let N be an integer parameter. Then, for any bounded functions $f_2, \dots, f_d : [N] \to [-1, 1]$,*

$$\sum_{\substack{\boldsymbol{n} \in [N]^d \\ \|L\boldsymbol{n}\|_\infty \leqslant \varepsilon}} \mu(n_1) \left(\prod_{j=2}^{d} f_j(n_j) \right) = o_{L,\varepsilon}(N^{d-m})$$

as $N \to \infty$. The same is true with μ replaced by the Liouville function λ.

Proof. This follows immediately from Corollary 1.4 and the deep facts (stated in [Green and Tao 2010a], proved in [Green and Tao 2012] and [Green et al. 2012]) that $\|\mu\|_{U^{s+1}[N]} = o_s(1)$ and $\|\lambda\|_{U^{s+1}[N]} = o_s(1)$ as $N \to \infty$. □

Corollary 1.9, when applied to the matrix (1-12), yields Corollary 1.1 from earlier in this introduction. It also yields cancellation in expressions such as

$$\sum_{\substack{\boldsymbol{n} \in [N]^4 \\ n_1 - n_2 = n_2 - n_3 \\ |(n_2 - n_3) - \sqrt{2}(n_3 - n_4)| \leqslant \frac{1}{2}}} \mu(n_1)\mu(n_2)\mu(n_3)\mu(n_4) = o(N^2) \tag{1-13}$$

as $N \to \infty$. There are of course many such examples; we chose (1-13) to emphasise that one can choose configurations that combine rational and irrational relations.

2. The main theorem

In our results from the previous section, the quantitative nature of the dependence of the error terms on the matrix L was hidden. Our main theorem (Theorem 2.12 below) addresses this point, which turns out to be surprisingly subtle.

To start off, let us introduce a multilinear form that will count solutions to a general version of (1-1).

Definition 2.1. Let N, m, d be natural numbers, and let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a linear map. Let $F : \mathbb{R}^d \to [0, 1]$ and $G : \mathbb{R}^m \to [0, 1]$ be two functions, with F supported on $[-N, N]^d$ and G compactly supported. Let $f_1, \dots, f_d : [N] \to [-1, 1]$ be arbitrary functions. We define

$$T^L_{F,G,N}(f_1, \dots, f_d) := \frac{1}{N^{d-m}} \sum_{\boldsymbol{n} \in \mathbb{Z}^d} \left(\prod_{j=1}^{d} f_j(n_j) \right) F(\boldsymbol{n}) G(L\boldsymbol{n}). \tag{2-1}$$

The normalisation factor of N^{d-m} is appropriate; in Lemma 3.2 we show that $T^L_{F,G,N}(f_1, \dots, f_d) \ll_G 1$.

In Theorem 2.12 we will bound $T^L_{F,G,N}(f_1, \dots, f_d)$ above by Gowers norms. The error term will depend on three further notions: the rational relations present in L; the "approximation function" A_L, which will measure the approximate rational relations present in L; and the nondegeneracy of L, which is related to the nondegeneracy conditions in Theorem 1.2. These three notions will be introduced in the next three subsections, before we (finally!) state Theorem 2.12 in Section 2D.

2A. *Rational relations.*

Let us consider some properties of the image $L(\mathbb{Z}^d)$. It is certainly true that if $L : \mathbb{R}^d \to \mathbb{R}^m$ is a surjective linear map then $\operatorname{span}(L(\mathbb{Z}^d)) = \mathbb{R}^m$. However, $L(\mathbb{Z}^d)$ need not be dense in $\mathbb{R}^m$, as the matrix L may satisfy some rational relations. These in turn restrict $L(\mathbb{Z}^d)$ to various affine subspaces of $\mathbb{R}^m$.

This observation motivates the following definitions:

Definition 2.2 (rational dimension, rational map, purely irrational). Let m and d be natural numbers, with $d \geqslant m + 1$. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map. Let u denote the largest integer for which there exists a surjective linear map $\Theta : \mathbb{R}^m \to \mathbb{R}^u$ for which $\Theta L(\mathbb{Z}^d) \subseteq \mathbb{Z}^u$. We call u the *rational dimension* of L, and we call any map Θ with the above property a *rational map* for L. We say that L is *purely irrational* if $u = 0$.

For example, suppose that $L : \mathbb{R}^4 \to \mathbb{R}^2$ is the linear map represented by the matrix

$$L := \begin{pmatrix} 1 & 0 & -\sqrt{2} & -\sqrt{3}+1 \\ 0 & 1 & 5\sqrt{2} & 5\sqrt{3} \end{pmatrix}$$

If $\Theta : \mathbb{R}^2 \to \mathbb{R}$ is given by the matrix

$$\Theta := (5 \;\; 1),$$

then $\Theta L(\mathbb{Z}^4) \subseteq \mathbb{Z}$, and in fact $\Theta L(\mathbb{Z}^4) = \mathbb{Z}$. So the rational dimension of L is at least 1. But the rational dimension of L cannot be 2, as if there were a surjective map $\Theta : \mathbb{R}^2 \to \mathbb{R}^2$ such that $\Theta L(\mathbb{Z}^4) \subseteq \mathbb{Z}^2$ then $L(\mathbb{Z}^4)$ would be a subset of a 2-dimensional lattice, which it is not. So the rational dimension of L is equal to 1.

Ours is certainly not the first paper on the topic of diophantine inequalities to have considered issues such as this. For example, in the previous section we remarked that Müller [2005] came across a similar phenomenon. Given quadratic forms $Q_1, \dots, Q_r$ he found infinitely many solutions $\boldsymbol{x}$ to the inequalities

$$|Q_1(\boldsymbol{x})| < \varepsilon, \dots, |Q_r(\boldsymbol{x})| < \varepsilon,$$

under the hypotheses that every quadratic form in the set

$$\left\{ \sum_{i=1}^{r} \alpha_i Q_i : \alpha_1, \dots, \alpha_r \in \mathbb{R},\ \boldsymbol{\alpha} \neq \mathbf{0} \right\}$$

was irrational and had rank greater than $8r$. One can use the coefficients of the quadratic forms to translate this problem into one of trying to understand $T^L_{F,G,N}(f_1, \dots, f_d)$ for a certain linear map L and for functions $f_1, \dots, f_d$ supported on the image of quadratic monomials. Then, Müller's hypothesis that all the linear combinations of the Q_i are irrational is transformed into the hypothesis that L is purely irrational. In this paper we consider all L, not just those which are purely irrational, and this causes some added complications.

In our definition of rational dimension, there is some flexibility over the exact choice of map Θ. The next lemma identifies an invariant.

Lemma 2.3. *Let m and d be natural numbers, with $d \geqslant m+1$. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map, and let u be the rational dimension of L. Then, if $\Theta_1, \Theta_2 : \mathbb{R}^m \to \mathbb{R}^u$ are two rational maps for L, $\ker \Theta_1 = \ker \Theta_2$.*

Proof. Suppose that $\Theta_1, \Theta_2 : \mathbb{R}^m \to \mathbb{R}^u$ are two rational maps for L for which $\ker \Theta_1 \neq \ker \Theta_2$. Then consider the map $(\Theta_1, \Theta_2) : \mathbb{R}^m \to \mathbb{R}^{2u}$. The kernel of this map has dimension at most $m-u-1$, as it is the intersection of two different subspaces of dimension $m-u$. Therefore the image has dimension at least $u+1$.

Also, $((\Theta_1, \Theta_2) \circ L)(\mathbb{Z}^d) \subseteq \mathbb{Z}^{2u}$. Let Φ be any surjective map from $\mathrm{im}((\Theta_1, \Theta_2))$ to $\mathbb{R}^{u+1}$ for which $\Phi(\mathbb{Z}^{2u} \cap \mathrm{im}((\Theta_1, \Theta_2))) \subseteq \mathbb{Z}^{u+1}$. Then $\Phi \circ (\Theta_1, \Theta_2) : \mathbb{R}^m \to \mathbb{R}^{u+1}$ is surjective and $(\Phi \circ (\Theta_1, \Theta_2) \circ L)(\mathbb{Z}^d) \subseteq \mathbb{Z}^{u+1}$. This contradicts the definition of u as the rational dimension. □

We will also need to identify the quantitative aspects of these rational relations, in order to properly state the main theorem.

Definition 2.4 (rational complexity). Let m and d be natural numbers, with $d \geqslant m+1$. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map, and let u denote the rational dimension of L. We say that L has *rational complexity at most C* if there exists a map Θ that is a rational map for L and for which $\|\Theta\|_\infty \leqslant C$. If L is purely irrational, then L has rational complexity 0.

In this definition, recall that for a linear map $\Theta : \mathbb{R}^m \to \mathbb{R}^u$ we use $\|\Theta\|_\infty$ to denote the maximum absolute value of the coefficients of its matrix with respect to the standard bases.

We observe that a linear map with maximal rational dimension, i.e., with rational dimension m, is equivalent to a linear map with integer coefficients, in the following sense.

Lemma 2.5 (maximal rational dimension). *Let m and d be natural numbers, with $d \geqslant m+1$. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map, and suppose that L has rational dimension m and rational complexity at most C. Then there exists an invertible m-by-m matrix Θ and an m-by-d matrix S with integer coefficients such that, as matrices, $\Theta L = S$. Furthermore, $\|\Theta\|_\infty \leqslant C$.*

Proof. Let $\Theta : \mathbb{R}^m \to \mathbb{R}^m$ be a rational map for L for which $\|\Theta\|_\infty \leqslant C$. □

We will use this lemma in Section 5, to reduce the study of maps L with maximal rational dimension to the study of maps L with integer coefficients.

2B. ***Approximation function.*** We must also quantify the rational relations in a second way. Indeed, L might have rational dimension u but be extremely close to having rational dimension at least $u+1$, in the sense that there might exist some surjective linear map $\Theta : \mathbb{R}^m \to \mathbb{R}^{u+1}$ such that the matrix of ΘL is very close to having integer coefficients. This phenomenon, essentially a notion of diophantine approximation, will also have a quantitative effect on our final bounds. The critical place where it enters the argument is Lemma 3.4, whose content we briefly sketch here, so as to further motivate our introduction of the "approximation function" below.

This will be the first, of many, places in the paper in which we have to manipulate Lipschitz functions. For the reader's benefit, in Appendix B we have collected together the definitions and results that we will use.

Let L be an m-by-d matrix, which may depend on the asymptotic parameter N. Suppose that one is seeking an upper bound on $T^L_{F,G,N}(1,\dots,1)$, where G is a Lipschitz function supported on $[-1,1]^m$ and F is a function supported on $[-N,N]^d$. We note that this task is a special case of bounding $T^L_{F,G,N}(f_1,\dots,f_d)$ above by the Gowers norms of the functions f_i. In our main proof, bounds on $T^L_{F,G,N}(1,\dots,1)$ will be useful when controlling some error terms which occur when the inequality is perturbed (see Section 2E for a full discussion of this point).

Also suppose, for simplicity, that the first m columns of L form the identity matrix, and let $\boldsymbol{v_j} \in \mathbb{R}^m$ denote the j-th column of L. Then, by summing over the variables $n_1,\dots,n_m \in \mathbb{Z}$, one quickly derives that

$$T^L_{F,G,N}(1,\dots,1) \ll \frac{1}{N^{d-m}} \sum_{\substack{n_{m+1},\dots,n_d \in \mathbb{Z} \\ |n_{m+1}|,\dots,|n_d| \leqslant N}} \tilde{G}\Big(\sum_{j=m+1}^{d} \boldsymbol{v_j} n_j\Big),$$

where $\tilde{G}$ is a $\mathbb{Z}^m$-periodic Lipschitz function formed by taking translates of G.

We consider $\tilde{G}$ as a function on $\mathbb{R}^m/\mathbb{Z}^m$, and approximate it by a short exponential sum (as one may do for all such Lipschitz functions).[5] As is familiar in these kind of problems, one is left with having to bound the expression that arises from the nonzero Fourier modes. Here, one ends up with terms

$$\frac{1}{N^{d-m}} \sum_{\substack{n_{m+1},\dots,n_d \in \mathbb{Z} \\ |n_{m+1}|,\dots,|n_d| \leqslant N}} e\Big(\boldsymbol{k} \cdot \sum_{j=m+1}^{d} \boldsymbol{v_j} n_j\Big)$$

[5] See Lemma B.3.

with $\boldsymbol{k} \in \mathbb{Z}^m \setminus \{\mathbf{0}\}$, which one may sum as geometric progressions over n_{m+1} to n_d. This means we have to control

$$\max_{\substack{\boldsymbol{k}\in\mathbb{Z}^m \\ 0<\|\boldsymbol{k}\|_\infty \leqslant X}} \Bigg(\prod_{j=m+1}^{d} \min(1, N^{-1}\|\boldsymbol{k}\cdot\boldsymbol{v}_j\|_{\mathbb{R}/\mathbb{Z}}^{-1}) \Bigg),$$

where X is some threshold, and the above expression is certainly bounded by

$$N^{-1} \max_{\substack{\boldsymbol{k}\in\mathbb{Z}^m \\ 0<\|\boldsymbol{k}\|_\infty \leqslant X}} \|L^T\boldsymbol{k}\|_{\mathbb{R}^d/\mathbb{Z}^d}^{-1}, \tag{2-2}$$

as the first m columns of L have integer coordinates. One hopes to bound expression (2-2) by $o(1)$ as $N \to \infty$.

We observe two facts about (2-2). Firstly, if L is not purely irrational and if X is larger than the rational complexity of L, then the expression (2-2) is infinite! Secondly, even if L is purely irrational then it could still be the case that (2-2) tends to infinity with N, as L may depend on N. We conclude that, with the state of our current argument, the size of expression (2-2) — or an expression like it — must be included in our error terms.

Motivated by the above discussion, we introduce the "approximation function". The definition is phrased in terms of dual functions — this will make linear algebraic manipulations more straightforward later on — and for real vectors φ rather than for integer vectors $\boldsymbol{k}$, which reflects the general situation in which the first m columns of L are not the identity matrix. We also generalise to the case of arbitrary rational dimension u, rather than just $u = 0$.

Following this definition and some remarks, we will show how to calculate the approximation function in an explicit example. This should hopefully serve to clarify the properties of this somewhat technical object.

Definition 2.6 (approximation function). Let m and d be natural numbers, with $d \geqslant m+1$. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map, and let u denote the rational dimension of L. Let $\Theta : \mathbb{R}^m \to \mathbb{R}^u$ be any rational map for L. Suppose that $u \leqslant m-1$. Then we define the *approximation function of* L, denoted $A_L : (0,1] \times (0,1] \to (0,\infty)$, by

$$A_L(\tau_1,\tau_2) := \inf_{\substack{\varphi\in(\mathbb{R}^m)^* \\ \operatorname{dist}(\varphi,\Theta^*((\mathbb{R}^u)^*))\geqslant\tau_1 \\ \|\varphi\|_\infty\leqslant\tau_2^{-1}}} \operatorname{dist}(L^*\varphi, (\mathbb{Z}^d)^T),$$

where $(\mathbb{Z}^d)^T$ denotes the set of those $\varphi \in (\mathbb{R}^d)^*$ that have integer coordinates with respect to the standard dual basis.

If $u = m$, we define $A_L(\tau_1,\tau_2)$ to be identically equal to τ_1.[6]

[6]When $u = m$ we've already seen (Lemma 2.5) that L may be transformed into a matrix S with integer coefficients, and thus L is somewhat degenerate from the point of view of diophantine approximation. We define $A_L(\tau_1,\tau_2)$ for such matrices only to avoid having to discuss this special case in the statement of Theorem 2.12 later on.

From our discussion of (2-2) above, one sees that upper bounds on $A_L(\tau_1, \tau_2)^{-1}$ will be the main focus. The threshold τ_2^{-1} plays the role of the threshold X in (2-2), and the condition $\operatorname{dist}(\varphi, \Theta^*((\mathbb{R}^u)^*)) \geqslant \tau_1$ corresponds to the condition $\|\boldsymbol{k}\|_\infty \geqslant 1$ which is implicit in (2-2).

There is some notation to unpack in Definition 2.6. Regarding the notion "dist", we remind the reader of some material from Section 1B, namely that we consider a-by-b matrices M as elements of $\mathbb{R}^{ab}$, simply by identifying the coefficients of M with coordinates in $\mathbb{R}^{ab}$. The ℓ^∞ norm and the dist operator may then be defined on matrices, i.e., if V is a set of a-by-b matrices, and L is an a-by-b matrix, then

$$\operatorname{dist}(L, V) := \inf_{L' \in V} \|L - L'\|_\infty.$$

In this instance we are working with 1-by-d matrices, i.e., elements of $(\mathbb{R}^d)^*$.

Note that Definition 2.6 is independent of the choice of Θ. Indeed, by basic linear algebra $\Theta^*((\mathbb{R}^u)^*) = (\ker \Theta)^0$, where $(\ker \Theta)^0$ is the annihilator of $\ker \Theta$ (see Section 1B). By Lemma 2.3, $\ker \Theta$ is independent of the choice of Θ, and therefore so is $(\ker \Theta)^0$.

Example 2.7. Suppose that, as a matrix,

$$L := (1 \quad -\sqrt{2} \quad -1+\sqrt{2})$$

as in Example 1.5. Then L is purely irrational, i.e., $u = 0$, since its coefficients are not all in rational ratio. Therefore $A_L(\tau_1, \tau_2)$ is equal to

$$\inf_{k \in \mathbb{R}: \tau_1 \leqslant |k| \leqslant \tau_2^{-1}} \max(\|k\|_{\mathbb{R}/\mathbb{Z}}, \|-k\sqrt{2}\|_{\mathbb{R}/\mathbb{Z}}, \|-k+k\sqrt{2}\|_{\mathbb{R}/\mathbb{Z}}).$$

As we said above, we seek an upper bound on $A_L(\tau_1, \tau_2)^{-1}$. To this end, we claim that, for this particular L and for all $\tau_1, \tau_2 \in (0, 1]$, one has

$$A_L(\tau_1, \tau_2) \gg \min(\tau_1, \tau_2).$$

Indeed, we know that, for all $q \in \mathbb{N}$,

$$\|q\sqrt{2}\|_{\mathbb{R}/\mathbb{Z}} \geqslant \frac{1}{10q}.$$

This is the statement that $\sqrt{2}$ is a badly approximable irrational. The proof is straightforward: if there were some natural number p for which $|q\sqrt{2} - p| < 1/(10q)$, then

$$1 \leqslant |2q^2 - p^2| < \frac{\sqrt{2}}{10} + \frac{p}{10q} < \frac{\sqrt{2}}{5} + \frac{1}{10},$$

which is a contradiction.

Suppose first that $\|k\|_{\mathbb{R}/\mathbb{Z}} \leqslant \tau_2/100$ and $\frac{1}{2} \leqslant |k| \leqslant \tau_2^{-1}$. Then, replacing k by $[k]$ (the nearest integer to k), we can conclude that

$$\max(\|-k\sqrt{2}\|_{\mathbb{R}/\mathbb{Z}}, \|-k+k\sqrt{2}\|_{\mathbb{R}/\mathbb{Z}}) \geqslant \|[k]\sqrt{2}\|_{\mathbb{R}/\mathbb{Z}} - \frac{\tau_2}{50} \geqslant \frac{1}{10[k]} - \frac{\tau_2}{50} \geqslant \frac{1}{10\tau_2^{-1}+10} - \frac{\tau_2}{50} \gg \tau_2.$$

Otherwise, one has

$$\|k\|_{\mathbb{R}/\mathbb{Z}} \gg \min(\tau_1, \tau_2).$$

Therefore,

$$A_L(\tau_1, \tau_2) \gg \min(\tau_1, \tau_2)$$

as claimed.

It is not too difficult to show that if L is an m-by-d matrix with rank m and with algebraic coefficients, then

$$A_L(\tau_1, \tau_2) \gg_L \min(\tau_1, \tau_2^{O_L(1)}), \tag{2-3}$$

where the $O_L(1)$ term in the exponent depends on the algebraic degree of the coefficients of L.[7] We shall give a proof of this statement in Appendix E. In general, however, $A_L(\tau_1, \tau_2)$ could tend to zero arbitrarily quickly as τ_2 tends to zero, for example in the case when $L = (1 \ \ -\lambda \ \ -1+\lambda)$ and λ is a Liouville number (an irrational number that may be very well-approximated by rationals).

Yet, however fast $A_L(\tau_1, \tau_2)$ decays, we have the following critical claim.

Claim 2.8. For all permissible choices of L, τ_1 and τ_2 in Definition 2.6, $A_L(\tau_1, \tau_2)$ is positive.

Proof. Let u be the rational dimension of L. Without loss of generality we may assume that $u \leqslant m-1$. Then, for all $\varphi \in (\mathbb{R}^m)^* \setminus \Theta^*((\mathbb{R}^u)^*)$ we have that $\operatorname{dist}(L^*\varphi, (\mathbb{Z}^d)^T) > 0$. (If this were not the case then the map $(\Theta, \varphi) : \mathbb{R}^m \to \mathbb{R}^{u+1}$ would contradict the definition of u.) Therefore, as the definition of $A_L(\tau_1, \tau_2)$ involves taking the infimum of a positive continuous function over a compact set, $A_L(\tau_1, \tau_2)$ is positive. □

The expression $A_L(\tau_1, \tau_2)^{-1}$ will appear in the error term of our main theorem; Claim 2.8 shows that such an error term still has content.

2C. ***Nondegeneracy.*** In the statement of Theorem 1.2, which we remind the reader was the result of Green and Tao that used Gowers norms to control the number of solutions to linear equations with integer coefficients, one recalls that there were certain linear-algebraic notions of nondegeneracy for the matrix L. These concerned the rank of L and the properties of its row space. In the setting of diophantine inequalities it will transpire that the same notions of nondegeneracy are important — this much was obvious from the statement of Corollary 1.4 — except that, in order to control the error terms when L depends on N, one must assume that L is not even "approximately" degenerate.

In order to make these notions precise, we will first give some names to the sets of degenerate maps that we wish to avoid.

Definition 2.9 (low-rank variety). Let m, d be natural numbers satisfying $d \geqslant m+1$. Let $V_{\mathrm{rank}}(m, d)$ denote the set of all linear maps $L : \mathbb{R}^d \to \mathbb{R}^m$ whose rank is less than m. We call $V_{\mathrm{rank}}(m, d)$ the *low-rank variety*.

[7] One could perhaps remove this dependence by using the Schmidt subspace theorem, though as there are power losses throughout the rest of the argument there does not seem to be a great advantage in doing so.

Let $V^{\text{unif}}_{\text{rank}}(m,d)$ denote the set of all linear maps $L:\mathbb{R}^d\to\mathbb{R}^m$ for which there exists a standard basis vector of $\mathbb{R}^d$, say $\boldsymbol{e_i}$, for which $L|_{\text{span}(\boldsymbol{e_j}:j\neq i)}$ has rank less than m. We call $V^{\text{unif}}_{\text{rank}}(m,d)$ the *uniform low-rank variety*.

We make the trivial observation that $V^{\text{unif}}_{\text{rank}}(m,d)$ contains $V_{\text{rank}}(m,d)$. For certain technical reasons it will be much more convenient to work with matrices $L\notin V^{\text{unif}}_{\text{rank}}(m,d)$, as opposed to merely working with matrices $L\notin V_{\text{rank}}(m,d)$, as we will be able to fix an arbitrary coordinate and still be left with a full rank linear map.

Definition 2.10 (dual degeneracy variety). Let m,d be natural numbers satisfying $d\geqslant m+2$. Let $\boldsymbol{e_1},\dots,\boldsymbol{e_d}$ denote the standard basis vectors of $\mathbb{R}^d$, and let $\boldsymbol{e_1^*},\dots,\boldsymbol{e_d^*}$ denote the dual basis of $(\mathbb{R}^d)^*$. Then let $V^*_{\text{degen}}(m,d)$ denote the set of all linear maps $L:\mathbb{R}^d\to\mathbb{R}^m$ for which there exist two indices $i,j\leqslant d$, and some real number λ, such that $\boldsymbol{e_i^*}-\lambda\boldsymbol{e_j^*}$ is nonzero and $(\boldsymbol{e_i^*}-\lambda\boldsymbol{e_j^*})\in L^*((\mathbb{R}^m)^*)$. We call $V^*_{\text{degen}}(m,d)$ the *dual degeneracy variety*.

It may be easily verified that this definition does nothing more than rephrase the condition that appeared in the statements of Corollary 1.4 and Theorem 1.2 concerning the row-space of a degenerate map L, namely that there exists a nonzero row-vector in the row-space of L that has two or fewer nonzero coordinates. The formulation in terms of dual spaces will be particularly convenient, however, for some of the algebraic manipulations in Section 5. This is the reason why we use the term "dual" in the name of $V^*_{\text{degen}}(m,d)$.[8]

Having introduced $V_{\text{rank}}(m,d)$, $V^{\text{unif}}_{\text{rank}}(m,d)$ and $V^*_{\text{degen}}(m,d)$, we can articulate the relationship between the nondegeneracy conditions in Theorem 1.2 (for linear equations given by L) and our nondegeneracy conditions in Theorem 2.12 below (for linear inequalities given by L). Indeed, for equations, L is nondegenerate if

$$L\notin V_{\text{rank}}(m,d)\quad\text{or}\quad L\notin V^*_{\text{degen}}(m,d). \tag{2-4}$$

For inequalities, L is nondegenerate if

$$\text{dist}(L,V^{\text{unif}}_{\text{rank}}(m,d))\geqslant c\quad\text{or}\quad\text{dist}(L,V^*_{\text{degen}}(m,d))\geqslant c', \tag{2-5}$$

for some fixed parameters c and c'. One can see immediately how the conditions for inequalities are "approximate" versions of the conditions for equations.

Example 2.11. It may be instructive to consider a matrix such as

$$L=\begin{pmatrix}1+N^{-1} & \sqrt{3}+N^{-\frac{1}{2}} & \pi & -\pi+\sqrt{2}\\ 2 & 2\sqrt{3}+N^{-\frac{1}{2}} & -\sqrt{5} & e\end{pmatrix}.$$

We observe that L has rank 2 and $L\notin V^*_{\text{degen}}(2,4)$. If one knew Theorem 1.2 and the conditions (2-4), then one might perhaps have hoped to apply the theory of Gowers norms to bound the number of solutions

[8] Later on (in Definition 4.4) we will have a set of degenerate maps $V_{\text{degen}}(d-m,d)$ which will parametrise the kernel of maps in $V^*_{\text{degen}}(m,d)$. Since these maps feel somewhat dual to those maps in $V^*_{\text{degen}}(m,d)$, we will come to call $V_{\text{degen}}(d-m,d)$ the "degeneracy variety".

to inequalities given by L. However, by considering perturbations of the first two columns, we see that $\operatorname{dist}(L, V^*_{\text{degen}}(2,4)) = o(1)$ as $N \to \infty$. Indeed, one may perturb L by $O(N^{-1/2})$ such that there is a vector $(0, 0, x_3, x_4)$ in the row space. So, despite the fact that L is nondegenerate from the point of view of equations, L *is* degenerate from the point of view of inequalities and the conditions (2-5). Thus, our main theorem on inequalities will not apply to this L.

Furthermore, we have another result (Theorem 2.14 below) which shows that one cannot possibly use Gowers norms to control inequalities given by such an L. Therefore, whatever methods we use to prove Theorem 2.12, these methods must necessarily break down when applied to this example.

2D. ***The main theorem and a partial converse.*** Having laid the groundwork, we may now state the main theorem of this paper.

Theorem 2.12 (Main Theorem). *Let m, d be natural numbers, satisfying $d \geqslant m+2$, and let ε, c, C, C' be positive reals. Let N be an integer parameter and let $L = L(N) : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map that satisfies $\|L\|_\infty \leqslant C$. Let $A_L : (0,1] \times (0,1] \to (0,\infty)$ be the approximation function of L. Suppose further that $\operatorname{dist}(L, V^{\text{unif}}_{\text{rank}}(m,d)) \geqslant c$, that $\operatorname{dist}(L, V^*_{\text{degen}}(m,d)) \geqslant c'$, and that L has rational complexity at most C'. Then there exists a natural number s at most $d-2$, independent of ε, such that the following is true. Let $F : \mathbb{R}^d \to [0,1]$ be the indicator function of $[1,N]^d$, and let $G : \mathbb{R}^m \to [0,1]$ be the indicator function of a convex domain contained in $[-\varepsilon,\varepsilon]^m$. Let $f_1, \dots, f_d : [N] \to [-1,1]$ be arbitrary functions, and suppose that*

$$\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho,$$

for some parameter ρ in the range $0 < \rho \leqslant 1$. Then

$$T^L_{F,G,N}(f_1, \dots, f_d) \ll_{c,c',C,C',\varepsilon} \rho^{\Omega(1)} + o_{\rho,A_L,c,c',C,C'}(1) \tag{2-6}$$

as $N \to \infty$. The $o_{\rho,A_L,c,c',C,C'}(1)$ term may be bounded above by

$$N^{-\Omega(1)}\rho^{-O(1)}A_L(\Omega_{c,c',C,C'}(1), \rho)^{-1}.$$

We remind the reader that the implied constants may depend on the dimensions m and d. Also note that in the above statement one may replace C and C' by a single constant C, and c and c' by a single constant c, without weakening the conclusion. We proceed with this assumption.

Let us note some consequences of this theorem. Firstly, since $A_L(\Omega_{c,C}(1), \rho)^{-1}$ is finite (by Claim 2.8), Theorem 2.12 immediately implies Corollary 1.4 (this was the qualitative statement around which we structured Section 1). Hence Theorem 2.12 also implies all the other corollaries from Section 1. Secondly, from (2-3), or rather from our full quantitative version in Lemma E.1, we have another corollary for matrices L with algebraic coefficients.

Corollary 2.13 (inequalities with algebraic coefficients). *Assume the same hypotheses as Theorem 2.12, and assume further that L has algebraic coefficients with algebraic degree at most k. Let H denote the*

maximum absolute value of all of the coefficients of all of the minimal polynomials of the coefficients of L. Then

$$T^L_{F,G,N}(f_1, \dots, f_d) \ll_{c,C,\varepsilon,H} \rho^{\Omega(1)} + N^{-\Omega(1)}\rho^{-O_k(1)}.$$

The reader may wonder how the implied constant in these statements depends on ε. Ultimately the implied constant in (2-6) tends to infinity as ε tends to zero, as our approximation argument in Section 6 will not be efficient in powers of ε. Yet, to prevent our notation becoming too unreadable, we choose not to keep track of the precise behaviour of implied constants involving ε.

As we remarked in Section 2C and Example 2.11, we can also prove a partial converse to Theorem 2.12. This result demonstrates that the nondegeneracy condition $\operatorname{dist}(L, V^*_{\text{degen}}(m, d)) \geqslant c$ is necessary in order to use Gowers norms to control inequalities given by L.

Theorem 2.14. *Let m, d be natural numbers, satisfying $d \geqslant m+2$, and let ε, c, C be positive constants. For each natural number N, let $L = L(N) : \mathbb{R}^d \to \mathbb{R}^m$ be a linear map satisfying $\|L\|_\infty \leqslant C$. Let $F : \mathbb{R}^d \to [0,1]$ denote the indicator function of $[1,N]^d$ and $G : \mathbb{R}^m \to [0,1]$ denote the indicator function of $[-\varepsilon,\varepsilon]^m$. Assume further that $\operatorname{dist}(L, V_{\text{rank}}(m,d)) \geqslant c$ and that $T^L_{F,G,N}(1,\dots,1) \gg_{c,C,\varepsilon} 1$ for large enough N.*

Suppose that

$$\liminf_{N\to\infty} \operatorname{dist}(L, V^*_{\text{degen}}(m,d)) = 0.$$

Let s be a natural number, let $H : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ be any function satisfying $H(\rho) = \kappa(\rho)$, and for each N let $E_\rho(N)$ denote some error term depending on a parameter ρ and satisfying $E_\rho(N) = o_\rho(1)$ as $N \to \infty$. Then one can find infinitely many natural numbers N such that there exist functions $f_1, \dots, f_d : [N] \to [-1,1]$ and some ρ at most 1 such that both

$$\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho$$

and

$$|T^L_{F,G,N}(f_1,\dots,f_d)| > H(\rho) + E_\rho(N). \tag{2-7}$$

In other words, the conclusion of Theorem 2.12 cannot possibly hold if $\operatorname{dist}(L, V^*_{\text{degen}}(m,d))$ is arbitrarily close to 0, even if one replaces the $\rho^{\Omega(1)}$ dependence in (2-6) with a function $H(\rho)$ that could potentially decay to zero arbitrarily slowly as ρ tends to zero.

The proof of Theorem 2.14 is contained in Section 9, which can be read independently of the rest of the paper.

2E. ***The proof strategy.*** All the corollaries from Sections 1 and 2 are implied by Theorem 2.12, so our remaining task is to prove this theorem. Speaking somewhat informally, we wish to bound $T^L_{F,G,N}(f_1,\dots,f_d)$ in terms of some Gowers norms $\|f_j\|_{U^{s+1}[N]}$ when the functions F and G are the indicator functions of certain convex domains. Now, one might expect the proof to be easier if, instead, F and G were functions with nicer analytic properties — Lipschitz functions, for example. This is indeed the case, and thus our proof splits naturally into two parts. The first part, contained in Sections 3 and 5, reduces Theorem 2.12

to a similar statement in which the functions F and G are Lipschitz — this will be Theorem 5.6. The second part of the paper is devoted to proving Theorem 5.6. For the rest of this subsection we will try to articulate the strategies for each part, and to elucidate the main technical difficulties.

In [Green and Tao 2010a], replacing convex cutoffs with Lipschitz cutoffs was an easy operation, accomplished in a couple of pages in Appendices A and C of that paper. Somewhat surprisingly, this part turns out to be the trickiest element in the setting of inequalities, at least when L is not purely irrational.

Replacing F with a Lipschitz cutoff is no issue, but the difficulty comes from replacing G. Consider the example

$$L := \begin{pmatrix} 1 & 0 & -\sqrt{2} & -\sqrt{3}+1 \\ 0 & 1 & 5\sqrt{2} & 5\sqrt{3} \end{pmatrix}$$

from Section 2A, in which we established that L has rational dimension 1 and that

$$L(\mathbb{Z}^4) \subset \left\{ \boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x} \cdot \begin{pmatrix} 5 \\ 1 \end{pmatrix} \in \mathbb{Z} \right\}.$$

Take G to be the indicator of the compact convex domain

$$\left\{ \boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_\infty \leqslant 10,\ -1 \leqslant \boldsymbol{x} \cdot \begin{pmatrix} 5 \\ 1 \end{pmatrix} \leqslant 1 \right\}.$$

Then

$$T^L_{F,G,N}(f_1, \dots, f_4) = \frac{1}{N^2} \sum_{\substack{\boldsymbol{n} \in \mathbb{Z}^4 \\ \|L\boldsymbol{n}\|_\infty \leqslant 10 \\ (5\ 1)L\boldsymbol{n} = -1,0,1}} \left(\prod_{j=1}^4 f_j(n_j) \right) F(\boldsymbol{n}). \tag{2-8}$$

To replace the convex cutoff G by a Lipschitz cutoff, a natural approach is to take a Lipschitz function $\tilde{G}$ that is a minorant for G,[9] with $\tilde{G}$ supported on the set

$$\left\{ \boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_\infty \leqslant 10 - \delta, \quad -1 + \delta \leqslant \boldsymbol{x} \cdot \begin{pmatrix} 5 \\ 1 \end{pmatrix} \leqslant 1 - \delta \right\}$$

for some small positive parameter δ, and $\tilde{G}$ identically equal to 1 on the set

$$\left\{ \boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_\infty \leqslant 10 - 2\delta, \quad -1 + 2\delta \leqslant \boldsymbol{x} \cdot \begin{pmatrix} 5 \\ 1 \end{pmatrix} \leqslant 1 - 2\delta \right\}.$$

One has $\|G - \tilde{G}\|_1 = \kappa(\delta)$, so one might hope that for any functions $f_1, \dots, f_4$ one would have

$$|T^L_{F,G,N}(f_1, \dots, f_4) - T^L_{F,\tilde{G},N}(f_1, \dots, f_4)| = \kappa(\delta). \tag{2-9}$$

However, no matter how small we choose δ,

$$T^L_{F,\tilde{G},N}(f_1, \dots, f_4) \approx \frac{1}{N^2} \sum_{\substack{\boldsymbol{n} \in \mathbb{Z}^4 \\ \|L\boldsymbol{n}\|_\infty \leqslant 10-\delta \\ (5\ 1)L\boldsymbol{n} = 0}} \left(\prod_{j=1}^4 f_j(n_j) \right) F(\boldsymbol{n}). \tag{2-10}$$

[9] One also finds a majorant Lipschitz function, but that won't feature in this example.

In moving from G to $\tilde{G}$ the range of summation for $\boldsymbol{n}$ between expressions (2-8) and (2-10) has been cut by a factor of two-thirds! Thus we have no reason to expect that (2-9) should hold for all functions $f_1, \dots, f_4$.

We circumvent these difficulties by employing the following idea. Rather than replacing G with a Lipschitz cutoff straight away, when faced with an expression such as (2-8) we can perform some initial reparametrisation, observing that there is a linear map $\Xi : \mathbb{R}^3 \to \mathbb{R}^4$ with integer coefficients which gives a lattice parametrisation of those $\boldsymbol{n} \in \mathbb{Z}^4$ for which $(5\ \ 1)L\boldsymbol{n} = 0$, namely

$$\Xi \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} = \begin{pmatrix} m_1 \\ -5m_1 - 5m_2 \\ m_3 \\ m_2 \end{pmatrix}.$$

Moreover, $\boldsymbol{n} \in \mathbb{Z}^4$ with $(5\ \ 1)L\boldsymbol{n} = \pm 1$ if and only if there are integers m_1, m_2, m_3 for which

$$\boldsymbol{n} = \Xi \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} + \begin{pmatrix} 0 \\ \pm 1 \\ 0 \\ 0 \end{pmatrix}.$$

This enables us to decompose $T^L_{F,G,N}(f_1, \dots, f_4)$ into three separate expressions, each of the form

$$\frac{1}{N^2} \sum_{\boldsymbol{m} \in \mathbb{Z}^3} \left(\prod_{j=1}^{4} f_j(\Xi(\boldsymbol{m})_j + \tilde{\boldsymbol{r}}_j) \right) F(\Xi(\boldsymbol{m}) + \tilde{\boldsymbol{r}}) 1_{[-10,10]^2}(L(\Xi\boldsymbol{m} + \tilde{\boldsymbol{r}})) \tag{2-11}$$

for some different vector $\tilde{\boldsymbol{r}} \in \mathbb{Z}^4$, where $\Xi(\boldsymbol{m})_j$ denotes the j-th coordinate of $\Xi(\boldsymbol{m})$. Now, replace the convex cutoff function $1_{[-10,10]^2}$ with some Lipschitz minorant $\tilde{G}$ which is supported on $[-10+\delta, 10-\delta]^2$ and equal to 1 on $[-10+2\delta, 10-2\delta]^2$, in each of the three expressions (2-11) separately. Then the size of these expressions *will* stay roughly constant.

To quantify this step, the approximation function A_L enters the picture. Indeed, if $\tilde{\boldsymbol{r}} = \boldsymbol{0}$ the error term introduced by applying such an approximation to (2-11) is bounded above by

$$T^{L\Xi}_{F,G^*,N}(1, \dots, 1),$$

where G^* is some other Lipschitz function supported on

$$\{\boldsymbol{x} \in \mathbb{R}^2 : 10 - 2\delta \leqslant \|\boldsymbol{x}\|_\infty \leqslant 10 + 2\delta\}.$$

Finding an upper bound on expressions such as $T^{L\Xi}_{F,G^*,N}(1, \dots, 1)$ is exactly the endeavour we discussed in Section 2B, when motivating the introduction of the approximation function A_L. The only difference is that now we are dealing with the function $A_{L\Xi}$, rather than A_L.

It turns out that the map $L\Xi$ is most naturally viewed as a map from $\mathbb{R}^3$ into a one dimensional space, i.e.,

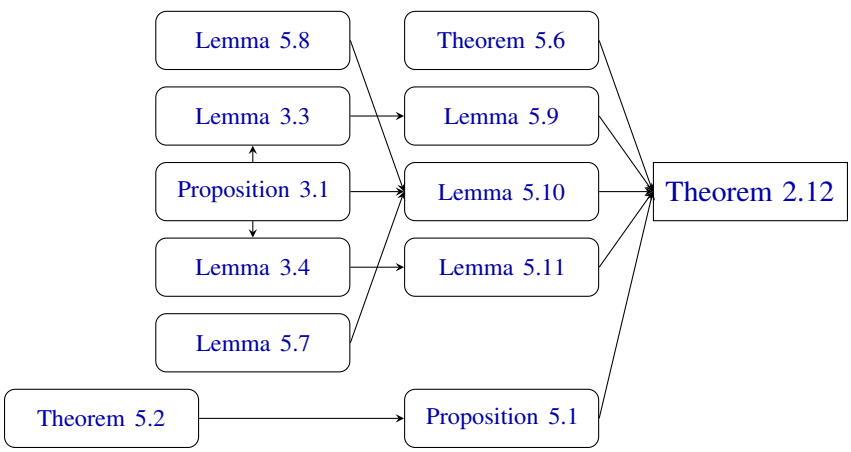
$$L\Xi : \mathbb{R}^3 \to \left\{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x} \cdot \begin{pmatrix} 5 \\ 1 \end{pmatrix} = 0\right\},$$

whereas L maps from $\mathbb{R}^4$ to $\mathbb{R}^2$. This is the "dimension reduction" which gives Section 5 its name.

The reader will have noticed that, after replacing $1_{[-10,10]^2}$ with the Lipschitz function $\tilde{G}$, expression (2-11) is not equal to an expression of the form $T^L_{F,\tilde{G},N}(f_1, \dots, f_4)$, since the map Ξ and the shift $\tilde{\boldsymbol{r}}$ are now both on the scene. This complicates matters in the second half of the proof, and thus Theorem 5.6 will not be exactly the same statement as Theorem 2.12 apart from the Lipschitz cutoffs. Rather, Theorem 5.6 will bound an object that we will come to denote by $T^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1, \dots, f_d)$, which will be a general version of expression (2-11). The reader may consult Definition 5.3 for the full definition.

In order to make this argument rigorous we will have to verify that in replacing the map L with the map $L\Xi$ we haven't introduced any extra rational relations;[10] to work out how to relate A_L and $A_{L\Xi}$; and to work out how to identify a suitable Ξ in the general case. Furthermore, we will have to carry the nondegeneracy relations (such as $\operatorname{dist}(L, V^*_{\text{degen}}(m, d)) \geqslant c$) through this reparametrisation by Ξ, and then establish what the new nondegeneracy notions should be for the pairs $(\Xi, L\Xi)$. This is all done in the (somewhat alarming) Lemma 5.10, which has 9 parts. The upper bounds on expressions like $T^{L\Xi}_{F,G^*,N}(1, \dots, 1)$ are established earlier, in Lemma 3.4, with everything combined at the end of Section 5.

The diagram of the dependency of the various lemmas — excluding those which are found in Appendices A, B and D, which are somewhat standard — is as follows:



It remains to resolve Theorem 5.6, and it turns out that this second part of the proof is significantly more straightforward than the first. In particular neither the statement of Theorem 5.6 nor its proof make any reference to the rational dimension of L nor to the approximation function A_L.

[10] This is essentially the statement that $L\Xi$ should be purely irrational.

The idea is as follows. For a function $f : [N] \to [-1, 1]$ and a small parameter η, let $\tilde{f} : \mathbb{R} \to [-1, 1]$ denote the function

$$\tilde{f}(x) = \begin{cases} f(n) & \text{if } |n - x| \leqslant \eta, \\ 0 & \text{otherwise,} \end{cases}$$

i.e., $\tilde{f}$ is a "fattened" version of f. Then, for Lipschitz functions F and G, let

$$\tilde{T}^L_{F,G,N}(\tilde{f}_1, \dots, \tilde{f}_d) := \frac{1}{N^{d-m}} \int_{\boldsymbol{x} \in \mathbb{R}^d} \Big(\prod_{j=1}^{d} \tilde{f}_j(x_j) \Big) F(\boldsymbol{x}) G(L\boldsymbol{x})\, d\boldsymbol{x}$$

represent the "real solution density" for the inequality weighted by the functions $\tilde{f}_j$. The expression $\tilde{T}^L_{F,G,N}(\tilde{f}_1, \dots, \tilde{f}_d)$ is more convenient to work with than $T^L_{F,G,N}(f_1, \dots, f_d)$, as we are now working in a setting in which the coefficients of L are invertible.[11]

The expression $\tilde{T}^L_{F,G,N}(\tilde{f}_1, \dots, \tilde{f}_d)$ enjoys the following two properties. Firstly, it is closely related to $T^L_{F,G,N}(f_1, \dots, f_d)$. Indeed, just by expanding out the definition of $\tilde{f}_j$, we see that

$$\begin{aligned} \tilde{T}^L_{F,G,N}(\tilde{f}_1, \dots, \tilde{f}_d) &= \frac{1}{N^{d-m}} \sum_{\boldsymbol{n} \in \mathbb{Z}^d} \Big(\prod_{j=1}^{d} f_j(n_j) \Big) \int_{\boldsymbol{y} \in \mathbb{R}^d} F(\boldsymbol{y}) G(L\boldsymbol{y}) 1_{[-\eta,\eta]^d}(\boldsymbol{y} - \boldsymbol{n})\, d\boldsymbol{y} \\ &\approx \frac{1}{N^{d-m}} \sum_{\boldsymbol{n} \in \mathbb{Z}^d} \Big(\prod_{j=1}^{d} f_j(n_j) \Big) F(\boldsymbol{n}) G(L\boldsymbol{n}) \int_{\boldsymbol{y} \in \mathbb{R}^d} 1_{[-\eta,\eta]^d}(\boldsymbol{y} - \boldsymbol{n})\, d\boldsymbol{y} \\ &\approx (2\eta)^d T^L_{F,G,N}(f_1, \dots, f_d), \end{aligned} \tag{2-12}$$

by using the Lipschitz properties of F and G to replace $F(\boldsymbol{y})$ and $G(L\boldsymbol{y})$ by $F(\boldsymbol{n})$ and $G(L\boldsymbol{n})$ respectively. This analysis is performed rigorously in Section 6, and is the only place in the proof where the Lipschitz property of G is used.

Secondly, $\tilde{T}^L_{F,G,N}(\tilde{f}_1, \dots, \tilde{f}_d)$ may be bounded above by expressions involving the Gowers norms (over the reals) of the functions $\tilde{f}_j$. Indeed, after some small manipulations using the compact support of G, one ends up with the bound

$$|\tilde{T}^L_{F,G,N}(\tilde{f}_1, \dots, \tilde{f}_d)| \ll_G \frac{1}{N^{d-m}} \int_{\substack{\boldsymbol{x} \in \mathbb{R}^d \\ L\boldsymbol{x} = \boldsymbol{0}}} \Big(\prod_{j=1}^{d} \tilde{f}_j(x_j) \Big) F(\boldsymbol{x})\, d\mu(\boldsymbol{x}), \tag{2-13}$$

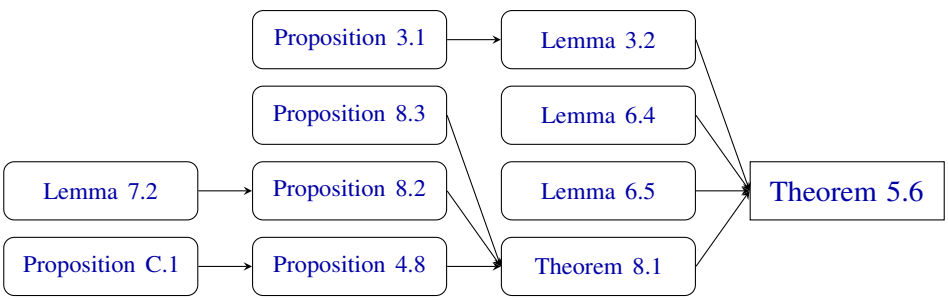
where $\mu(\boldsymbol{x})$ is a suitable measure supported on $\ker L$. The reader will then notice that the right-hand side of (2-13) bears a structural similarity to the expression considered in Theorem 1.2 above, i.e., in the generalised von Neumann theorem for equations with integer coefficients. One may then rejig Green and Tao's proof of Theorem 1.2 to apply in this setting, and thereby bound $\tilde{T}^L_{F,G,N}(\tilde{f}_1, \dots, \tilde{f}_d)$ by the Gowers norms of the functions $\tilde{f}_j$. This is done in Section 8. Finally, there is an elementary argument (Lemma 6.5) that relates the Gowers norms of the functions $\tilde{f}_j$ to the Gowers norms of the original functions f_j, thus completing the proof of our result.

[11] This manoeuvre is somewhat analogous to the device used by Green and Tao [2010a] of passing from $[N]$ to some cyclic group $\mathbb{Z}/N'\mathbb{Z}$, where N' is a prime number larger than N.

As will be familiar to readers of [Green and Tao 2010a], the key manoeuvre in analysing (2-13) is parametrising $\ker L$ in a certain special way (in *normal form*, see Section 4), in order to facilitate repeated applications of the Cauchy–Schwarz inequality. When working over the reals, maintaining quantitative control over the size of the coefficients after this reparametrisation is no longer trivial, and requires the assumption that $\operatorname{dist}(L, V^*_{\mathrm{degen}}(m,d)) \geqslant c$. The details of this piece of quantitative linear algebra are given in Proposition 4.8 and Appendix C. It is this part of the argument which would break down were one to attempt to use Gowers norms to bound inequalities such as the one given by the matrix L in Example 2.11.

We have already remarked that Theorem 5.6 does not just concern the objects $T^{L}_{F,G,N}(f_1,\dots,f_d)$ but actually concerns the more general objects $T^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1,\dots,f_d)$, which are similar to (2-11). This adds an extra veneer of complication, centred largely around the notion of degeneracy for the pair of maps $(\Xi, L\Xi)$. Matters are resolved by a linear algebra argument in Section 7, relating different notions of degeneracy.

The diagram of the dependency of the lemmas used in the proof of Theorem 5.6 is as follows:



The appendices contain some extra material which we felt to be best kept apart from the main narrative. In the case of the first two appendices, they comprise standard results from the literature on Gowers norms and Lipschitz functions, which we include to assist any readers who are unfamiliar with these topics. In the case of Appendices C and D, we present a handful of arguments of a linear algebraic nature which, though perhaps not already present in the literature in the exact form we require, are nonetheless easy to establish. Finally, Appendix E concerns the analysis of the approximation function A_L when L has algebraic coefficients. This argument has a similar flavour to Example 2.7, and is included for the sake of completeness.

3. Upper bounds

This section is devoted to proving three upper bounds on the expression $T^{L}_{F,G,N}(1,\dots,1)$. For the definition of this quantity, the reader may refer to Definition 2.1.

The following proposition, which represents a quantitative version of the "row-rank equals column-rank" principle, will be useful throughout.

Proposition 3.1 (rank matrix). *Let m, d be natural numbers, with $d \geqslant m+1$. Let c, C be positive constants. Then there are positive constants $D_{c,C}, D'_{c,C}$ for which the following holds. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective*

linear map, denoted by matrix $(\lambda_{ij})_{i\leqslant m, j\leqslant d}$, *and assume that* $\|L\|_\infty \leqslant C$ *and* $\operatorname{dist}(L, V_{\mathrm{rank}}(m,d)) \geqslant c$. *Then there exists a matrix* M *that is an* m*-by-*m *submatrix of* L *and enjoys the following properties*:

(1) $|\det M| \geqslant D_{c,C}$.

(2) $\|M^{-1}\|_\infty \leqslant D'_{c,C}$.

We call such a matrix M *a* rank matrix *of* L. *Furthermore*:

(3) *Let* $\boldsymbol{v} \in \mathbb{R}^d$ *be a vector such that* $\boldsymbol{v}^T$ *is in the row-space of* L, *and suppose that* $\|\boldsymbol{v}\|_\infty \leqslant C_1$ *for some positive constant* C_1. *Then for* i *in the range* $1 \leqslant i \leqslant m$ *there exist coefficients* a_i *satisfying* $|a_i| = O_{c,C,C_1}(1)$ *such that* $\sum_{i=1}^m a_i \lambda_{ij} = v_j$ *for all* j *in the range* $1 \leqslant j \leqslant d$.

Finally:

(4) *If* L *satisfies the stronger hypothesis* $\operatorname{dist}(L, V^{\mathrm{unif}}_{\mathrm{rank}}(m,d)) \geqslant c$, *then, for each* j, *there exists a rank matrix of* L *that does not include the* j*-th column of* L.

We defer the proof to Appendix C.

Our first upper bound is exceptionally crude, but will nonetheless be useful in Section 6.

Lemma 3.2. *Let* N, m, d *be natural numbers, satisfying* $d \geqslant m+1$, *and let* c, C, ε *be positive constants. Let* $L : \mathbb{R}^d \to \mathbb{R}^m$ *be a surjective linear map, and suppose that* $\|L\|_\infty \leqslant C$ *and* $\operatorname{dist}(L, V_{\mathrm{rank}}(m,d)) \geqslant c$. *Let* $F : \mathbb{R}^d \to [0,1]$ *and* $G : \mathbb{R}^m \to [0,1]$ *be two functions, with* F *supported on* $[-N, N]^d$ *and* G *supported on* $[-\varepsilon, \varepsilon]^m$. *Then*

$$T^L_{F,G,N}(1, \dots, 1) \ll_{c,C,\varepsilon} \|G\|_\infty.$$

Proof. Let M be a rank matrix of L (Proposition 3.1), and suppose without loss of generality that M consists of the first m columns of L. For j in the range $m+1 \leqslant j \leqslant d$, let the vector $\boldsymbol{v_j} \in \mathbb{R}^m$ be the j-th column of the matrix $M^{-1}L$. Then $N^{d-m} T^L_{F,G,N}(1, \dots, 1) \leqslant \|G\|_\infty \cdot Z$, where Z is the number of solutions to

$$\begin{pmatrix} n_1 \\ \vdots \\ n_m \end{pmatrix} + \sum_{j=m+1}^{d} \boldsymbol{v_j} n_j \in M^{-1}([-\varepsilon, \varepsilon]^m)$$

in which $n_1, \dots, n_d$ are integers that satisfy $|n_1|, \dots, |n_d| \leqslant N$. Fixing a choice of the variables $n_{m+1}, \dots, n_d$ forces the vector $(n_1, \dots, n_m)^T$ to lie in a convex region of diameter $O_{c,C,\varepsilon}(1)$. There are at most $O_{c,C,\varepsilon}(1)$ such points, so $Z \ll_{c,C,\varepsilon} N^{d-m}$. The claimed bound follows. □

Our second estimate is a slight strengthening of the above, albeit under stronger hypotheses.

Lemma 3.3. *Let* N, m, d *be natural numbers, with* $d \geqslant m+1$, *and let* c, C, ε *be positive constants. Let* $L : \mathbb{R}^d \to \mathbb{R}^m$ *be a surjective linear map, and suppose that* $\|L\|_\infty \leqslant C$ *and* $\operatorname{dist}(L, V^{\mathrm{unif}}_{\mathrm{rank}}(m,d)) \geqslant c$. *Let* σ *be a real number in the range* $0 < \sigma < \frac{1}{2}$. *Let* $F : \mathbb{R}^d \to [0,1]$ *and* $G : \mathbb{R}^m \to [0,1]$ *be two functions, with* F *supported on*

$$\{\boldsymbol{x} \in \mathbb{R}^d : \operatorname{dist}(\boldsymbol{x}, \partial([1,N]^d)) \leqslant \sigma N\}$$

and G supported on $[-\varepsilon,\varepsilon]^m$. Then

$$T^L_{F,G,N}(1,\dots,1) \ll_{c,C,\varepsilon} \sigma \|G\|_\infty.$$

Proof. Without loss of generality, we may assume that F is supported on

$$\{\boldsymbol{x}\in\mathbb{R}^d : \|\boldsymbol{x}\|_\infty \leqslant 2N,\ |x_d-1|\leqslant \sigma N\} \quad \text{or} \quad \{\boldsymbol{x}\in\mathbb{R}^d : \|\boldsymbol{x}\|_\infty \leqslant 2N,\ |x_d-N|\leqslant \sigma N\}.$$

Consider the first case. By Proposition 3.1 there exists a rank matrix M that does not contain the column d. By reordering columns, we can assume without loss of generality that M consists of the first m columns of L. Continuing as in the proof of Lemma 3.2, for j in the range $m+1\leqslant j\leqslant d$, let the vector $\boldsymbol{v_j}\in\mathbb{R}^m$ be the j-th column of the matrix $M^{-1}L$. Then the expression $N^{d-m}T^L_{F,G,N}(1,\dots,1)$ may be bounded above by $\|G\|_\infty$ times the number of solutions to

$$\begin{pmatrix} n_1\\ \vdots\\ n_m\end{pmatrix} + \sum_{j=m+1}^{d} \boldsymbol{v_j} n_j \in M^{-1}([-\varepsilon,\varepsilon]^m)$$

satisfying $|n_1|,\dots,|n_{d-1}|\leqslant 2N$ and $|n_d|\leqslant \sigma N$. We conclude as in the previous proof.

In the second case, the relevant equation is

$$\begin{pmatrix} n_1\\ \vdots\\ n_m\end{pmatrix} + \sum_{j=m+1}^{d} \boldsymbol{v_j} n_j + (N-1)\boldsymbol{v_d} \in M^{-1}([-\varepsilon,\varepsilon]^m),$$

in which we count solutions satisfying $|n_1|,\dots,|n_{d-1}|\leqslant 2N$ and $|n_d-1|\leqslant \sigma N$. We conclude as in the previous proof. □

Our third estimate is more refined, and will be needed in Section 5 when we replace the sharp cutoff $1_{[-\varepsilon,\varepsilon]^m}$ with a Lipschitz cutoff. For the definition of the approximation function A_L, we refer the reader to Definition 2.6.

Lemma 3.4. *Let N, m, d be natural numbers, with $d\geqslant m+1$. Let c, C, ε be positive constants, and let σ_G be a parameter in the range $0<\sigma_G<\frac{1}{2}$. Suppose that $L:\mathbb{R}^d\to\mathbb{R}^m$ is a purely irrational surjective linear map, satisfying $\|L\|_\infty\leqslant C$ and $\operatorname{dist}(L, V_{\mathrm{rank}}(m,d))\geqslant c$. Let A_L denote the approximation function of L. Let $F:\mathbb{R}^d\to[0,1]$ be supported on $[-N,N]^d$, and let $G:\mathbb{R}^m\to[0,1]$ be a Lipschitz function, with Lipschitz constant $O(1/\sigma_G)$, supported on $[-\varepsilon,\varepsilon]^m$. Assume further that $\int_{\boldsymbol{x}} G(\boldsymbol{x})\,d\boldsymbol{x} = O_\varepsilon(\sigma_G)$. Then for all τ_2 in the range $0<\tau_2\leqslant 1$,*

$$T^L_{F,G,N}(1,\dots,1) \ll_{c,C,\varepsilon} \sigma_G + \frac{\tau_2^{1/2}}{\sigma_G} + \frac{\tau_2^{-O(1)} A_L(\Omega_{c,C}(1),\tau_2)^{-1}}{N}.$$

Proof. Following the proof of Lemma 3.2 verbatim, we arrive at the bound

$$T^L_{F,G,N}(1,\dots,1) \ll_{c,C,\varepsilon} \frac{1}{N^{d-m}} \sum_{\substack{n_{m+1},\dots,n_d\in\mathbb{Z}\\ |n_{m+1}|,\dots,|n_d|\leqslant N}} \tilde{G}\Big(\sum_{j=m+1}^{d} \boldsymbol{v}_j n_j\Big), \tag{3-1}$$

where $\boldsymbol{v}_j$ denotes the j-th column of the matrix $M^{-1}L$, and $\tilde{G}:\mathbb{R}^m\to[0,1]$ denotes the function

$$\tilde{G}(\boldsymbol{x}) = \sum_{\boldsymbol{a}\in\mathbb{Z}^m} (G\circ M)(\boldsymbol{a}+\boldsymbol{x}).$$

It remains to estimate the right-hand side of (3-1).

We may consider $\tilde{G}$ as a function on $\mathbb{R}^m/\mathbb{Z}^m$. With respect to the metric $\|\boldsymbol{x}\|_{\mathbb{R}^m/\mathbb{Z}^m}$, $\tilde{G}$ is Lipschitz with Lipschitz constant $O_{c,C,\varepsilon}(1/\sigma_G)$. Also,

$$\int_{\boldsymbol{x}\in[0,1)^m} \tilde{G}(\boldsymbol{x})\,d\boldsymbol{x} = \int_{\boldsymbol{x}\in\mathbb{R}^m} (G\circ M)(\boldsymbol{x})\,d\boldsymbol{x} = O_{c,C,\varepsilon}(\sigma_G).$$

By [Green and Tao 2008b, Lemma A.9], which we recall in Lemma B.3, for any X at least 2 we may write

$$\tilde{G}(\boldsymbol{x}) = \sum_{\substack{\boldsymbol{k}\in\mathbb{Z}^m\\ \|\boldsymbol{k}\|_\infty\leqslant X}} b_X(\boldsymbol{k}) e(\boldsymbol{k}\cdot\boldsymbol{x}) + O_{c,C,\varepsilon}\Big(\frac{\log X}{\sigma_G X}\Big), \tag{3-2}$$

where $b_X(\boldsymbol{k})\in\mathbb{C}$ and satisfies $|b_X(\boldsymbol{k})| = O(1)$. Moreover $b_X(\boldsymbol{0}) = \int_{\boldsymbol{x}\in[0,1)^m}\tilde{G}(\boldsymbol{x})\,d\boldsymbol{x}$.[12]

Returning to (3-1), we see that for any X at least 2 we may write

$$T^L_{F,G,N}(1,\dots,1) \ll_{c,C,\varepsilon} \sigma_G + \frac{\log X}{\sigma_G X} + X^{O(1)} \max_{\substack{\boldsymbol{k}\in\mathbb{Z}^m\\ 0<\|\boldsymbol{k}\|_\infty\leqslant X}} \Big(\prod_{j=m+1}^{d} \min(1, N^{-1}\|\boldsymbol{k}\cdot\boldsymbol{v}_j\|^{-1}_{\mathbb{R}/\mathbb{Z}})\Big), \tag{3-3}$$

where the final error term comes from summing over the arithmetic progressions $[-N,N]\cap\mathbb{Z}$.

It remains to relate the final error term of (3-3) to the approximation function A_L. Since L is purely irrational,

$$A_L(\tau_1,\tau_2) = \inf_{\substack{\varphi\in(\mathbb{R}^m)^*\\ \tau_1\leqslant\|\varphi\|_\infty\leqslant\tau_2^{-1}}} \operatorname{dist}(L^*\varphi, (\mathbb{Z}^d)^T).$$

[12] This final fact is not given explicitly in the statement of [Green and Tao 2008b, Lemma A.9], although it is given in the proof. In any case, it may be immediately deduced from (3-2), by letting X tend to infinity and integrating (3-2) over all $\boldsymbol{x}\in\mathbb{R}^m/\mathbb{Z}^m$.

Let τ_2 be in the range $0 < \tau_2 \leqslant 1$. Then there exist positive parameters D and D', depending only on c and C, for which

$$\begin{aligned}\min_{\substack{\boldsymbol{k}\in\mathbb{Z}^m\\ 0<\|\boldsymbol{k}\|_\infty\leqslant D\tau_2^{-1}}} \max(\{\|\boldsymbol{k}\cdot\boldsymbol{v_j}\|_{\mathbb{R}/\mathbb{Z}} : m+1\leqslant j\leqslant d\}) &= \min_{\substack{\boldsymbol{k}\in\mathbb{Z}^m\\ 0<\|\boldsymbol{k}\|_\infty\leqslant D\tau_2^{-1}}} \operatorname{dist}(\boldsymbol{k}^T M^{-1}L, (\mathbb{Z}^d)^T)\\ &\geqslant \inf_{\substack{\boldsymbol{k}\in\mathbb{R}^m\\ 1\leqslant\|\boldsymbol{k}\|_\infty\leqslant D\tau_2^{-1}}} \operatorname{dist}(\boldsymbol{k}^T M^{-1}L, (\mathbb{Z}^d)^T)\\ &\geqslant \inf_{\substack{\boldsymbol{k}\in\mathbb{R}^m\\ D'\leqslant\|\boldsymbol{k}\|_\infty\leqslant\tau_2^{-1}}} \operatorname{dist}(\boldsymbol{k}^T L, (\mathbb{Z}^d)^T)\\ &= A_L(D',\tau_2). \end{aligned} \tag{3-4}$$

Letting $X = D\tau_2^{-1}$, and substituting the bound (3-4) into (3-3), one derives

$$T^L_{F,G,N}(1,\dots,1) \ll_{c,C,\varepsilon} \sigma_G + \frac{\tau_2^{1/2}}{\sigma_G} + \frac{\tau_2^{-O(1)}A_L(\Omega_{c,C}(1),\tau_2)^{-1}}{N}$$

as required. □

The relations (3-4) formalise the estimate (2-2), which we first discussed when introducing the approximation function A_L in Definition 2.6. With the details all here, one can now see that it would have been enough to define the approximation function, at least if L is purely irrational, to be the function

$$\tau_2 \mapsto \min_{\substack{\boldsymbol{k}\in\mathbb{Z}^m\\ 0<\|\boldsymbol{k}\|_\infty\leqslant\tau_2^{-1}}} \operatorname{dist}(\boldsymbol{k}^T M^{-1}L, (\mathbb{Z}^d)^T). \tag{3-5}$$

One might now be concerned that, in defining A_L using real vectors φ rather than integer vectors $\boldsymbol{k}$, we might have constructed a much weaker object than (3-5), making (3-4) a wasteful step in our estimation. This is not the case, because if $\varphi \in (\mathbb{R}^m)^*$ and

$$\operatorname{dist}(L^*\varphi, (\mathbb{Z}^d)^T) \leqslant \delta$$

then $\operatorname{dist}(L^*(M^*)^{-1}M^*\varphi, (\mathbb{Z}^d)^T) \leqslant \delta$, and so in particular $\operatorname{dist}(M^*\varphi, (\mathbb{Z}^m)^T) \leqslant \delta$ (as M is a rank matrix). Letting $\boldsymbol{k}^T \in (\mathbb{Z}^m)^T$ be the nearest integer vector to $M^*\varphi$, we have that

$$\operatorname{dist}(\boldsymbol{k}^T M^{-1}L, (\mathbb{Z}^d)^T) \ll_{c,C} \delta.$$

So, up to some constants depending on c and C, there is essentially no difference between working with Definition 2.6 or with (3-5).

Restricting to integer vectors $\boldsymbol{k}$ may seem more natural from the point of view of diophantine approximation, but on the other hand the expression (3-5) depends on the choice of the particular rank matrix M, which is not canonical. It was more to our taste to present a definition of A_L which was intrinsic to L. Lemma 8.1 of our follow-up paper [Walker 2019] is also a setting in which having real vectors in the definition of A_L seems to be more natural.

It is also worth highlighting the exact moment in the proof of Lemma 3.4 in which it was vital that L was purely irrational. Considering expression (3-3), if L was not purely irrational and X was bigger than the rational complexity of L then the final error term is just $X^{O(1)}$, which is not $o(1)$ as $N \to \infty$.

4. Normal form

In this section we recall a technical notion from [Green and Tao 2010a] that those authors refer to as *normal form*. In Section 8 we will need to appeal to a quantitative refinement of this notion, which we also develop here.

Let $\Psi : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Putting the standard coordinates on $\mathbb{R}^n$ and $\mathbb{R}^m$, we may write $(\psi_1, \dots, \psi_m) := \Psi : \mathbb{R}^n \to \mathbb{R}^m$ as a system of homogeneous linear forms. The crux of the theory from [Green and Tao 2010a] is that, provided Ψ is of so-called "finite Cauchy–Schwarz complexity", Ψ may be reparametrised in such a way that it interacts particularly well with certain applications of the Cauchy–Schwarz inequality (see Proposition 8.3). Below we will give a brief overview of this terminology, before introducing our own quantitative versions; a much fuller discussion may be found in [Green and Tao 2010a, Section 1] and [Gowers and Wolf 2010].

In words, a reparametrisation into normal form is one in which each linear form is the only one that mentions all of its particular collection of variables. For example, the forms

$$\begin{aligned}\psi_1(t,u,v) &= u+v\\ \psi_2(t,u,v) &= v+t\\ \psi_3(t,u,v) &= u+t\\ \psi_4(t,u,v) &= u+v+t\end{aligned} \tag{4-1}$$

are in normal form with respect to ψ_4, since ψ_4 is the only form to utilise all three of the variables. However, this system is not in normal form with respect to ψ_3, say. However, the system

$$\begin{aligned}\psi_1(t,u,v,w) &= u+v+2w\\ \psi_2(t,u,v,w) &= v+t-w\\ \psi_3(t,u,v,w) &= u+t-w\\ \psi_4(t,u,v,w) &= u+v+t,\end{aligned} \tag{4-2}$$

which parametrises the same subspace of $\mathbb{R}^4$, *is* in normal form for all i.

We repeat the precise definition from [Green and Tao 2010a].

Definition 4.1. Let m, n be natural numbers, and let $(\psi_1, \dots, \psi_m) = \Psi : \mathbb{R}^n \to \mathbb{R}^m$ be a system of homogeneous linear forms. Let $i \in [m]$. We say that Ψ is in normal form with respect to ψ_i if there exists a nonnegative integer s and a collection $J_i \subseteq \{\boldsymbol{e_1}, \dots, \boldsymbol{e_n}\}$ of the standard basis vectors, satisfying $|J_i| = s+1$, such that

$$\prod_{\boldsymbol{e} \in J_i} \psi_{i'}(\boldsymbol{e})$$

is nonzero when $i' = i$ and vanishes otherwise. We say that Ψ is in normal form if it is in normal form with respect to ψ_i for every i.

Let us also recall what it means for a certain system of forms Ψ' to extend the system of forms Ψ.

Definition 4.2. For a system of homogeneous linear forms $(\psi_1, \dots, \psi_m) = \Psi : \mathbb{R}^n \to \mathbb{R}^m$, an extension $(\psi'_1, \dots, \psi'_m) = \Psi' : \mathbb{R}^{n'} \to \mathbb{R}^m$ is a system of homogeneous linear forms on $\mathbb{R}^{n'}$, for some n' with $n' \geqslant n$, such that

(1) $\Psi'(\mathbb{R}^{n'}) = \Psi(\mathbb{R}^n)$;

(2) if we identify $\mathbb{R}^n$ with the subset $\mathbb{R}^n \times \{0\}^{n'-n}$ in the obvious manner, then Ψ is the restriction of Ψ' to this subset.

The paper [Green and Tao 2010a] includes a result (Lemma 4.4) on the existence of extensions in normal form, but we will need a quantitative refinement of this analysis.

The reader will note from examples (4-1) and (4-2) that the property of "being in normal form" is a property of the parametrisation, and not of the underlying space that is being parametrised. It is natural to wonder whether there is some property of a space that can enable one to find a parametrisation in normal form, even if the original parametrisation is not. Fortunately there is such a notion, and it is the notion of finite Cauchy–Schwarz complexity introduced in [Green and Tao 2010a].[13] We introduce this notion in the following definitions, which we have phrased in such a way as to help us formulate a quantitative version.

Definition 4.3 (suitable partitions). Let m, n be natural numbers, with $m \geqslant 2$, and let $(\psi_1, \dots, \psi_m) = \Psi : \mathbb{R}^n \to \mathbb{R}^m$ be a system of homogeneous linear forms. Fix $i \in [m]$. Let $\mathcal{P}_i$ be a partition of $[m] \setminus \{i\}$, i.e.,

$$[m] \setminus \{i\} = \bigcup_{k=1}^{s+1} \mathcal{C}_k$$

for some s satisfying $0 \leqslant s \leqslant m-2$ and some disjoint sets $\mathcal{C}_k$. We say that $\mathcal{P}_i$ is *suitable* for Ψ if

$$\psi_i \notin \operatorname{span}_{\mathbb{R}}(\psi_j : j \in \mathcal{C}_k)$$

for any k.

Definition 4.4 (degeneracy varieties). Let m, n be natural numbers, with $m \geqslant 2$. Let $\mathcal{P}_i$ be a partition of $[m] \setminus \{i\}$. We define the *$\mathcal{P}_i$-degeneracy variety* $V_{\mathcal{P}_i}$ to be the set of all the systems of homogeneous linear forms $\Psi : \mathbb{R}^n \to \mathbb{R}^m$ for which $\mathcal{P}_i$ is not suitable for Ψ. Finally, the *degeneracy variety* $V_{\text{degen}}(n, m)$ is given by

$$V_{\text{degen}}(n, m) := \bigcup_{i=1}^{m} \bigcap_{\mathcal{P}_i} V_{\mathcal{P}_i},$$

where the inner intersection is over all possible partitions $\mathcal{P}_i$.

[13] In [Green and Tao 2010a] this is just called "complexity".

It is easy to observe that $\Psi \in V_{\text{degen}}(n, m)$ if and only if, for some $i \neq j$, ψ_i is a real multiple of ψ_j. This also yields the following:

Proposition 4.5 (relating $V^*_{\text{degen}}(m, d)$ and $V_{\text{degen}}(d-m, d)$). *Let m, d be natural numbers with $d \geqslant m+2$, and let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map. Let $\Psi : \mathbb{R}^{d-m} \to \mathbb{R}^d$ be any system of homogeneous linear forms whose image is* $\ker L$. *Then $L \in V^*_{\text{degen}}(m, d)$ if and only if $\Psi \in V_{\text{degen}}(d-m, d)$.*

Proof. We know that $L \in V^*_{\text{degen}}(m, d)$ if and only if there exist some nonzero vector $\boldsymbol{e}_i^* - \lambda \boldsymbol{e}_j^* \in L^*((\mathbb{R}^m)^*)$. But $L^*((\mathbb{R}^m)^*) = (\ker L)^0 = (\Psi(\mathbb{R}^{d-m}))^0$, so this occurs if and only if $\psi_i = \lambda \psi_j$ for some i and j. □

We will prove a more general version of this statement in Lemma 7.1.

Green and Tao [2010a, Definition 1.5] refer to those $\Psi \in V_{\text{degen}}(n, m)$ as having infinite Cauchy–Schwarz complexity, and develop their theory for $\Psi \notin V_{\text{degen}}(n, m)$. As we did for describing degeneracy properties of L, we need to quantify such a notion.

Definition 4.6 (c_1-Cauchy–Schwarz complexity). Let m, n be natural numbers, with $m \geqslant 3$, and let c_1 be a positive constant. Let $(\psi_1, \dots, \psi_m) = \Psi : \mathbb{R}^n \to \mathbb{R}^m$ be a system of homogeneous linear forms. For $i \in [m]$, we define a quantity s_i either by defining $s_i + 1$ to be the minimal number of parts in a partition $\mathcal{P}_i$ of $[m] \setminus \{i\}$ such that $\operatorname{dist}(\Psi, V_{\mathcal{P}_i}) \geqslant c_1$, or by $s_i = \infty$ if no such partition exists. Then we define $s := \max(1, \max_i s_i)$. We say that s is the c_1-Cauchy–Schwarz complexity of Ψ.

We remark, for readers familiar with [Green and Tao 2010a], that we preclude the "complexity 0" case. This is for a mundane technical reason, that occurs when absorbing the exponential phases in Section 8, when it will be convenient that $s + 1 \geqslant 2$. This is why we need the condition $m \geqslant 3$ in the above definition. We also take this opportunity to note that if s satisfies the above definition, and $s \neq \infty$, then $2 \leqslant s + 1 \leqslant m - 1$.

We note an easy consequence of these definitions.

Lemma 4.7. *Let m and n be natural numbers, with $m \geqslant 3$, and let c_1 be a positive constant. Let $(\psi_1, \dots, \psi_m) = \Psi : \mathbb{R}^n \to \mathbb{R}^m$ be a system of homogeneous linear forms. Furthermore, suppose that* $\operatorname{dist}(\Psi, V_{\text{degen}}(n, m)) \geqslant c_1$. *Then Ψ has finite c_1-Cauchy–Schwarz complexity.*

Proof. We have already observed that $\Psi \in V_{\text{degen}}(n, m)$ if and only if, for some $i \neq j$, ψ_i is a real multiple of ψ_j. From now until the end of the proof, fix $\mathcal{P}_i$ to be the partition of $[m] \setminus \{i\}$ in which every form ψ_k is in its own part. Our initial observation then implies that $\Psi \in V_{\text{degen}}(n, m)$ if and only if $\Psi \in V_{\mathcal{P}_i}$ for some i. So $\operatorname{dist}(\Psi, V_{\text{degen}}(n, m)) \geqslant c_1$ implies that $\operatorname{dist}(\Psi, V_{\mathcal{P}_i}) \geqslant c_1$ for all i. Therefore, by using these partitions $\mathcal{P}_i$ in Definition 4.6, we conclude that Ψ has finite c_1-Cauchy–Schwarz complexity. □

After having built up these definitions, we state the key proposition on the existence of normal form extensions to systems of real linear forms. We remind the reader that all implied constants may depend on the dimensions of the underlying spaces.

Proposition 4.8 (normal form algorithm). *Let m, n be natural numbers, with $m \geqslant 3$, and let c_1, C_1 be positive constants. Let $(\psi_1, \dots, \psi_m) = \Psi : \mathbb{R}^n \to \mathbb{R}^m$ be a system of homogeneous linear forms, and*

suppose that the coefficients of Ψ are bounded above in absolute value by C_1. Furthermore, suppose that Ψ has c_1-Cauchy–Schwarz complexity s, for some finite s. Then, for each $i \in [m]$, there is an extension $\Psi' : \mathbb{R}^{n'} \to \mathbb{R}^m$ such that:

(1) $n' = n + s + 1 \leqslant n + m - 1$.

(2) Ψ' *is of the form*

$$\Psi'(\boldsymbol{u}, w_1, \dots, w_{s+1}) := \Psi(\boldsymbol{u} + w_1 \boldsymbol{f_1} + \cdots + w_{s+1} \boldsymbol{f_{s+1}})$$

for some vectors $\boldsymbol{f_k} \in \mathbb{R}^n$, such that $\|\boldsymbol{f_k}\|_\infty = O_{c_1,C_1}(1)$ for every k.

(3) Ψ' *is in normal form with respect to ψ_i'.*

(4) $\psi_i'(\mathbf{0}, \boldsymbol{w}) = w_1 + \cdots + w_{s+1}$.

The proof is deferred to Appendix C, as it is very similar to the proof from [Green and Tao 2010a] (although with one important extra subtlety, which we mention in the appendix).

We conclude this discussion of normal form by noting an example of a system of homogeneous linear forms that may be reparametrised in normal form, but without quantitative control over the resulting extension.

Indeed, take $\iota(N)$ to be some function such that $\iota(N) \to \infty$ as $N \to \infty$. Consider the forms

$$\begin{aligned}\psi_1(u_1, u_2, u_3) &= (1 + \iota(N)^{-1})u_1 + u_2\\ \psi_2(u_1, u_2, u_3) &= u_1 + u_2\\ \psi_3(u_1, u_2, u_3) &= u_3.\end{aligned}$$

and let $\Psi := (\psi_1, \psi_2, \psi_3)$. Notice that $\operatorname{dist}(\Psi, V_{\text{degen}}(3, 3)) \to 0$ as $N \to \infty$. Therefore, for any $c_1 > 0$, if N is large enough then Ψ does not have finite c_1-Cauchy–Schwarz complexity. One may nonetheless construct a normal form reparametrisation

$$\begin{aligned}\psi_1'(u_1, u_2, u_3, w_1, w_2) &= (1 + \iota(N)^{-1})u_1 + u_2 + w_1\\ \psi_2'(u_1, u_2, u_3, w_1, w_2) &= u_1 + u_2 + w_2\\ \psi_3'(u_1, u_2, u_3, w_1, w_2) &= u_3.\end{aligned}$$

However, since

$$\Psi'(u_1, u_2, u_3, w_1, w_2) = \Psi(u_1 + \iota(N)w_1 - \iota(N)w_2, u_2 - \iota(N)w_1 + (\iota(N) + 1)w_2, u_3),$$

Ψ' is not obtained by bounded shifts of the u_i variables, and so (if N is large enough) it fails to satisfy part (2) of the conclusion of the above proposition. Such an extension Ψ' would not be suitable for our requirements in Section 8.

Remark 4.9. In [Green and Tao 2010a], the simple algorithm that constructs normal form extensions with respect to a fixed i may easily be iterated, and so the authors work with systems that are in normal form with respect to every index i. A careful analysis of the proof in Appendix C of [loc. cit.] demonstrates

that it is sufficient for Ψ merely to admit, for each i separately, an extension that is in normal form with respect to ψ_i, but this is of little consequence in [loc. cit.]. Yet certain quantitative aspects of the iteration of the normal form algorithm, critical to our application of these ideas, are not immediately clear to us. We have stated Proposition 4.8 for normal forms only with respect to a single i, in order to avoid this technical annoyance.

5. Dimension reduction

As we described in our proof strategy (Section 2E), in this section we reduce Theorem 2.12 to a different result, namely Theorem 5.6. This second theorem will be simpler in one key respect: the replacement of sharp cutoffs by Lipschitz cutoffs. It is the proof of Theorem 5.6 in which the Lipschitz property is actually used, and this will begin in Section 6. Any reader only wishing to consider the case of diophantine inequalities with Lipschitz cutoffs may eschew Section 5 of this paper entirely.

We begin by dismissing the case of maximal rational dimension.

Proposition 5.1. *Theorem 2.12 holds under the additional assumption that L has rational dimension m.*

To prove this, we will appeal to a quantitative version of Theorem 1.2.

Theorem 5.2 (generalised von Neumann theorem for rational forms (quantitative version)). *Let N, m, d be natural numbers, satisfying $d \geqslant m+2$, and let C_1, C_2 be positive constants. Let $S = S(N)$ be an m-by-d matrix with integer coefficients, satisfying $\|S\|_\infty \leqslant C_1$, and let $\boldsymbol{r} \in \mathbb{Z}^m$ be some vector with $\|\boldsymbol{r}\|_\infty \leqslant C_2 N$. Suppose S has rank m, and $S \notin V^*_{\mathrm{degen}}(m,d)$. Let $K \subseteq [-N,N]^d$ be convex. Then there exists some natural number s at most $d-2$ that satisfies the following. Let $f_1, \dots, f_d : [N] \to \mathbb{C}$ be arbitrary functions with $\|f_j\|_\infty \leqslant 1$ for all j, and assume that*

$$\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho$$

for some ρ in the range $0 < \rho \leqslant 1$. Then

$$\frac{1}{N^{d-m}} \sum_{\substack{\boldsymbol{n} \in \mathbb{Z}^d \cap K \\ S\boldsymbol{n} = \boldsymbol{r}}} \prod_{j=1}^{d} f_j(n_j) \ll_{C_1,C_2} \rho^{\Omega(1)} + o_\rho(1)$$

as $N \to \infty$. Furthermore, the $o_\rho(1)$ term may be bounded above by $\rho^{-O(1)} N^{-\Omega(1)}$.

Let us sketch a proof of this result, assuming a certain familiarity with the methods and terminology of [Green and Tao 2010a].

Proof sketch of Theorem 5.2. One follows the proof of Theorem 1.8 of [Green and Tao 2010a]. Firstly, recall that in our language, the nondegeneracy condition in the statement of Theorem 1.8 of [loc. cit.] is exactly the condition that $S \notin V^*_{\mathrm{degen}}(m,d)$. One then follows the same linear algebraic reductions as those used in Section 4 of [loc. cit.] to reduce Theorem 1.8 to Theorem 7.1 of the same paper (the generalised von Neumann theorem).

Theorem 7.1 may then be considered solely in the case of bounded functions f_j, as in [Tao 2012, Exercise 1.3.23], rather than in the more general case of functions bounded by a pseudorandom measure. It is clear from the proof that, in this more restricted setting, the $\kappa(\rho)$ term that appears in the statement may be replaced by a polynomial dependence, and the $o_\rho(1)$ term may be bounded above by $\rho^{-O(1)}N^{-\Omega(1)}$.

This settles Theorem 5.2, where s is the Cauchy–Schwarz complexity of some system of forms $(\psi_1, \dots, \psi_d)$ that parametrises $\ker S$. But s is at most $d-2$, as any system of d forms with finite Cauchy–Schwarz complexity has Cauchy–Schwarz complexity at most $d-2$. Therefore Theorem 5.2 is proved. □

Now let us use Theorem 5.2 to resolve Proposition 5.1.

Proof of Proposition 5.1. Let L be as in Theorem 2.12, and assume that L has rational dimension m and rational complexity at most C. Let $\Theta : \mathbb{R}^m \to \mathbb{R}^m$ be some linear isomorphism satisfying $\Theta L(\mathbb{Z}^d) \subseteq \mathbb{Z}^m$ and $\|\Theta\|_\infty \leqslant C$. Let M be a rank matrix of L (Proposition 3.1). Then the matrix $M^{-1}L$ satisfies $\|M^{-1}L\|_\infty \ll_{c,C} 1$ and has rational dimension m, since $((\Theta M) \circ (M^{-1}L))(\mathbb{Z}^d) = \Theta L(\mathbb{Z}^d) \subseteq \mathbb{Z}^m$. The matrix $M^{-1}L$ also has rational complexity $O_{c,C}(1)$. Therefore, replacing L with $M^{-1}L$, we may assume that the first m columns of L form the identity matrix.

As in Lemma 2.5, we write $\Theta L = S$, where S has integer coefficients and $\|\Theta\|_\infty \ll_{c,C} 1$. Hence $\|S\|_\infty \ll_{c,C} 1$. But Θ must also have integer coefficients, as the first m columns of L form the identity matrix, and hence $\|\Theta^{-1}\|_\infty \ll_{c,C} 1$ as well. Note finally that $S \notin V^*_{\mathrm{degen}}(m, d)$, since $L \notin V^*_{\mathrm{degen}}(m, d)$.

Now, suppose that $G : \mathbb{R}^m \to [0, 1]$ is the indicator function of some convex domain D, with $D \subseteq [-\varepsilon, \varepsilon]^m$. Then there are at most $O_{c,C,\varepsilon}(1)$ possible vectors $\boldsymbol{r} \in \mathbb{Z}^m$ such that $\boldsymbol{r} \in S(\mathbb{Z}^d) \cap \Theta(D)$. Let R be the set of all such vectors. Therefore, with F being the indicator function of the set $[1, N]^d$, we have

$$T^L_{F,G,N}(f_1, \dots, f_d) = \sum_{\boldsymbol{r}\in R} \sum_{\substack{\boldsymbol{n}\in[N]^d \\ S\boldsymbol{n}=\boldsymbol{r}}} \prod_{j=1}^{d} f_j(n_j) \ll_{c,C,\varepsilon} \rho^{\Omega(1)} + o_\rho(1) \tag{5-1}$$

as $N \to \infty$, by Theorem 5.2. The $o_\rho(1)$ term may be bounded above by $\rho^{-O(1)}N^{-\Omega(1)}$. This is the desired conclusion of Theorem 2.12 in the case when L has rational dimension m. □

Having dismissed this case, we prepare to state Theorem 5.6. We begin with a definition that generalises Definition 2.1.

Definition 5.3. Let N, m, d, h be natural numbers, with $d \geqslant h \geqslant m+2$. Let ε be positive. Let $\Xi = (\xi_1, \dots, \xi_d) : \mathbb{R}^h \to \mathbb{R}^d$ and $L : \mathbb{R}^h \to \mathbb{R}^m$ be linear maps. Let $F : \mathbb{R}^h \to [0, 1]$ and $G : \mathbb{R}^m \to [0, 1]$ be two functions, with F supported on $[-N, N]^h$ and G compactly supported. Let $\tilde{\boldsymbol{r}} \in \mathbb{Z}^d$ be some vector, and let $f_1, \dots, f_d : \mathbb{R} \to [-1, 1]$ be arbitrary functions. We then define

$$T^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1, \dots, f_d) := \frac{1}{N^{h-m}} \sum_{\boldsymbol{n}\in\mathbb{Z}^h} \left(\prod_{j=1}^{d} f_j(\xi_j(\boldsymbol{n}) + \tilde{\boldsymbol{r}}_j) \right) F(\boldsymbol{n})G(L\boldsymbol{n}). \tag{5-2}$$

In the paper so far we have introduced many degeneracy relations (Definitions 2.9, 2.10, 4.4). In order to state Theorem 5.6, we must introduce another.

Definition 5.4 (dual pair degeneracy variety). Let m, d, h be natural numbers satisfying $d \geqslant h \geqslant m+2$. Let $\boldsymbol{e_1}, \dots, \boldsymbol{e_d}$ denote the standard basis vectors of $\mathbb{R}^d$, and let $\boldsymbol{e_1^*}, \dots, \boldsymbol{e_d^*}$ denote the dual basis of $(\mathbb{R}^d)^*$. Then let $V^*_{\text{degen},2}(m, d, h)$ denote the set of all pairs of linear maps $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ and $L : \mathbb{R}^h \to \mathbb{R}^m$ for which there exist two indices $i, j \leqslant d$, and some real number λ, such that $(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*})$ is nonzero and $\Xi^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}) \in L^*((\mathbb{R}^m)^*)$. We call $V^*_{\text{degen},2}(m, d, h)$ the *dual pair degeneracy variety*.

One can motivate this definition as follows. We noted in Proposition 4.5 that, if $L : \mathbb{R}^d \to \mathbb{R}^m$ is a surjective linear map, then saying that $L \notin V^*_{\text{degen}}(m, d)$ is equivalent to saying that any parametrisation $\Psi = (\psi_1, \dots, \psi_d) : \mathbb{R}^{d-m} \to \mathbb{R}^d$ of $\ker L$ has finite Cauchy–Schwarz complexity. In this paper, following our sketched idea in expression (2-11), we will end up needing to replace the map L with two maps, an injective map $\Xi : \mathbb{R}^{d-u} \to \mathbb{R}^d$ and a purely irrational surjective map $L' : \mathbb{R}^{d-u} \to \mathbb{R}^{m-u}$ (here u will be the rational dimension of L). It will turn out that after this manipulation the system of forms that we will require to have finite Cauchy–Schwarz complexity (in order to bring in Gowers norms) will be $\Xi\Psi' : \mathbb{R}^{d-m} \to \mathbb{R}^d$, where $\Psi' : \mathbb{R}^{d-m} \to \mathbb{R}^{d-u}$ is a parametrisation of $\ker L'$. One can easily show (and we do, in Lemma 7.1), that $(\Xi, L') \notin V^*_{\text{degen}}(m-u, d, d-u)$ is the exactly the right condition to ensure that $\Xi\Psi' : \mathbb{R}^{d-m} \to \mathbb{R}^d$ has finite Cauchy–Schwarz complexity.

As ever, we need a quantitative version of nondegeneracy.

Definition 5.5 (distance metric for pairs of matrices). Let m, d, h be natural numbers, with $d \geqslant h \geqslant m+2$, and let $V^*_{\text{degen},2}(m, d, h)$ be the dual pair degeneracy variety. Let $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ and $L : \mathbb{R}^h \to \mathbb{R}^m$ be linear maps. We say that $\operatorname{dist}((\Xi, L), V^*_{\text{degen},2}(m, d, h)) \geqslant c$ if $(\Xi + Q, L) \notin V^*_{\text{degen},2}(m, d, h)$ for all $Q : \mathbb{R}^h \to \mathbb{R}^d$ with $\|Q\|_\infty < c$.

Although this is no great subtlety, we should emphasise that in the above definition we only consider perturbations to Ξ, and not perturbations to L as well.

We are now ready to state our theorem on linear inequalities with Lipschitz cutoffs.

Theorem 5.6 (Lipschitz case). *Let N, m, d, h be natural numbers, with $d \geqslant h \geqslant m+2$, and let c, C, ε be positive constants. Let $\Xi = \Xi(N) : \mathbb{R}^h \to \mathbb{R}^d$ be an injective linear map with integer coefficients, and assume that $\Xi(\mathbb{Z}^h) = \mathbb{Z}^d \cap \operatorname{im} \Xi$. Let $L = L(N) : \mathbb{R}^h \to \mathbb{R}^m$ be a surjective linear map. Assume that $\|\Xi\|_\infty \leqslant C$, $\|L\|_\infty \leqslant C$, $\operatorname{dist}(L, V_{\text{rank}}(m, d)) \geqslant c$ and $\operatorname{dist}((\Xi, L), V^*_{\text{degen},2}(m, d, h)) \geqslant c$. Then there exists a natural number s at most $d-2$, independent of ε, such that the following holds. Let σ_F, σ_G be any two parameters in the range $0 < \sigma_F, \sigma_G < \frac{1}{2}$. Let $F : \mathbb{R}^h \to [0, 1]$ be a Lipschitz function supported on $[-N, N]^h$ with Lipschitz constant $O(1/\sigma_F N)$, and let $G : \mathbb{R}^m \to [0, 1]$ be a Lipschitz function supported on $[-\varepsilon, \varepsilon]^m$ with Lipschitz constant $O(1/\sigma_G)$. Let $\tilde{\boldsymbol{r}}$ be a fixed vector in $\mathbb{Z}^d$, satisfying $\|\tilde{\boldsymbol{r}}\|_\infty = O_{c,C,\varepsilon}(1)$. Suppose that $f_1, \dots, f_d : [N] \to [-1, 1]$ are arbitrary bounded functions satisfying*

$$\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho,$$

for some ρ in the range $0 < \rho \leqslant 1$. Then

$$T^{L,\Xi,\tilde{r}}_{F,G,N}(f_1,\dots,f_d) \ll_{c,C,\varepsilon} \rho^{\Omega(1)}(\sigma_F^{-O(1)} + \sigma_G^{-O(1)}) + \sigma_F^{-O(1)} N^{-\Omega(1)}. \tag{5-3}$$

Although the above theorem contains more technical conditions than even Theorem 2.12 did, it does represent a significant reduction in complexity from the original problem. Note in particular that the approximation function A_L does not feature in the estimate (5-3).

As we described in Section 2E, the presence of Lipschitz cutoffs rather than convex cutoffs will be especially convenient when approximating the discrete solution count by a continuous solution count. This will be done in Section 6.

The remainder of this section will be devoted to proving the main theorem (Theorem 2.12), assuming the truth of Theorem 5.6.

We begin with two lemmas: one concerning lattices, and the other concerning a quantitative decomposition of the dual space $(\mathbb{R}^d)^*$. Their proofs are entirely standard, but we state them prominently, as we will need to refer to them often in the dimension reduction argument of Lemma 5.10.

Lemma 5.7 (parametrising the image lattice). *Let u, d be integers with $d \geqslant u+1$. Let $S : \mathbb{R}^d \to \mathbb{R}^u$ be a surjective linear map with $S(\mathbb{Z}^d) \subseteq \mathbb{Z}^u$, and suppose that $\|S\|_\infty \leqslant C$. Then there exists a set $\{\boldsymbol{a_1},\dots,\boldsymbol{a_u}\} \subset \mathbb{Z}^u$ that is a basis for the lattice $S(\mathbb{Z}^d)$ and for which $\|\boldsymbol{a_i}\|_\infty = O_C(1)$ for every i. Furthermore there exist $\boldsymbol{x_1},\dots,\boldsymbol{x_u} \in \mathbb{Z}^d$ such that, for every i, $S(\boldsymbol{x_i}) = \boldsymbol{a_i}$ and $\|\boldsymbol{x_i}\|_\infty = O_C(1)$.*

Proof. The lattice $S(\mathbb{Z}^d)$ is u dimensional, as S is surjective. If $\{\boldsymbol{e_j} : j \leqslant d\}$ denotes the standard basis of $\mathbb{R}^d$ then integer combinations of elements from the set $\{S(\boldsymbol{e_j}) : j \leqslant d\}$ span $S(\mathbb{Z}^d)$. Since $\|S\|_\infty \leqslant C$, these vectors also satisfy $\|S(\boldsymbol{e_j})\|_\infty = O_C(1)$. Therefore the u successive minima of the lattice $S(\mathbb{Z}^d)$ are all $O_C(1)$, and so, by Mahler's theorem [Tao and Vu 2006, Theorem 3.34] the lattice $S(\mathbb{Z}^d)$ has a basis $\{\boldsymbol{a_1},\dots,\boldsymbol{a_u}\}$ of the required form.

Note that S has integer coefficients. The construction of suitable $\boldsymbol{x_1},\dots,\boldsymbol{x_u}$ may be achieved by applying any of the standard algorithms. For example, using Gaussian elimination one may find a basis for $\ker S$ that, by inspection of the algorithm, consists of vectors with rational coordinates of naive height $O_C(1)$. By clearing denominators, one gets vectors $\boldsymbol{v_1},\dots,\boldsymbol{v_{d-u}} \in \mathbb{Z}^d$ whose integer span is a full-dimensional sublattice of the $d-u$ dimensional lattice $\mathbb{Z}^d \cap \ker S$, and that satisfy $\|\boldsymbol{v_i}\|_\infty = O_C(1)$ for all i. Now given some $\boldsymbol{a_i}$, by its construction there must be some $\boldsymbol{x_i} \in \mathbb{Z}^d$ that satisfies $S(\boldsymbol{x_i}) = \boldsymbol{a_i}$. Write $\boldsymbol{x_i} = \boldsymbol{x_i}|_{\ker S} + \boldsymbol{x_i}|_{(\ker S)^\perp}$ as the sum of the obvious projections. By adding a suitable integer combination of the vectors $\boldsymbol{v_1},\dots,\boldsymbol{v_{d-u}}$ to $\boldsymbol{x_i}$ one may find such an $\boldsymbol{x_i}$ that satisfies $\|\boldsymbol{x_i}|_{\ker S}\|_\infty = O_C(1)$. Furthermore, $\operatorname{dist}(S, V_{\mathrm{rank}}(m,d)) = \Omega_C(1)$, since S has integer coordinates, and so (by Lemma D.1) $\|\boldsymbol{x_i}|_{(\ker S)^\perp}\|_\infty = O_C(1)$. Hence $\|\boldsymbol{x_i}\|_\infty = O_C(1)$, as desired. □

Having established that such a lattice basis $\{\boldsymbol{a_1},\dots,\boldsymbol{a_u}\}$ exists, we can now use it to quantitatively decompose $(\mathbb{R}^d)^*$.

Lemma 5.8 (dual space decomposition). *Let u, d, be integers with $d \geqslant u+1$, and let C, η be constants. Let $S : \mathbb{R}^d \to \mathbb{R}^u$ be a surjective linear map with $S(\mathbb{Z}^d) \subseteq \mathbb{Z}^u$, and suppose that $\|S\|_\infty \leqslant C$. Let $\{\boldsymbol{a_1}, \dots, \boldsymbol{a_u}\}$ be a basis for the lattice $S(\mathbb{Z}^d)$ that satisfies $\|\boldsymbol{a_i}\|_\infty = O_C(1)$ for every i. Let $\boldsymbol{x_1}, \dots, \boldsymbol{x_u} \in \mathbb{Z}^d$ be vectors such that, for every i, $S(\boldsymbol{x_i}) = \boldsymbol{a_i}$ and $\|\boldsymbol{x_i}\|_\infty = O_C(1)$. Suppose that $\Xi : \mathbb{R}^{d-u} \to \mathbb{R}^d$ is an injective linear map such that* $\operatorname{im} \Xi = \ker S$ *and such that $\Xi(\mathbb{Z}^{d-u}) = \mathbb{Z}^d \cap \operatorname{im} \Xi$. Suppose further that $\|\Xi\|_\infty \leqslant C$.*

Let $\boldsymbol{w_1}, \dots, \boldsymbol{w_{d-u}}$ denote the standard basis vectors in $\mathbb{R}^{d-u}$. Then:

(1) *The set $\mathcal{B} := \{\boldsymbol{x_i} : i \leqslant u\} \cup \{\Xi(\boldsymbol{w_j}) : j \leqslant d-u\}$ is a basis for $\mathbb{R}^d$, and a lattice basis for $\mathbb{Z}^d$.*

(2) *Writing $\mathcal{B}^* := \{\boldsymbol{x_i}^* : i \leqslant u\} \cup \{\Xi(\boldsymbol{w_j})^* : j \leqslant d-u\}$ for the dual basis, both the change of basis matrix between the standard dual basis and $\mathcal{B}^*$ and the inverse of this matrix have integer coordinates. The coefficients of both of these matrices are bounded in absolute value by $O_C(1)$.*

Write $V := \operatorname{span}(\boldsymbol{x_i}^* : i \leqslant u)$ and $W := \operatorname{span}(\Xi(\boldsymbol{w_j})^* : j \leqslant d-u)$. Then:

(3) $V = S^*((\mathbb{R}^u)^*)$.

(4) *Suppose that $\varphi \in (\mathbb{R}^d)^*$ satisfies $\|\Xi^*(\varphi)\|_\infty \leqslant \eta$. Then, writing $\varphi = \varphi_V + \varphi_W$ with $\varphi_V \in V$ and $\varphi_W \in W$, we have $\|\varphi_W\|_\infty = O_C(\eta)$.*

Proof. For part (1), the fact that $\mathcal{B}$ is a basis for $\mathbb{R}^d$ is just a manifestation of the familiar principle $\mathbb{R}^d \cong \ker S \oplus \operatorname{im} S$. To show that $\mathcal{B}$ is a lattice basis for $\mathbb{Z}^d$, let $\boldsymbol{n} \in \mathbb{Z}^d$ and write

$$\boldsymbol{n} = \sum_{i=1}^{u} \lambda_i \boldsymbol{x_i} + \sum_{j=1}^{d-u} \mu_j \Xi(\boldsymbol{w_j})$$

for some $\lambda_i, \mu_j \in \mathbb{R}$. Applying S, we see $S(\boldsymbol{n}) = \sum_i^u \lambda_i \boldsymbol{a_i}$, and hence $\lambda_i \in \mathbb{Z}$ for all i, as $\{\boldsymbol{a_1}, \dots, \boldsymbol{a_u}\}$ is a basis for the lattice $S(\mathbb{Z}^d)$. But this implies $\sum_{j=1}^{d-u} \mu_j \Xi(\boldsymbol{w_j}) \in \mathbb{Z}^d \cap \operatorname{im}(\Xi)$. Therefore, as $\Xi(\mathbb{Z}^{d-u}) = \mathbb{Z}^d \cap \ker S$, $\mu_j \in \mathbb{Z}$ for all j.

Part (2) follows immediately from part (1). Part (3) is immediate from the definitions.

For part (4), let j be at most $d-u$. Then the assumption $\|\Xi^*(\varphi)\|_\infty \leqslant \eta$ means that $|\Xi^*(\varphi)(\boldsymbol{w_j})| \leqslant \eta$. Hence $|\varphi(\Xi(\boldsymbol{w_j}))| \leqslant \eta$. But, writing $\varphi_W = \sum_{j=1}^{d-u} \mu_j \Xi(\boldsymbol{w_j})^*$, this implies that $|\mu_j| \leqslant \eta$. Since the coefficients of the change of basis matrix between $\mathcal{B}^*$ and the standard dual basis are bounded in absolute value by $O_C(1)$, this implies that $\|\varphi_W\|_\infty \leqslant O_C(\eta)$. □

We now begin the attack on Theorem 2.12 in earnest. Assume the hypotheses of Theorem 2.12. As a reminder, we have natural numbers m, d satisfying $d \geqslant m+2$, and positive reals ε, c, C. For a natural number N, we have $L = L(N) : \mathbb{R}^d \to \mathbb{R}^m$ being a surjective linear map with approximation function A_L, with $\operatorname{dist}(L, V^{\text{unif}}_{\text{rank}}(m,d)) \geqslant c$, $\operatorname{dist}(L, V^*_{\text{degen}}(m,d)) \geqslant c$, and with rational complexity at most C. We have $F : \mathbb{R}^d \to [0,1]$ being the indicator function of $[1,N]^d$ and $G : \mathbb{R}^m \to [0,1]$ being the indicator function of a convex domain contained in $[-\varepsilon, \varepsilon]^m$. For some $s \leqslant d-2$, to be determined, we also have functions $f_1, \dots, f_d : [N] \to [-1,1]$ that satisfy $\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho$ for some ρ in the range $0 < \rho \leqslant 1$.

The proof has four parts:

- Lemma 5.9, in which we replace the indicator function of $[1, N]^d$ with a Lipschitz cutoff.
- Lemma 5.10, in which we replace L by a pair of maps (Ξ, L') where L' is purely irrational.
- Lemma 5.11, in which we replace the function G by a Lipschitz cutoff (using Lemma 3.4).
- Finally, the application of Theorem 5.6 to the pair (Ξ, L').

The second of these steps is by far the most technically intricate, and, as we mentioned when discussing our proof strategy in Section 2E, Lemma 5.10 will have 9 subparts. One might well ask why it is necessary to expend so much effort creating a purely irrational map L', given that Theorem 5.6 does not include this condition in its hypotheses. The point is that in order to replace G with a Lipschitz cutoff (and thus in order to be able to apply Theorem 5.6 at all) it is vital that L' is purely irrational. If $L' : \mathbb{R}^{d-u} \to \mathbb{R}^{m-u}$ failed to be purely irrational then $L'\mathbb{Z}^{d-u}$ would not equidistribute in $\mathbb{R}^{m-u}$; it would instead be restricted to certain proper affine subspaces. This would affect our ability to perturb the function G without drastically altering the number of solutions to the inequality. For more on this issue, the reader may consult Section 2E.

One does note from the above discussion, however, that in order to deduce Theorem 2.12 it would be enough to prove Theorem 5.6 under the additional assumption that L is purely irrational. Yet it turns out that the general version of Theorem 5.6 that we have stated is no harder to prove than the restricted version.

We begin with the first of our four parts.

Lemma 5.9 (replacing variable cutoff). *Assume the hypotheses of Theorem 2.12 (in particular let F be the indicator function $1_{[1,N]^d}$), and let σ_F be any parameter in the range $0 < \sigma_F < \frac{1}{2}$. Then there exists a Lipschitz function $F_{1,\sigma_F} : \mathbb{R}^d \to [0, 1]$, supported on $[-2N, 2N]^d$ and with Lipschitz constant $O(1/\sigma_F N)$, such that*

$$|T^L_{F,G,N}(f_1, \dots, f_d)| \ll |T^L_{F_{1,\sigma_F},G,N}(f_1, \dots, f_d)| + O_{c,C}(\sigma_F).$$

Proof. By Lemma B.2, for any parameter σ_F in the range $0 < \sigma_F < \frac{1}{2}$ we may write

$$1_{[1,N]^d} = F_{1,\sigma_F} + O(F_{2,\sigma_F}),$$

where F_{1,σ_F}, F_{2,σ_F} are Lipschitz functions supported on $[-2N, 2N]^d$, with Lipschitz constants $O(1/\sigma_F N)$, and with $\int_{\boldsymbol{x}} F_{2,\sigma_F}(\boldsymbol{x})\, d\boldsymbol{x} = O(\sigma_F N^d)$. Moreover, F_{2,σ_F} is supported on

$$\{\boldsymbol{x} \in \mathbb{R}^d : \operatorname{dist}(\boldsymbol{x}, \partial([1, N]^d)) = O(\sigma_F N)\}.$$

Therefore

$$T^L_{F,G,N}(f_1, \dots, f_d) \ll |T^L_{F_{1,\sigma_F},G,N}(f_1, \dots, f_d)| + |T^L_{F_{2,\sigma_F},G,N}(1, \dots, 1)|.$$

Therefore, since $\operatorname{dist}(L, V^{\text{unif}}_{\text{rank}}(m, d)) \geqslant c$, by Lemma 3.3 we have

$$|T^L_{F_{2,\sigma_F},G,N}(f_1, \dots, f_d)| = O_{c,C}(\sigma_F).$$

This gives the lemma. □

Next comes the critical lemma, in which we successfully replace the map L by a purely irrational map L'. For the definition of the approximation function A_L, one may consult Definition 2.6.

Lemma 5.10 (generating a purely irrational map). *Let σ_F be a parameter in the range $0 < \sigma_F < \frac{1}{2}$. Assume the hypotheses of Theorem 2.12, with the exception that $F : \mathbb{R}^d \to [0,1]$ now denotes a Lipschitz function supported on $[-2N, 2N]^d$ and with Lipschitz constant $O(1/\sigma_F N)$. Let u be the rational dimension of L, and assume that $u \leqslant m-1$. Then there exists a surjective linear map $L' : \mathbb{R}^{d-u} \to \mathbb{R}^{m-u}$, an injective linear map $\Xi : \mathbb{R}^{d-u} \to \mathbb{R}^d$, a finite subset $\tilde{R} \subset \mathbb{Z}^d$, and, for each $\tilde{\boldsymbol{r}} \in \tilde{R}$, functions $F_{\tilde{\boldsymbol{r}}} : \mathbb{R}^{d-u} \to [0,1]$ and $G_{\tilde{\boldsymbol{r}}} : \mathbb{R}^{m-u} \to [0,1]$, that together satisfy the following properties*:

(1) *Ξ has integer coefficients, $\|\Xi\|_\infty = O_{c,C}(1)$, and $\Xi(\mathbb{Z}^{d-u}) = \mathbb{Z}^d \cap \operatorname{im} \Xi$.*

(2) *$|\tilde{R}| = O_{c,C}(1)$, and $\|\tilde{\boldsymbol{r}}\|_\infty = O_{c,C}(1)$ for all $\tilde{\boldsymbol{r}} \in \tilde{R}$.*

(3) *$F_{\tilde{\boldsymbol{r}}}$ is supported on $[-O_{c,C}(N), O_{c,C}(N)]^{d-u}$, with Lipschitz constant $O_{c,C}(1/\sigma_F N)$, and $G_{\tilde{\boldsymbol{r}}}$ is the indicator function of a convex domain contained in $[-O_{c,C,\varepsilon}(1), O_{c,C,\varepsilon}(1)]^{m-u}$.*

(4) *$T^L_{F,G,N}(f_1, \dots, f_d) = \sum_{\tilde{\boldsymbol{r}} \in \tilde{R}} T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}}},N}(f_1, \dots, f_d)$.*

(5) *L' is purely irrational.*

(6) *$\|L'\|_\infty = O_{c,C}(1)$ and $\operatorname{dist}(L', V_{\mathrm{rank}}(m-u, d-u)) = \Omega_{c,C}(1)$.*

(7) *$\operatorname{dist}((\Xi, L'), V^*_{\mathrm{degen},2}(m-u, d, d-u)) = \Omega_{c,C}(1)$.*

(8) *For all $\tau_1, \tau_2 \in (0,1]$, $A_{L'}(\tau_1, \tau_2) \gg_{c,C} A_L(\Omega_{c,C}(\tau_1), \Omega_{c,C}(\tau_2))$.*

(9) *For all $\tau_1, \tau_2 \in (0,1]$, $A_{L'}(\tau_1, \tau_2) \ll_{c,C} A_L(\Omega_{c,C}(\tau_1), \Omega_{c,C}(\tau_2))$.*

The fundamental aspect of this lemma is part (4), of course, as this directly concerns how we control the number of solutions to the diophantine inequality itself when passing from L to L'. However, we do need to establish parts (1)–(8), in order to be able to ensure that the hypotheses of Lemma 3.4 and Theorem 5.6 are satisfied. Part (9) is included for completeness, and to assist the calculations in Appendix E.

Before giving the full details of the proof, we sketch the idea. Let $\Theta : \mathbb{R}^m \to \mathbb{R}^u$ be a rational map for L. The space $\ker(\Theta L)$ has dimension $d-u$, and so we may parametrise it by some injective map $\Xi : \mathbb{R}^{d-u} \to \ker(\Theta L)$. Without too much difficultly, Ξ can be chosen to satisfy $\Xi(\mathbb{Z}^{d-u}) = \mathbb{Z}^d \cap \operatorname{im} \Xi$. Then

$$L\Xi : \mathbb{R}^{d-u} \to \ker \Theta,$$

is a map from a $d-u$ dimensional space to an $m-u$ dimensional space, and it turns out that $L\Xi$ is purely irrational, and $L' = L\Xi$ may be used in Lemma 5.10.

Of course this is not quite possible, as we only defined the notion of purely irrational maps between vector spaces of the form $\mathbb{R}^a$. But it is true after choosing a judicious isomorphism from $\ker \Theta$ to $\mathbb{R}^{m-u}$ (though this does complicate the notation).

Let us complete the details.

Proof. First we note that the lemma is obvious when $u = 0$, since one may take $\Xi : \mathbb{R}^d \to \mathbb{R}^d$ to be the identity map, $\tilde{\boldsymbol{r}}$ to be $\mathbf{0}$, and L' to be L. So assume that $u > 1$.

We proceed with a general reduction, familiar from our proof of Proposition 5.1, in which we may assume that the first m columns of L form the identity matrix.

Indeed, let $\Theta : \mathbb{R}^m \to \mathbb{R}^u$ be a rational map for L with $\|\Theta\|_\infty \leqslant C$. Now let $\tilde{L} := M^{-1}L$, where M is a rank matrix of L (Proposition 3.1), which, without loss of generality, consists of the first m columns of L. Let $\tilde{\Theta} := \Theta M$ and let $\tilde{G} := G \circ M$. Then

$$T^L_{F,G,N}(f_1, \dots, f_d) = T^{\tilde{L}}_{F,\tilde{G},N}(f_1, \dots, f_d),$$

and, considering $\tilde{\Theta}$, $\tilde{L}$ has rational complexity $O_{c,C}(1)$. Furthermore, $\tilde{G}$ is the indicator function of a convex domain contained in $[-O_{c,C}(\varepsilon), O_{c,C}(\varepsilon)]^m$. We also have $\operatorname{dist}(\tilde{L}, V^*_{\text{degen}}(m, d)) = \Omega_{c,C}(1)$. Finally, for all $\tau_1, \tau_2 \in (0, 1]$, we have that $A_{\tilde{L}}(\tau_1, \tau_2) \asymp_{c,C} A_L(\Omega_{c,C}(\tau_1), \Omega_{c,C}(\tau_2))$.

Therefore, by replacing L with $\tilde{L}$ and G with $\tilde{G}$, we may assume throughout the proof of Lemma 5.10 that the first m columns of L form the identity matrix. This is at the cost of replacing ε by $O_{c,C}(\varepsilon)$, C by $O_{c,C}(1)$, and c by $\Omega_{c,C}(1)$.

Now let $\Theta : \mathbb{R}^m \to \mathbb{R}^u$ be a rational map for L with $\|\Theta\|_\infty = O_{c,C}(1)$. Since the first m columns of L form the identity matrix, Θ must have integer coefficients.

Part (1): By rank-nullity $\ker(\Theta L)$ is a $d - u$ dimensional subspace of $\mathbb{R}^d$. The matrix of ΘL has integer coefficients and $\|\Theta L\|_\infty = O_{c,C}(1)$. Combining these two facts, we see that $\ker(\Theta L) \cap \mathbb{Z}^d$ is a $d - u$ dimensional lattice, and by the standard algorithms one can find a lattice basis $\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(d-u)} \in \mathbb{Z}^d$ that satisfies $\|\boldsymbol{v}^{(i)}\|_\infty = O_{c,C}(1)$ for every i. Define $\Xi : \mathbb{R}^{d-u} \to \mathbb{R}^d$ by

$$\Xi(\boldsymbol{w}) := \sum_{i=1}^{d-u} w_i \boldsymbol{v}^{(i)}.$$

Then Ξ satisfies property (1) of the lemma. Note that the image of the map $L\Xi : \mathbb{R}^{d-u} \to \mathbb{R}^m$ is exactly $\ker \Theta$.

Part (2): Since $\|\Theta\|_\infty = O_{c,C}(1)$, if $\boldsymbol{y} \in \mathbb{R}^m$ and $\Theta(\boldsymbol{y}) = \boldsymbol{r}$ then $\|\boldsymbol{y}\|_\infty \gg_{c,C} \|\boldsymbol{r}\|_\infty$. Recall that the support of G is contained within $[-O_{c,C,\varepsilon}(1), O_{c,C,\varepsilon}(1)]^m$, and that $\Theta L(\mathbb{Z}^d) \subseteq \mathbb{Z}^u$. It follows that there are at most $O_{c,C,\varepsilon}(1)$ possible vectors $\boldsymbol{r} \in \mathbb{Z}^u$ for which there exists a vector $\boldsymbol{n} \in \mathbb{Z}^d$ for which both $G(L\boldsymbol{n}) \neq 0$ and $\Theta L\boldsymbol{n} = \boldsymbol{r}$. Let R denote the set of all such vectors $\boldsymbol{r}$.

For each $\boldsymbol{r} \in R$, there exists a vector $\tilde{\boldsymbol{r}} \in \mathbb{Z}^d$ such that $\Theta L\tilde{\boldsymbol{r}} = \boldsymbol{r}$ and $\|\tilde{\boldsymbol{r}}\|_\infty = O_{c,C,\varepsilon}(1)$. Let $\tilde{R}$ denote the set of these $\tilde{\boldsymbol{r}}$. Then $\tilde{R}$ satisfies part (2).

Before proceeding to prove part (3) of the lemma, we pause to apply Lemmas 5.7 and 5.8. Indeed, applying these lemmas to the map $S := \Theta L$, there exists a set $\{\boldsymbol{a_1}, \dots, \boldsymbol{a_u}\} \subset \mathbb{Z}^u$ that is a basis for the lattice $\Theta L(\mathbb{Z}^d)$ and for which $\|\boldsymbol{a_i}\|_\infty = O_{c,C}(1)$ for each i. Also, there exists a set of vectors $\{\boldsymbol{x_1}, \dots, \boldsymbol{x_u}\} \subset \mathbb{Z}^d$ such that $\Theta L(\boldsymbol{x_i}) = \boldsymbol{a_i}$ for each i, and $\|\boldsymbol{x_i}\|_\infty = O_{c,C}(1)$. By Lemma 5.8,

$$\mathcal{B} := \{\boldsymbol{x_i} : i \leqslant u\} \cup \{\Xi(\boldsymbol{w_j}) : j \leqslant d - u\} \tag{5-4}$$

is a basis for $\mathbb{R}^d$ and a lattice basis for $\mathbb{Z}^d$, where $\boldsymbol{w_1}, \dots, \boldsymbol{w_{d-u}}$ denotes the standard basis of $\mathbb{R}^{d-u}$.

Part (3): By the definition of $\tilde{R}$, and the fact that $\Xi(\mathbb{Z}^{d-u}) = \mathbb{Z}^d \cap \ker(\Theta L)$, we have

$$T^L_{F,G,N}(f_1, \dots, f_d) = \sum_{\tilde{\boldsymbol{r}} \in \tilde{R}} \frac{1}{N^{d-m}} \sum_{\boldsymbol{n} \in \mathbb{Z}^{d-u}} \left(\prod_{j=1}^{d} f_j(\xi_j(\boldsymbol{n}) + \tilde{\boldsymbol{r}}_j) \right) F(\Xi(\boldsymbol{n}) + \tilde{\boldsymbol{r}}) G(L\Xi(\boldsymbol{n}) + L\tilde{\boldsymbol{r}}), \tag{5-5}$$

where $\tilde{\boldsymbol{r}}_j$ denotes the j-th coordinates of $\tilde{\boldsymbol{r}}$. Now by an easy linear algebraic argument (recorded in Lemma D.4),

$$\mathbb{R}^m = \operatorname{span}(L\boldsymbol{x_i} : i \leqslant u) \oplus \ker \Theta \tag{5-6}$$

as an algebraic direct sum, and there exists an invertible linear map $P : \mathbb{R}^m \to \mathbb{R}^m$ such that

$$P((\operatorname{span}(L\boldsymbol{x_i} : i \leqslant u))) = \mathbb{R}^u \times \{0\}^{m-u}, \tag{5-7}$$

$$P(\ker \Theta) = \{0\}^u \times \mathbb{R}^{m-u}, \tag{5-8}$$

and both $\|P\|_\infty = O_{c,C}(1)$ and $\|P^{-1}\|_\infty = O_{c,C}(1)$.

We have

$$G(L\Xi(\boldsymbol{n}) + L\tilde{\boldsymbol{r}}) = (G \circ P^{-1})(PL\Xi(\boldsymbol{n}) + PL\tilde{\boldsymbol{r}}),$$

and we note that $PL\Xi(\boldsymbol{n}) \in \{0\}^u \times \mathbb{R}^{m-u}$ for every $\boldsymbol{n} \in \mathbb{Z}^{d-u}$. Define $G_{\tilde{\boldsymbol{r}}} : \mathbb{R}^{m-u} \to [0, 1]$ by

$$G_{\tilde{\boldsymbol{r}}}(\boldsymbol{x}) := (G \circ P^{-1})(\boldsymbol{x_0} + PL\tilde{\boldsymbol{r}}),$$

where $\boldsymbol{x_0}$ is the extension of $\boldsymbol{x}$ by 0 in the first u coordinates. Then the function $G_{\tilde{\boldsymbol{r}}}$ is the indicator function of a convex set contained in $[-O_{c,C,\varepsilon}(1), O_{c,C,\varepsilon}(1)]^{m-u}$.

Define

$$F_{\tilde{\boldsymbol{r}}}(\boldsymbol{n}) := F(\Xi(\boldsymbol{n}) + \tilde{\boldsymbol{r}}).$$

Then $F_{\tilde{\boldsymbol{r}}}$ has Lipschitz constant $O_{c,C}(1/\sigma_F N)$ and $F_{\tilde{\boldsymbol{r}}}$ is supported on $[-O_{c,C,\varepsilon}(N), O_{c,C,\varepsilon}(N)]^{d-u}$. (For a full proof of this fact, apply Lemma D.3 to the map Ξ). So $F_{\tilde{\boldsymbol{r}}}$ and $G_{\tilde{\boldsymbol{r}}}$ satisfy part (3).

Part (4): Writing $\pi_{m-u} : \mathbb{R}^m \to \mathbb{R}^{m-u}$ for the projection onto the final $m-u$ coordinates, expression (5-5) is equal to

$$\sum_{\tilde{\boldsymbol{r}} \in \tilde{R}} \frac{1}{N^{d-m}} \sum_{\boldsymbol{n} \in \mathbb{Z}^{d-u}} \left(\prod_{j=1}^{d} f_j(\xi_j(\boldsymbol{n}) + \tilde{\boldsymbol{r}}_j) \right) F_{\tilde{\boldsymbol{r}}}(\boldsymbol{n}) G_{\tilde{\boldsymbol{r}}}(\pi_{m-u} PL\Xi(\boldsymbol{n})). \tag{5-9}$$

Let

$$L' := \pi_{m-u} PL\Xi. \tag{5-10}$$

Then $L' : \mathbb{R}^{d-u} \to \mathbb{R}^{m-u}$ is surjective, and

$$T^L_{F,G,N}(f_1, \dots, f_d) = \sum_{\tilde{\boldsymbol{r}} \in \tilde{R}} T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}}},N}(f_1, \dots, f_d).$$

This resolves part (4).

Part (5): We wish to show that L' is purely irrational. Suppose for contradiction that there exists some surjective linear map $\varphi : \mathbb{R}^{m-u} \to \mathbb{R}$ with $\varphi L'(\mathbb{Z}^{d-u}) \subseteq \mathbb{Z}$, i.e., with $\varphi\pi_{m-u}PL\Xi(\mathbb{Z}^{d-u}) \subseteq \mathbb{Z}$. Then define the map $\Theta' : \mathbb{R}^m \to \mathbb{R}^{u+1}$ by

$$\Theta'(\boldsymbol{x}) := (\Theta(\boldsymbol{x}), \varphi\pi_{m-u}P(\boldsymbol{x})).$$

Then Θ' is surjective, and $\Theta' L(\mathbb{Z}^d) \subseteq \mathbb{Z}^{u+1}$. (This second fact is immediately seen by writing $\mathbb{Z}^d$ with respect to the lattice basis $\mathcal{B}$ from (5-4)). This contradicts the assumption that L has rational dimension u. So L' is purely irrational.

Part (6): The bound $\|L'\|_\infty = O_{c,C}(1)$ follows immediately from the bounds on the coefficients of Ξ, L, P, and π_{m-u} separately.

We wish to prove that

$$\operatorname{dist}(L', V_{\text{rank}}(m-u, d-u)) \gg_{c,C} 1,$$

i.e., that

$$\operatorname{dist}(\pi_{m-u}PL\Xi, V_{\text{rank}}(m-u, d-u)) \gg_{c,C} 1.$$

Suppose for contradiction that, for a small parameter η, there exists a linear map $Q : \mathbb{R}^{d-u} \to \mathbb{R}^{m-u}$ such that $\|Q\|_\infty < \eta$ and $\pi_{m-u}PL\Xi + Q$ has rank less than $m-u$. Recall that $PL\Xi(\mathbb{R}^{d-u}) = \{0\}^u \times \mathbb{R}^{m-u}$. So, extending Q by zeros to a map $Q : \mathbb{R}^{d-u} \to \{0\}^u \times \mathbb{R}^{m-u}$, and applying P^{-1}, there is a map $Q' : \mathbb{R}^{d-u} \to \mathbb{R}^m$ such that $\|Q'\|_\infty = O_{c,C}(\eta)$ and $L\Xi + Q'$ has rank less than $m-u$.

We may factorise $Q' = H\Xi$ for some m-by-d matrix H. Indeed let

$$\mathcal{B} := \{\boldsymbol{x_i} : i \leqslant u\} \cup \{\Xi(\boldsymbol{w_j}) : j \leqslant d-u\}$$

be the basis of $\mathbb{R}^d$ from (5-4), i.e., the basis formed by applying Lemma 5.8 to the map $S := \Theta L$. Define the linear map H by $H(\Xi(\boldsymbol{w_j})) := Q'(\boldsymbol{w_j})$ for each j and $H(\boldsymbol{x_i}) := \boldsymbol{0}$ for each i. Since the change of basis matrix between $\mathcal{B}$ and the standard basis of $\mathbb{R}^d$ has integer coefficients with absolute values at most $O_{c,C}(1)$, it follows that the matrix representing H with respect to the standard bases satisfies $\|H\|_\infty = O_{c,C}(\eta)$.

So we know that $(L+H)\Xi$ has rank less than $m-u$. But $\Xi : \mathbb{R}^{d-u} \to \mathbb{R}^d$ is injective, so this implies that the rank of $L+H$ is less than m. Hence $\operatorname{dist}(L, V_{\text{rank}}(m,d)) = O_{c,C}(\eta)$, which contradicts the assumptions of the lemma (if η is small enough). So $\operatorname{dist}(L', V_{\text{rank}}(m-u, d-u)) \gg_{c,C} 1$ as required.

Part (7): We wish to show that $\operatorname{dist}((\Xi, L'), V^*_{\text{degen},2}(m-u, d, d-u)) = \Omega_{c,C}(1)$. Suppose for contradiction that, for a small parameter η, there exists a linear map $Q : \mathbb{R}^{d-u} \to \mathbb{R}^d$ such that $\|Q\|_\infty \leqslant \eta$ and $\operatorname{dist}((\Xi + Q, L'), V^*_{\text{degen},2}(m-u, d, d-u)) \leqslant \eta$. In other words, we suppose there exist two indices $i, j \leqslant d$, and a real number λ, such that $\boldsymbol{e}_i^* - \lambda\boldsymbol{e}_j^*$ is nonzero and

$$(\Xi + Q)^*(\boldsymbol{e}_i^* - \lambda\boldsymbol{e}_j^*) \in (L')^*((\mathbb{R}^{m-u})^*),$$

where $\{\boldsymbol{e_1}, \dots, \boldsymbol{e_d}\}$ denotes the standard basis of $\mathbb{R}^d$ and $\{\boldsymbol{e_1^*}, \dots, \boldsymbol{e_d^*}\}$ denotes the dual basis. Expanding out the definition of L', this means that there exists some $\varphi \in (\mathbb{R}^{m-u})^*$ such that

$$\Xi^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*} - L^*(P^*\pi_{m-u}^*(\varphi))) = -Q^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}).$$

Because $\|Q\|_\infty \leqslant \eta$, this means that

$$\|\Xi^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*} - L^*(P^*\pi_{m-u}^*(\varphi)))\|_\infty = O(\eta \|\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}\|_\infty). \tag{5-11}$$

Let

$$\mathcal{B}^* := \{\boldsymbol{x_i^*} : i \leqslant u\} \cup \{\Xi(\boldsymbol{w_j})^* : j \leqslant d-u\} \tag{5-12}$$

denote the basis of $(\mathbb{R}^d)^*$ that is dual to the basis $\mathcal{B}$ from (5-4). It follows from part (4) of Lemma 5.8 and (5-11) that

$$\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*} - L^*(P^*\pi_{m-u}^*(\varphi)) = \omega_V + \omega_W,$$

where $\omega_V \in L^*\Theta^*((\mathbb{R}^u)^*)$, $\omega_W \in \operatorname{span}(\Xi(\boldsymbol{w_j})^* : j \leqslant d-u)$, and $\|\omega_W\|_\infty = O_{c,C}(\eta \|\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}\|_\infty)$. So therefore

$$\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*} = L^*(\alpha) + \omega_W,$$

for some $\alpha \in (\mathbb{R}^m)^*$.

Therefore $\|\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*} - \omega_W\|_\infty \geqslant \frac{1}{2}\|\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}\|_\infty$, provided η is small enough. Since $\|L^*\|_\infty = O_{c,C}(1)$, we conclude that $\|\alpha\|_\infty = \Omega_{c,C}(\|\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}\|_\infty)$.

This means that there exists a linear map $E : \mathbb{R}^d \to \mathbb{R}^m$ with $\|E\|_\infty = O_{c,C}(\eta)$ for which $E^*(\alpha) = \omega_W$. Then

$$\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*} \in (L+E)^*((\mathbb{R}^m)^*),$$

and hence $\operatorname{dist}(L, V^*_{\text{degen}}(m,d)) = O_{c,C}(\eta)$. This is a contradiction to the hypotheses of Theorem 2.12, provided η is small enough, and hence $\operatorname{dist}((\Xi, L'), V^*_{\text{degen},2}(m-u, d, d-u)) = \Omega_{c,C}(1)$.

Part (8): Let $\tau_1, \tau_2 \in (0,1]$. We desire to prove the relationship

$$A_{L'}(\tau_1, \tau_2) \gg_{c,C} A_L(\Omega_{c,C}(\tau_1), \Omega_{c,C}(\tau_2)), \tag{5-13}$$

where L' is as in (5-10).

We have already proved that L' is purely irrational (that was part (5) of the lemma). So, if $A_{L'}(\tau_1, \tau_2) < \eta$, for some η, there exists some $\varphi \in (\mathbb{R}^{m-u})^*$ for which $\tau_1 \leqslant \|\varphi\|_\infty \leqslant \tau_2^{-1}$ and for which

$$\operatorname{dist}((\pi_{m-u} P L \Xi)^*(\varphi), (\mathbb{Z}^{d-u})^T) < \eta,$$

where, one recalls, we use $(\mathbb{Z}^{d-u})^T$ to denote the set of those functions in $(\mathbb{R}^{d-u})^*$ that have integer coordinates with respect to the standard dual basis.

We claim that

$$\operatorname{dist}(L^*(P^*\pi^*_{m-u}(\varphi)), (\mathbb{Z}^d)^T) \ll_{c,C} \eta; \tag{5-14}$$

$$\|P^*\pi^*_{m-u}(\varphi)\|_\infty \ll_{c,C} \tau_2^{-1}; \tag{5-15}$$

$$\operatorname{dist}(P^*\pi^*_{m-u}(\varphi), \Theta^*((\mathbb{R}^u)^*)) \gg_{c,C} \tau_1, \tag{5-16}$$

from which (5-13) immediately follows.

Let us prove (5-14). Indeed, we already know that $\operatorname{dist}(\Xi^*L^*P^*\pi^*_{m-u}(\varphi), (\mathbb{Z}^{d-u})^T) < \eta$, i.e., that

$$\|\Xi^*L^*P^*\pi^*_{m-u}(\varphi) - \alpha\|_\infty < \eta, \tag{5-17}$$

for some $\alpha \in (\mathbb{Z}^{d-u})^T$. Let us write $\alpha = \sum_{j=1}^{d-u} \lambda_j \boldsymbol{w}_j^*$ for some $\lambda_j \in \mathbb{Z}$, where $\boldsymbol{w_1}, \dots, \boldsymbol{w_{d-u}}$ denotes the standard basis for $\mathbb{R}^{d-u}$ and $\boldsymbol{w}_1^*, \dots, \boldsymbol{w}_{d-u}^*$ denotes the dual basis. Let $\mathcal{B}^*$ be as in (5-12). Then $\boldsymbol{w}_j^* = \Xi^*((\Xi(\boldsymbol{w}_j)^*)$, and so

$$\alpha = \Xi^*\Bigg(\sum_{j=1}^{d-u} \lambda_j \Xi(\boldsymbol{w}_j)^*\Bigg).$$

So from (5-17) and the final part of Lemma 5.8,

$$L^*P^*\pi^*_{m-u}(\varphi) - \sum_{j=1}^{d-u} \lambda_j \Xi(\boldsymbol{w}_j)^* = \omega_V + \omega_W, \tag{5-18}$$

where $\omega_V \in \operatorname{span}(\boldsymbol{x}_i^* : i \leqslant u)$, $\omega_W \in \operatorname{span}(\Xi(\boldsymbol{w}_j)^* : j \leqslant d-u)$, and $\|\omega_W\|_\infty = O_{c,C}(\eta)$.

But $L^*P^*\pi^*_{m-u}(\varphi) \in \operatorname{span}(\Xi(\boldsymbol{w}_j)^* : j \leqslant d-u)$ too. Indeed, for every i at most $d-u$,

$$L^*P^*\pi^*_{m-u}(\varphi)(\boldsymbol{x}_i) = \varphi(\pi_{m-u}PL\boldsymbol{x}_i) = \varphi(\boldsymbol{0}) = 0,$$

by the properties of P (see (5-7)). Therefore $\omega_V = \boldsymbol{0}$, and so

$$\Bigg\|L^*P^*\pi^*_{m-u}(\varphi) - \sum_{j=1}^{d-u} \lambda_j \Xi(\boldsymbol{w}_j)^*\Bigg\|_\infty = O_{c,C}(\eta).$$

Since $\sum_{j=1}^{d-u} \lambda_j \Xi(\boldsymbol{w}_j)^* \in (\mathbb{Z}^d)^T$, this implies (5-14) as claimed.

The bound (5-15) is immediate from the bounds on the coefficients of P^* and π^*_{m-u}, so it remains to prove (5-16). Suppose for contradiction that, for some small parameter δ,

$$P^*\pi^*_{m-u}(\varphi) = \alpha_1 + \alpha_2,$$

where $\alpha_1 \in \Theta^*((\mathbb{R}^u)^*)$ and $\|\alpha_2\|_\infty \leqslant \delta\tau_1$. We know that $\|\varphi\|_\infty \geqslant \tau_1$, which means that there is some standard basis vector $\boldsymbol{f_k} \in \mathbb{R}^{m-u}$ for which $|\varphi(\boldsymbol{f_k})| \geqslant \tau_1$. Let $\boldsymbol{b_{k+u}}$ be the standard basis vector of $\mathbb{R}^m$ for which $\pi_{m-u}(\boldsymbol{b_{k+u}}) = \boldsymbol{f_k}$. Recall the properties of P (given in (5-7) and (5-8)), in particular recall that $P : \ker\Theta \to \{0\}^u \times \mathbb{R}^{m-u}$ is an isomorphism. Then

$$|P^*\pi^*_{m-u}(\varphi)(P^{-1}(\boldsymbol{b_{k+u}}))| = |\pi^*_{m-u}(\varphi)(\boldsymbol{b_{k+u}})| = |\varphi(\boldsymbol{f_k})| \geqslant \tau_1.$$

Note that $\Theta^*((\mathbb{R}^u)^*) = (\ker\Theta)^0$, and so

$$|P^*\pi^*_{m-u}(\varphi)(P^{-1}(\boldsymbol{b_{k+u}}))| = |(\alpha_1+\alpha_2)(P^{-1}(\boldsymbol{b_{k+u}}))| = |\alpha_2(P^{-1}(\boldsymbol{b_{k+u}}))| \ll_{c,C} \delta\tau_1,$$

as $P^{-1}(\boldsymbol{b_{k+u}}) \in \ker\Theta$ and satisfies $\|P^{-1}(\boldsymbol{b_{k+u}})\|_\infty = O_{c,C}(1)$. This is a contradiction if δ is small enough, and so (5-16) holds. This resolves part (8).

Part (9): Let $\tau_1, \tau_2 \in (0, 1]$. We desire to prove the relationship

$$A_{L'}(\tau_1, \tau_2) \ll_{c,C} A_L(\Omega_{c,C}(\tau_1), \Omega_{c,C}(\tau_2)), \tag{5-19}$$

where L' is as in (5-10). This inequality is the reverse inequality of part (8), and in fact it will not be required in the proof of any of our main theorems. However, it will be required in order to analyse $A_L(\tau_1, \tau_2)$ when L has algebraic coefficients (in Appendix E), so we choose to state and prove it here, close to our argument for part (8).

Suppose that $A_L(\tau_1, \tau_2) < \eta$, for some parameter η. Then there exists some $\varphi \in (\mathbb{R}^m)^*$ such that $\operatorname{dist}(\varphi, \Theta^*((\mathbb{R}^u)^*)) \geqslant \tau_1$, $\|\varphi\|_\infty \leqslant \tau_2^{-1}$, and $\operatorname{dist}(L^*\varphi, (\mathbb{Z}^d)^T) < \eta$. So there exists some $\omega \in (\mathbb{Z}^d)^T$ for which

$$\|L^*\varphi - \omega\|_\infty < \eta.$$

We expand both $L^*\varphi$ and ω with respect to the dual basis $\mathcal{B}^*$ from (5-12). So,

$$L^*\varphi = \sum_{i=1}^{u} \lambda_i \boldsymbol{x_i^*} + \sum_{j=1}^{d-u} \mu_j \Xi(\boldsymbol{w_j})^* \quad \text{and} \quad \omega = \sum_{i=1}^{u} \lambda_i' \boldsymbol{x_i^*} + \sum_{j=1}^{d-u} \mu_j' \Xi(\boldsymbol{w_j})^*.$$

Since $\mathcal{B}^*$ is a lattice basis for $(\mathbb{Z}^d)^T$, we have $\lambda_i' \in \mathbb{Z}$ and $\mu_j' \in \mathbb{Z}$ for each i and j. Since the change of basis matrix between $\mathcal{B}^*$ and the standard dual basis has integer coefficients that are bounded in absolute value by $O_{c,C}(1)$ (part (2) of Lemma 5.8), one has $|\lambda_i - \lambda_i'| = O_{c,C}(\eta)$ and $|\mu_j - \mu_j'| = O_{c,C}(\eta)$ for each i and j.

Let $\boldsymbol{w_1^*}, \dots, \boldsymbol{w_{d-u}^*}$ denote the standard dual basis of $(\mathbb{R}^{d-u})^*$, and define

$$\omega' := \sum_{j=1}^{d-u} \mu_j' \boldsymbol{w_j^*}.$$

Certainly $\omega' \in (\mathbb{Z}^{d-u})^T$. We claim that there exists a map $\varphi' \in (\mathbb{R}^{m-u})^*$ such that $\tau_1 \ll_{c,C} \|\varphi'\|_\infty \ll_{c,C} \tau_2^{-1}$ and $\|(L')^*\varphi' - \omega'\|_\infty \ll_{c,C} \eta$, which will immediately resolve (5-19) and part (9).

Indeed, recall the decomposition $\mathbb{R}^m = (\operatorname{span}(L\boldsymbol{x_i} : i \leqslant u)) \oplus \ker\Theta$ as an algebraic direct sum from (5-6). Let $\varphi = \varphi_1 + \varphi_2$, where $\varphi_1 \in (\operatorname{span}(L\boldsymbol{x_i} : i \leqslant u))^0$ and $\varphi_2 \in (\ker\Theta)^0$. Since $\operatorname{dist}(\varphi, (\ker\Theta)^0) \geqslant \tau_1$, we have $\|\varphi_1\|_\infty \geqslant \tau_1$. By the properties of the matrix P ((5-7) and (5-8)) there exists some $\varphi' \in (\mathbb{R}^{m-u})^*$ such that

$$\varphi_1 = P^*\pi^*_{m-u}\varphi'.$$

Furthermore, by evaluating φ' at the standard basis vectors, one sees that

$$\tau_1 \ll_{c,C} \|\varphi'\|_\infty \ll_{c,C} \tau_2^{-1}.$$

We shall use this φ'.

By evaluating $L^*\varphi_1$ at the elements of $\mathcal{B}$ one immediately sees that

$$L^*\varphi_1 = \sum_{j=1}^{d-u} \mu_j \Xi(\boldsymbol{w}_j)^*.$$

Hence

$$\Xi^* L^* P^* \pi_{m-u}^* \varphi' = \sum_{j=1}^{d-u} \mu_j \boldsymbol{w}_j^*,$$

in other words $(L')^*\varphi' = \sum_{j=1}^{d-u} \mu_j \boldsymbol{w}_j^*$. But since $|\mu_j - \mu_j'| = O_{c,C}(\eta)$ for each j, one has $\|(L')^*\varphi' - \omega'\|_\infty = O_{c,C}(\eta)$ as required. This settles part (9).

The entire lemma is settled. □

The final lemma we need in order to deduce Theorem 2.12 involves removing the sharp cutoff G.

Lemma 5.11 (removing image cutoff). *Let m, d, h be natural numbers, satisfying $d \geqslant h \geqslant m+1$. Let c, C, ε be positive, and let σ_G be any parameter in the range $0 < \sigma_G < \frac{1}{2}$. Let $L' : \mathbb{R}^h \to \mathbb{R}^m$ be a purely irrational surjective map, and let $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ be an injective map. Suppose that $\|L'\|_\infty \leqslant C$ and that $\operatorname{dist}(L', V_{\mathrm{rank}}(m,h)) \geqslant c$. Let $F_{\tilde{\boldsymbol{r}}} : \mathbb{R}^h \to [0,1]$ be any function supported on $[-N, N]^h$, and let $G_{\tilde{\boldsymbol{r}}} : \mathbb{R}^m \to [0,1]$ be the indicator function of a convex set contained within $[-\varepsilon, \varepsilon]^m$. Then there exists a Lipschitz function $G_{\tilde{\boldsymbol{r}},\sigma_G,1}$ supported on $[-O_{c,C,\varepsilon}(1), O_{c,C,\varepsilon}(1)]^m$, and with Lipschitz constant $O_{c,C,\varepsilon}(1/\sigma_G)$, such that, for any parameter τ_2 in the range $0 < \tau_2 \leqslant 1$ and for any functions $f_1, \dots, f_d : [N] \to [-1,1]$,*

$$|T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}}},N}(f_1,\dots,f_d)| \ll_{c,C,\varepsilon} |T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}},\sigma_G,1},N}(f_1,\dots,f_d)| + \sigma_G + \frac{\tau_2^{1/2}}{\sigma_G} + \frac{\tau_2^{-O(1)} A_L(\Omega_{c,C}(1),\tau_2)^{-1}}{N}.$$

Proof. Applying Lemma B.2 to the function $G_{\tilde{\boldsymbol{r}}}$, we have

$$G_{\tilde{\boldsymbol{r}}} = G_{\tilde{\boldsymbol{r}},\sigma_G,1} + O(G_{\tilde{\boldsymbol{r}},\sigma_G,2}),$$

where $G_{\tilde{\boldsymbol{r}},\sigma_G,1}, G_{\tilde{\boldsymbol{r}},\sigma_G,2} : \mathbb{R}^m \to [0,1]$ are Lipschitz functions with Lipschitz constant $O_{c,C,\varepsilon}(1/\sigma_G)$, both supported on $[-O_{c,C,\varepsilon}(1), O_{c,C,\varepsilon}(1)]^m$, and with $\int_{\boldsymbol{x}} G_{\tilde{\boldsymbol{r}},\sigma_G,2}(\boldsymbol{x})\,d\boldsymbol{x} = O_{c,C,\varepsilon}(\sigma_G)$.

By the triangle inequality,

$$|T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}},\sigma_G,2},N}(1,\dots,1)| \leqslant T^{L'}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}},\sigma_G,2},N}(1,\dots,1).$$

We now apply Lemma 3.4, with linear map L' and Lipschitz function $G_{\tilde{\boldsymbol{r}},\sigma_G,2}$. Inserting the bound from Lemma 3.4, the present lemma follows. □

We conclude this section by combining the three previous lemmas, along with Theorem 5.6, to deduce our main result.

Proof of Theorem 2.12 assuming Theorem 5.6. Assume the hypotheses of Theorem 2.12. Let σ_F and σ_G be any parameters satisfying $0 < \sigma_F, \sigma_G < \frac{1}{2}$, and let τ_2 be any parameter satisfying $0 < \tau_2 \leqslant 1$.

By Lemma 5.9,

$$|T^L_{F,G,N}(f_1, \dots, f_d)| \leqslant |T^L_{F_{1,\sigma_F},G,N}(f_1, \dots, f_d)| + O_{c,C}(\sigma_F),$$

for some function $F_{1,\sigma_F} : \mathbb{R}^d \to [0, 1]$ supported on $[-2N, 2N]^d$ and with Lipschitz constant $O(1/\sigma_F N)$. By part (4) of Lemma 5.10, writing F_{1,σ_F} for F, we have

$$|T^L_{F_{1,\sigma_F},G,N}(f_1, \dots, f_d)| \leqslant \sum_{\tilde{\boldsymbol{r}} \in \tilde{R}} |T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}}},N}(f_1, \dots, f_d)|,$$

where the objects $F_{\tilde{\boldsymbol{r}}}$, $G_{\tilde{\boldsymbol{r}}}$, L', Ξ and $\tilde{R}$ satisfy all the conclusions of that lemma.

Parts (1), (5) and (6) of Lemma 5.10 show that Ξ and L' satisfy the hypotheses of Lemma 5.11, where in the notation of Lemma 5.11 we take $h := d - u$ and rewrite m for $m - u$. So, applying Lemma 5.11, there are some Lipschitz functions $G_{\tilde{\boldsymbol{r}},\sigma_G,1} : \mathbb{R}^{m-u} \to [0, 1]$ supported on $[-O_{c,C,\varepsilon}(1), O_{c,C,\varepsilon}(1)]^{m-u}$ and with Lipschitz constant $O_{c,C,\varepsilon}(1/\sigma_G)$ such that

$$|T^L_{F,G,N}(f_1, \dots, f_d)| \ll_{c,C,\varepsilon} \sum_{\tilde{\boldsymbol{r}} \in \tilde{R}} |T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}},\sigma_G,1},N}(f_1, \dots, f_d)| + \sigma_G + \frac{\tau_2^{1/2}}{\sigma_G} + \frac{\tau_2^{-O(1)} A_{L'}(\Omega_{c,C}(1), \tau_2)^{-1}}{N} + \sigma_F. \quad (5\text{-}20)$$

(Recall that $|\tilde{R}| = O_{c,C,\varepsilon}(1)$, by part (2) of Lemma 5.10.)

By conclusion (8) of Lemma 5.10, we may replace the term $A_{L'}(\Omega_{c,C}(1), \tau_2)^{-1}$ with the term $A_L(\Omega_{c,C}(1), \Omega_{c,C}(\tau_2))^{-1}$.

Since $F_{\tilde{\boldsymbol{r}}}$, L', Ξ, and $\tilde{R}$ together satisfy conclusions (1), (2), (3), (6), and (7) of Lemma 5.10, the hypotheses are satisfied so that we may apply Theorem 5.6 to the expression $T^{L',\Xi,\tilde{\boldsymbol{r}}}_{F_{\tilde{\boldsymbol{r}}},G_{\tilde{\boldsymbol{r}},\sigma_G,1},N}(f_1, \dots, f_d)$. (We take $h = d - u$ and rewrite m for $m - u$, as above). Therefore there exists an s at most $d - 2$, independent of $F_{\tilde{\boldsymbol{r}}}$, $G_{\tilde{\boldsymbol{r}}}$ and $\tilde{\boldsymbol{r}}$, such that, if

$$\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho,$$

for some ρ in the range $0 < \rho \leqslant 1$ then $|T^L_{F,G,N}(f_1, \dots, f_d)|$ is

$$\ll_{c,C,\varepsilon} \rho^{\Omega(1)}(\sigma_F^{-O(1)} + \sigma_G^{-O(1)}) + \sigma_F^{-O(1)} N^{-\Omega(1)} + \sigma_G + \frac{\tau_2^{1/2}}{\sigma_G} + \frac{\tau_2^{-O(1)} A_L(\Omega_{c,C}(1), \Omega_{c,C}(\tau_2))^{-1}}{N} + \sigma_F. \quad (5\text{-}21)$$

It remains to pick appropriate parameters. Let C_1 be a constant that is suitably large in terms of c, C, and all $O(1)$ constants, and let c_1 be a constant that is suitably small in terms of all $O(1)$ constants. Pick $\sigma_F := \sigma_G := \rho^{c_1}$ and $\tau_2 := C_1\rho$. Then

$$|T^L_{F,G,N}(f_1, \dots, f_d)| \ll_{c,C,\varepsilon} \rho^{\Omega(1)} + o_{\rho,A_L,c,C}(1)$$

as $N \to \infty$, where, after the combining the various error terms from (5-21), the $o_{\rho, A_L, c, C}(1)$ term may be bounded above by

$$N^{-\Omega(1)}\rho^{-O(1)}A_L(\Omega_{c,C}(1), \rho)^{-1},$$

as $A_L(\tau_1, \tau_2)$ is monotonically decreasing as τ_2 decreases. This is the desired conclusion of Theorem 2.12. □

6. Transfer from $\mathbb{Z}$ to $\mathbb{R}$

Our remaining task is to prove Theorem 5.6. We devote this section to the formulation and proof of a certain "transfer" argument, whereby we replace the discrete summation in the definition of $T^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1, \dots, f_d)$ with an integral $\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)$. This manoeuvre will be extremely useful in the sequel, as it gives us access to the standard techniques of manipulating real integrals (in particular reparametrisation of variables). These reparametrisations may be attempted directly in the context of the discrete summation $T^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1, \dots, f_d)$, but the results will be messy, and one will need to control the error term each time such a reparametrisation is undertaken. It is easier in our view to do a single approximation at the beginning, so that we may subsequently reparametrise at will. As we remarked in Section 2E, there is a somewhat analogous device in [Green and Tao 2010a], in which the authors transfer their combinatorial expressions into summations over a field (a finite field $\mathbb{Z}/N'\mathbb{Z}$ for some prime N', in their case), in order that their algebraic manipulations may be simplified. The natural field to use in our setting is $\mathbb{R}$.

Let us introduce some notation for the integral in question.

Definition 6.1. Let N, m, d, h be natural numbers, with $d \geqslant h \geqslant m+2$. Let ε be positive. Let $\Xi = (\xi_1, \dots, \xi_d) : \mathbb{R}^h \to \mathbb{R}^d$ and $L : \mathbb{R}^h \to \mathbb{R}^m$ be linear maps. Let $F : \mathbb{R}^h \to [0, 1]$ and $G : \mathbb{R}^m \to [0, 1]$ be two functions, with F supported on $[-N, N]^h$ and G supported on $[-\varepsilon, \varepsilon]^m$. Let $g_1, \dots, g_d : \mathbb{R} \to [-1, 1]$ be arbitrary functions. We define

$$\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d) := \frac{1}{N^{h-m}} \int_{\boldsymbol{x} \in \mathbb{R}^h} \Bigg(\prod_{j=1}^{d} g_j(\xi_j(\boldsymbol{x}) + \tilde{\boldsymbol{r}}_j) \Bigg) F(\boldsymbol{x}) G(L\boldsymbol{x})\, d\boldsymbol{x}. \tag{6-1}$$

Next, we determine a particular class of measurable functions that will be useful to us.

Definition 6.2 (η-supported). Let $\chi : \mathbb{R} \to [0, 1]$ be a measurable function, and let η be a positive parameter. We say that χ is *η-supported* if χ is supported on $[-\eta, \eta]$ and $\chi(x) \equiv 1$ for all $x \in [-\eta/2, \eta/2]$.

Definition 6.3 (convolution). If $f : \mathbb{Z} \to \mathbb{R}$ has finite support, and $\chi : \mathbb{R} \to [0, 1]$ is a measurable function, we may define the (rather singular) convolution $(f * \chi)(x) : \mathbb{R} \to \mathbb{R}$ by

$$(f * \chi)(x) := \sum_{n \in \mathbb{Z}} f(n)\chi(x - n).$$

We note that if χ is η-supported, for small enough η, then there is only one possible integer n that makes a nonzero contribution to above summation.

We now state the key lemma.

Lemma 6.4. *Let N, m, d, h be natural numbers, with $d \geqslant h \geqslant m+2$, and let c, C, ε, η be positive constants. Let $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ be an injective linear map with integer coefficients, and assume that $\Xi(\mathbb{Z}^h) = \mathbb{Z}^d \cap \operatorname{im} \Xi$. Let $L : \mathbb{R}^h \to \mathbb{R}^m$ be a surjective linear map. Assume that $\|\Xi\|_\infty \leqslant C$, $\|L\|_\infty \leqslant C$, and $\operatorname{dist}(L, V_{\operatorname{rank}}(m,h)) \geqslant c$. Let $F : \mathbb{R}^h \to [0,1]$ be a Lipschitz function supported on $[-N, N]^h$ with Lipschitz constant $O(1/\sigma_F N)$, and let $G : \mathbb{R}^m \to [0,1]$ be a Lipschitz function supported on $[-\varepsilon, \varepsilon]^m$ with Lipschitz constant $O(1/\sigma_G)$. Let $\tilde{\boldsymbol{r}}$ be a fixed vector in $\mathbb{Z}^d$, satisfying $\|\tilde{\boldsymbol{r}}\|_\infty = O_{c,C,\varepsilon}(1)$. Let $\chi : \mathbb{R} \to [0,1]$ be an η-supported measurable function. Then, if η is small enough (in terms of the dimensions m, d, h, C, and ε) there exists some positive real number $C_{\Xi,\chi}$ such that, if $f_1, \dots, f_d : [N] \to [-1,1]$ are arbitrary functions,*

$$T^{\Xi,L,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1, \dots, f_d) = \frac{1}{C_{\Xi,\chi}\eta^h} \tilde{T}^{\Xi,L,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1 * \chi, \dots, f_d * \chi) + O_{C,c,\varepsilon}(\eta/\sigma_G) + O_{C,c,\varepsilon}(\eta/\sigma_F N). \quad (6\text{-}2)$$

Moreover, $C_{\Xi,\chi} \asymp_C 1$.

This lemma is a rigorous formulation of (2-12) from the proof strategy in Section 2E. It is in fact the only part of the proof of Theorem 5.6 in which we use the fact that G is Lipschitz.

Proof. Let $\boldsymbol{\chi} : \mathbb{R}^d \to [0,1]$ denote the function $\boldsymbol{x} \mapsto \prod_{i=1}^d \chi(x_i)$. We choose

$$C_{\Xi,\chi} := \frac{1}{\eta^h} \int_{\boldsymbol{x} \in \mathbb{R}^h} \boldsymbol{\chi}(\Xi(\boldsymbol{x}))\, d\boldsymbol{x}.$$

Since χ is η-supported, $C_{\Xi,\chi} \asymp_C 1$.

Then, expanding the definition of the convolution,

$$\frac{1}{C_{\Xi,\chi}\eta^h} \tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1 * \chi, \dots, f_d * \chi)$$

equals

$$\frac{1}{N^{h-m}} \sum_{\boldsymbol{n} \in \mathbb{Z}^d} \left(\prod_{j=1}^d f_j(n_j) \right) \frac{1}{C_{\Xi,\chi}\eta^h} \int_{\boldsymbol{y} \in \mathbb{R}^h} F(\boldsymbol{y}) G(L\boldsymbol{y}) \boldsymbol{\chi}(\Xi(\boldsymbol{y}) + \tilde{\boldsymbol{r}} - \boldsymbol{n})\, d\boldsymbol{y}. \quad (6\text{-}3)$$

Note that any vector $\boldsymbol{n} \in \mathbb{Z}^d$ that gives a nonzero contribution to expression (6-3) satisfies

$$\|\boldsymbol{n} - \Xi(\boldsymbol{y}) - \tilde{\boldsymbol{r}}\|_\infty \ll \eta,$$

for some $\boldsymbol{y} \in \mathbb{R}^h$. Therefore, $\boldsymbol{n}$ must be of the form $\Xi(\boldsymbol{n}') + \tilde{\boldsymbol{r}}$ for some unique $\boldsymbol{n}' \in \mathbb{Z}^h$. (This is proved in full in Lemma D.2). Therefore, writing $\Xi = (\xi_1, \dots, \xi_d)$, we may reformulate (6-3) as

$$\frac{1}{N^{h-m}} \sum_{\boldsymbol{n} \in \mathbb{Z}^h} \left(\prod_{j=1}^d f_j(\xi_j(\boldsymbol{n}) + \tilde{\boldsymbol{r}}_j) \right) \frac{1}{C_{\Xi,\chi}\eta^h} \int_{\boldsymbol{y} \in \mathbb{R}^h} F(\boldsymbol{y}) G(L\boldsymbol{y}) \boldsymbol{\chi}(\Xi(\boldsymbol{y} - \boldsymbol{n}))\, d\boldsymbol{y},$$

which is equal to

$$\frac{1}{N^{h-m}}\sum_{\boldsymbol{n}\in\mathbb{Z}^h}\Big(\prod_{j=1}^{d} f_j(\xi_j(\boldsymbol{n})+\tilde{\boldsymbol{r}}_j)\Big)\frac{1}{C_{\Xi,\chi}\eta^h}\int_{\boldsymbol{y}\in\mathbb{R}^h}(F(\boldsymbol{n})+O_C(\eta/\sigma_F N))G(L\boldsymbol{y})\boldsymbol{\chi}(\Xi(\boldsymbol{y}-\boldsymbol{n}))\,d\boldsymbol{y}. \quad (6\text{-}4)$$

Indeed, the inner integral is only nonzero when $\|\Xi(\boldsymbol{y})-\Xi(\boldsymbol{n})\|_\infty \ll \eta$, and this implies that $\|\boldsymbol{y}-\boldsymbol{n}\|_\infty \ll C^{-O(1)}\eta$. (This is proved in full in Lemma D.3). Then recall that F has Lipschitz constant $O(1/\sigma_F N)$.

Continuing, expression (6-4) is equal to

$$\frac{1}{N^{h-m}}\sum_{\boldsymbol{n}\in\mathbb{Z}^h}\Big(\prod_{j=1}^{d} f_j(\xi_j(\boldsymbol{n})+\tilde{\boldsymbol{r}}_j)\Big)F(\boldsymbol{n})H(L\boldsymbol{n})+E \quad (6\text{-}5)$$

where

$$H(\boldsymbol{x})=\frac{1}{C_{\Xi,\chi}\eta^h}\int_{\boldsymbol{y}\in\mathbb{R}^h}\boldsymbol{\chi}(\Xi(\boldsymbol{y}))G(\boldsymbol{x}+L\boldsymbol{y})\;\,d\boldsymbol{y}$$

and E is a certain error, which may be bounded above by

$$\ll_C \frac{\eta}{\sigma_F N}\frac{1}{N^{h-m}}\sum_{\boldsymbol{n}\in[-O(N),O(N)]^h}H(L\boldsymbol{n}). \quad (6\text{-}6)$$

Let us deal with the first term of (6-5), in which we wish to replace H with G. We therefore consider

$$\Big|\frac{1}{N^{h-m}}\sum_{\boldsymbol{n}\in\mathbb{Z}^h}\Big(\prod_{j=1}^{d} f_j(\xi_j(\boldsymbol{n})+\tilde{\boldsymbol{r}}_j)\Big)F(\boldsymbol{n})(G(L\boldsymbol{n})-H(L\boldsymbol{n}))\Big|,$$

which is

$$\leqslant \frac{1}{N^{h-m}}\sum_{\boldsymbol{n}\in\mathbb{Z}^h}F(\boldsymbol{n})|G-H|(L\boldsymbol{n}). \quad (6\text{-}7)$$

Using Lemma D.3 again, the function H is supported on $[-\varepsilon-O_C(\eta),\varepsilon+O_C(\eta)]^m$. Thus, if η is small enough in terms of ε, the function $|G-H|:\mathbb{R}^m\to\mathbb{R}$ is supported on $[-O_C(\varepsilon),O_C(\varepsilon)]^m$. Furthermore, $\|G-H\|_\infty = O_C(\eta/\sigma_G)$. Indeed,

$$\begin{aligned}G(\boldsymbol{x})-\frac{1}{C_{\Xi,\chi}\eta^h}\int_{\boldsymbol{y}\in\mathbb{R}^h}G(\boldsymbol{x}+L\boldsymbol{y})\boldsymbol{\chi}(\Xi(\boldsymbol{y}))\,d\boldsymbol{y} &= G(\boldsymbol{x})-\frac{1}{C_{\Xi,\chi}\eta^h}\int_{\boldsymbol{y}\in\mathbb{R}^h}(G(\boldsymbol{x})+O_C(\eta/\sigma_G))\boldsymbol{\chi}(\Xi(\boldsymbol{y}))\,d\boldsymbol{y}\\ &= O_C(\eta/\sigma_G),\end{aligned}$$

by the definition of $C_{\Xi,\chi}$ and using the Lipschitz property of G. So, by the crude bound given in Lemma 3.2, (6-7) may be bounded above by $O_{c,C,\varepsilon}(\eta/\sigma_G)$.

Turning to the error E from (6-5), we've already remarked that it may be bounded above by expression (6-6). Applying Lemma 3.2 again, expression (6-6) may be bounded above by $O_{c,C,\varepsilon}(\eta/\sigma_F N)$.

Lemma 6.4 follows immediately upon substituting the estimates on (6-6) and (6-7) into (6-5). □

We finish this section by noting a simple relationship between the Gowers norms $\|f*\chi\|_{U^{s+1}(\mathbb{R},2N)}$ and the Gowers norms $\|f\|_{U^{s+1}[N]}$.

Lemma 6.5 (relating different Gowers norms). *Let s be a natural number, and assume that η is a positive parameter that is small enough in terms of s. Let $\chi : \mathbb{R} \to [0, 1]$ be an η-supported measurable function. Let N be a natural number, and let $f : [N] \to \mathbb{R}$ be an arbitrary function. View $f * \chi$ as a function supported on $[-2N, 2N]$. Then we have*

$$\|f * \chi\|_{U^{s+1}(\mathbb{R},2N)} \ll \eta^{(s+2)/2^{s+1}} \|f\|_{U^{s+1}[N]}. \tag{6-8}$$

The definition of the real Gowers norm $\|f * \chi\|_{U^{s+1}(\mathbb{R},2N)}$ is recorded in Definition A.3.

Proof. From expression (A-5), we have

$$\|f * \chi\|^{2^{s+1}}_{U^{s+1}(\mathbb{R},2N)} \ll \frac{1}{N^{s+2}} \int_{(x,\boldsymbol{h})\in\mathbb{R}^{s+2}} \prod_{\boldsymbol{\omega}\in\{0,1\}^{s+1}} (f * \chi)(x + \boldsymbol{h}\cdot\boldsymbol{\omega})\, dx\, d\boldsymbol{h}.$$

Substituting in the definition of $f * \chi$, this is equal to

$$\frac{1}{N^{s+2}} \sum_{(n_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\{0,1\}^{s+1}}\in\mathbb{Z}^{\{0,1\}^{s+1}}} \Bigg(\prod_{\boldsymbol{\omega}\in\{0,1\}^{s+1}} f(n_{\boldsymbol{\omega}})\Bigg) \int_{(x,\boldsymbol{h})\in\mathbb{R}^{s+2}} \boldsymbol{\chi}(\Psi(x,\boldsymbol{h}) - \boldsymbol{n})\, dx\, d\boldsymbol{h}, \tag{6-9}$$

where $\Psi : \mathbb{R}^{s+2} \to \mathbb{R}^{2^{s+1}}$ has coordinate functions $\psi_{\boldsymbol{\omega}}$, indexed by $\boldsymbol{\omega} \in \{0, 1\}^{s+1}$, where $\psi_{\boldsymbol{\omega}}(x, \boldsymbol{h}) := x + \boldsymbol{h}\cdot\boldsymbol{\omega}$. In similar notation to that used in the previous proof, for $\boldsymbol{x} \in \mathbb{R}^{2^{s+1}}$, we let $\boldsymbol{\chi}(\boldsymbol{x}) := \prod_{i=1}^{2^{s+1}} \chi(x_i)$. Note that Ψ is injective, $\Psi(\mathbb{Z}^{s+2}) = \mathbb{Z}^{2^{s+1}} \cap \operatorname{im}\Psi$, and $\|\Psi\|_\infty = O_s(1)$.

The contribution to the inner integral of (6-9) from a particular $\boldsymbol{n}$ is zero unless $\|\boldsymbol{n} - \Psi(x, \boldsymbol{h})\|_\infty \ll \eta$, for some $(x, \boldsymbol{h}) \in \mathbb{R}^{s+2}$. Therefore, if η is small enough we can conclude that $\boldsymbol{n}$ must be of the form $\Psi(p, \boldsymbol{k})$, for some unique $(p, \boldsymbol{k}) \in \mathbb{Z}^{s+2}$. (To spell it out, apply Lemma D.2 with the map Ψ in place of the map Ξ). So (6-9) is equal to

$$\frac{1}{N^{s+2}} \sum_{(p,\boldsymbol{k})\in\mathbb{Z}^{s+2}} \Bigg(\prod_{\boldsymbol{\omega}\in\{0,1\}^{s+1}} f(p + \boldsymbol{k}\cdot\boldsymbol{\omega})\Bigg) \int_{(x,\boldsymbol{h})\in\mathbb{R}^{s+2}} \boldsymbol{\chi}(\Psi(x - p, \boldsymbol{h} - \boldsymbol{k}))\, dx\, d\boldsymbol{h}, \tag{6-10}$$

which, after a change of variables, is equal to

$$\frac{C}{N^{s+2}} \sum_{(p,\boldsymbol{k})\in\mathbb{Z}^{s+2}} \prod_{\boldsymbol{\omega}\in\{0,1\}^{s+1}} f(p + \boldsymbol{k}\cdot\boldsymbol{\omega}), \tag{6-11}$$

where

$$C := \int_{(x,\boldsymbol{h})\in\mathbb{R}^{s+2}} \boldsymbol{\chi}(\Psi(x, \boldsymbol{h}))\, dx\, d\boldsymbol{h}.$$

Since $\boldsymbol{\chi}$ has support contained within $[-\eta, \eta]^{2^{s+1}}$, a vector $(x, \boldsymbol{h})$ only makes a nonzero contribution to the above integral if $\|\Psi(x, \boldsymbol{h})\|_\infty \ll \eta$. This implies that $\|(x, \boldsymbol{h})\|_\infty \ll \eta$. (To prove this is full, apply Lemma D.3 to the linear map Ψ). Since $\|\boldsymbol{\chi}\|_\infty = O(1)$, this means $C = O(\eta^{s+2})$. The lemma then follows from (6-11). □

7. Degeneracy relations

Our aim for this short section is to establish a quantitative relationship between the dual pair degeneracy variety $V^*_{\mathrm{degen},2}(m,d,h)$ and the degeneracy variety $V_{\mathrm{degen}}(h-m,d)$ (see Definitions 5.4 and 4.4 respectively), which will be needed in the next section. It is here that we show that $V^*_{\mathrm{degen},2}(m,d,h)$ was indeed the appropriate notion for guaranteeing finite Cauchy–Schwarz complexity of the relevant system of homogeneous linear forms. We direct the reader to Proposition 4.5 and the discussion after Definition 5.4 for more on this issue.

To introduce the ideas, we first prove a nonquantitative proposition (which is a generalisation of Proposition 4.5).

Lemma 7.1. *Let m,d,h be natural numbers, with $d \geqslant h \geqslant m+2$. Let $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ be an injective linear map, let $L : \mathbb{R}^h \to \mathbb{R}^m$ be a surjective linear map, and suppose that $(\Xi, L) \notin V^*_{\mathrm{degen},2}(m,d,h)$. Let $\Phi : \mathbb{R}^{h-m} \to \ker L$ be any surjective linear map. Then the linear map $\Xi\Phi : \mathbb{R}^{h-m} \to \mathbb{R}^d$, viewed as a system of homogeneous linear forms, is not in $V_{\mathrm{degen}}(h-m,d)$.*

Proof. Let $\boldsymbol{e_1},\dots,\boldsymbol{e_d}$ denote the standard basis vectors in $\mathbb{R}^d$, and let $\boldsymbol{e_1^*},\dots,\boldsymbol{e_d^*}$ denote the dual basis of $(\mathbb{R}^d)^*$. Suppose for contradiction that $\Xi\Phi \in V_{\mathrm{degen}}(h-m,d)$. Then by Proposition 4.5 there exist two indices $i,j \leqslant d$, and a real number λ, such that $\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}$ is nonzero and $\Xi\Phi(\mathbb{R}^{h-m}) \subset \ker(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*})$.

But then $\Phi(\mathbb{R}^{h-m}) \subset \ker(\Xi^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}))$, i.e., $\Xi^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}) \in (\ker L)^0$. But $(\ker L)^0 = L^*((\mathbb{R}^m)^*)$, and so $\Xi^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}) \in L^*((\mathbb{R}^m)^*)$.

Then, by the definition of $V^*_{\mathrm{degen},2}(m,d,h)$, we have $(\Xi, L) \in V^*_{\mathrm{degen},2}(m,d,h)$, which is a contradiction. □

The ideas having been introduced, we state the quantitative version we require.

Lemma 7.2. *Let m,d,h be natural numbers, with $d \geqslant h \geqslant m+2$, and let c, C be positive constants. Let $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ be a linear map, and let $L : \mathbb{R}^h \to \mathbb{R}^m$ be a surjective linear map. Suppose that $\|\Xi\|_\infty \leqslant C$, and* $\mathrm{dist}((\Xi, L), V^*_{\mathrm{degen},2}(m,d,h)) \geqslant c$. *Let K denote* $\ker L$, *choose any orthonormal basis $\{\boldsymbol{v^{(1)}},\dots,\boldsymbol{v^{(h-m)}}\}$ for K, and let $\Phi : \mathbb{R}^{h-m} \to K$ denote the associated parametrisation, i.e., $\Phi(\boldsymbol{x}) := \sum_{i=1}^{h-m} x_i \boldsymbol{v^{(i)}}$. Then $\|\Xi\Phi\|_\infty = O(C)$ and* $\mathrm{dist}(\Xi\Phi, V_{\mathrm{degen}}(h-m,d)) = \Omega(c)$.

For the definition of $\mathrm{dist}((\Xi, L), V^*_{\mathrm{degen},2}(m,d,h))$, consult Definition 5.5.

Proof. Certainly $\|\Phi\|_\infty = O(1)$, as the chosen basis $\{\boldsymbol{v^{(1)}},\dots,\boldsymbol{v^{(h-m)}}\}$ is orthonormal. Therefore $\|\Xi\Phi\|_\infty = O(C)$.

Let $\boldsymbol{e_1},\dots,\boldsymbol{e_d}$ denote the standard basis vectors in $\mathbb{R}^d$, and let $\boldsymbol{e_1^*},\dots,\boldsymbol{e_d^*}$ denote the dual basis of $(\mathbb{R}^d)^*$. Suppose for contradiction that $\mathrm{dist}(\Xi\Phi, V_{\mathrm{degen}}(h-m,d)) \leqslant \eta$ for some small parameter η. In other words, assume that there exists a linear map $P : \mathbb{R}^{h-m} \to \mathbb{R}^d$ with $\|P\|_\infty \leqslant \eta$ such that $\Xi\Phi + P \in V_{\mathrm{degen}}(h-m,d)$. By definition, this means that

$$(\Xi\Phi + P)(\mathbb{R}^{h-m}) \subset \ker(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}),$$

for some two indices $i,j \leqslant d$, and some real number λ, such that $\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}$ is nonzero.

We can factorise $P = Q\Phi$, for some linear map $Q : \mathbb{R}^h \to \mathbb{R}^d$ with $\|Q\|_\infty \ll \eta$. Indeed, let $\boldsymbol{f_1}, \dots, \boldsymbol{f_{h-m}}$ denote the standard basis vectors in $\mathbb{R}^{h-m}$, and for all k at most $h-m$ define

$$Q(\boldsymbol{v^{(k)}}) := P(\boldsymbol{f_k}).$$

(If the notation for the indices seems odd here, it is designed to match the notation in Proposition 8.2, in which having superscript on the vectors $\boldsymbol{v^{(k)}}$ seems to be natural). Complete $\{\boldsymbol{v^{(1)}}, \dots, \boldsymbol{v^{(h-m)}}\}$ to an orthonormal basis $\{\boldsymbol{v^{(1)}}, \dots, \boldsymbol{v^{(h)}}\}$ for $\mathbb{R}^h$ and, for k in the range $h-m+1 \leqslant k \leqslant h-m$, define $Q(\boldsymbol{v^{(k)}}) := \boldsymbol{0}$. Then $P = Q\Phi$, and $\|Q\|_\infty = O(\eta)$, since $\{\boldsymbol{v^{(1)}}, \dots, \boldsymbol{v^{(h)}}\}$ is an orthonormal basis.

Thus,

$$(\Xi\Phi + Q\Phi)(\mathbb{R}^{h-m}) \subset \ker(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}).$$

So

$$\Phi(\mathbb{R}^{h-m}) \subset \ker((\Xi + Q)^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*})).$$

Like the previous proof, we conclude that

$$(\Xi + Q)^*(\boldsymbol{e_i^*} - \lambda \boldsymbol{e_j^*}) \in L^*((\mathbb{R}^m)^*).$$

Hence $((\Xi + Q), L) \in V^*_{\text{degen},2}(m, d, h)$, which, if η is small enough, contradicts the assumption that $\text{dist}((\Xi, L), V^*_{\text{degen},2}(m, d, h)) \geqslant c$. □

8. A generalised von Neumann theorem

In this section we complete the proof of Theorem 5.6, and therefore complete the proof of Theorem 2.12. It will be enough to prove the following statement.

Theorem 8.1. *Let N, m, d, h be natural numbers, with $d \geqslant h \geqslant m+2$, and let c, C, ε be positive reals. Let $\Xi = \Xi(N) : \mathbb{R}^h \to \mathbb{R}^d$ be an injective linear map with integer coefficients, and let $L = L(N) : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map. Suppose further that $\|L\|_\infty \leqslant C$, $\|\Xi\|_\infty \leqslant C$, $\text{dist}(L, V_{\text{rank}}(m, d)) \geqslant c$ and $\text{dist}((\Xi, L), V^*_{\text{degen},2}(m, d, h)) \geqslant c$. Then there is some natural number s at most $d-2$, independent of ε, such that the following holds. Let $\tilde{\boldsymbol{r}} \in \mathbb{Z}^d$ be some vector with $\|\tilde{\boldsymbol{r}}\|_\infty = O_{c,C,\varepsilon}(1)$, and let σ_F be a parameter in the range $0 < \sigma_F < \frac{1}{2}$. Let $F : \mathbb{R}^h \to [0,1]$ be a Lipschitz function supported on $[-N, N]^h$, with Lipschitz constant $O(1/\sigma_F N)$, and let $G : \mathbb{R}^m \to [0,1]$ be any function supported on $[-\varepsilon, \varepsilon]^m$. Let $g_1, \dots, g_d : [-2N, 2N]^d \to [-1, 1]$ be arbitrary measurable functions. Suppose*

$$\min_{j \leqslant d} \|g_j\|_{U^{s+1}(\mathbb{R}, 2N)} \leqslant \rho$$

for some ρ at most 1. Then

$$|\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)| \ll_{c,C,\varepsilon} \rho^{\Omega(1)} \sigma_F^{-1}. \tag{8-1}$$

Proof that Theorem 8.1 implies Theorem 5.6. Assume the hypotheses of Theorem 5.6. This gives natural numbers N, m, d, h, linear maps $L : \mathbb{R}^h \to \mathbb{R}^m$ and $\Xi : \mathbb{R}^h \to \mathbb{R}^d$, and functions $F : \mathbb{R}^h \to [0,1]$ and

$G : \mathbb{R}^m \to [0,1]$. Let $f_1, \dots, f_d : [N] \to [-1,1]$ be arbitrary functions, and for ease of notation let

$$\delta := T^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(f_1, \dots, f_d).$$

From Lemma 3.2 and the triangle inequality, we have the crude bound $\delta = O_{c,C,\varepsilon}(1)$.

Let $\eta := c_1\delta\sigma_G$, where c_1 is small enough depending on m, d, h, c, C, and ε, and let $\chi : \mathbb{R} \to [0,1]$ be an η-supported measurable function (see Definition 6.2). For all j at most d, let $g_j := f_j * \chi$. Finally, suppose $\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho$, for some parameter ρ in the range $0 < \rho \leqslant 1$.

We proceed by bounding $\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)$. Indeed, by Lemma 6.5, if c_1 is small enough

$$\min_j \|g_j\|_{U^{s+1}(\mathbb{R})} \ll \eta^{\frac{s+2}{2^{s+1}}} \min_j \|f_j\|_{U^{s+1}[N]} \ll_{c,C,\varepsilon} \rho.$$

Applying Theorem 8.1 to these functions $g_1, \dots, g_d$, the above implies

$$\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d) \ll_{c,C,\varepsilon} \rho^{\Omega(1)}\sigma_F^{-1}. \tag{8-2}$$

Now we use this to bound δ by Gowers norms. Indeed, by Lemma 6.4, we have

$$\delta \ll_{c,C,\varepsilon} \frac{1}{(c_1\delta\sigma_G)^h}\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d) + c_1\delta + c_1\delta\sigma_G\sigma_F^{-1}N^{-1}.$$

Picking c_1 small enough, we may move the $c_1\delta$ term to the left-hand side to get an $\Omega(\delta)$ term. The bound (8-2) then yields

$$\delta^{h+1} \ll_{c,C,\varepsilon} \rho^{\Omega(1)}\sigma_F^{-1}\sigma_G^{-h} + \sigma_F^{-1}N^{-1},$$

and so

$$\delta \ll_{c,C,\varepsilon} \rho^{\Omega(1)}(\sigma_F^{-O(1)} + \sigma_G^{-O(1)}) + \sigma_F^{-O(1)}N^{-\Omega(1)}.$$

This yields the desired conclusion of Theorem 5.6. □

So it remains to prove Theorem 8.1, for which the bulk of the work will be done in the following two propositions. In Proposition 8.2, we will reduce the integral in $\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)$ to an integral over the kernel of L. This kernel will be parametrised by a map Ψ, which will have finite c_1-Cauchy–Schwarz complexity for some suitable c_1. In Proposition 8.3 we will then work out the details of applying the Cauchy–Schwarz inequality to such a map, thereby producing Gowers norms.

Proposition 8.2 (separating out the kernel). *Let N, m, d, h be natural numbers, with $d \geqslant h \geqslant m+2$, and let c, C, ε be positive constants. Let σ_F be a parameter in the range $0 < \sigma_F < 1/2$. Let $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ be an injective linear map with integer coefficients, and let $L : \mathbb{R}^h \to \mathbb{R}^m$ be a surjective linear map. Assume further that $\|L\|_\infty \leqslant C$, $\|\Xi\|_\infty \leqslant C$, $\operatorname{dist}(L, V_{\mathrm{rank}}(m,h)) \geqslant c$ and $\operatorname{dist}((\Xi, L), V^*_{\mathrm{degen},2}(m,d,h)) \geqslant c$. Let $F : \mathbb{R}^h \to [0,1]$ be a Lipschitz function supported on $[-CN, CN]^h$, with Lipschitz constant $O_C(1/\sigma_F N)$, and let $G : \mathbb{R}^m \to [0,1]$ be a measurable function supported on $[-\varepsilon, \varepsilon]^m$. Let $\tilde{\boldsymbol{r}}$ be a fixed vector in $\mathbb{Z}^d$, satisfying $\|\tilde{\boldsymbol{r}}\|_\infty = O_C(1)$. Then there exists a system of linear forms $(\psi_1, \dots, \psi_d) = \Psi : \mathbb{R}^{h-m} \to \mathbb{R}^d$ satisfying $\|\Psi\|_\infty = O_C(1)$, and a Lipschitz function $F_1 : \mathbb{R}^{h-m} \to [0,1]$ supported on $[-O_{c,C,\varepsilon}(N), O_{c,C,\varepsilon}(N)]^{h-m}$*

with Lipschitz constant $O(1/\sigma_F N)$, *such that, if* $g_1, \dots, g_d : [-2N, 2N] \to [-1, 1]$ *are arbitrary functions,*

$$|\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)| \ll_{c,C,\varepsilon} \left| \frac{1}{N^{h-m}} \int_{\boldsymbol{x}} \prod_{j=1}^{d} g_j(\psi_j(\boldsymbol{x}) + a_j) F_1(\boldsymbol{x})\, d\boldsymbol{x} \right|, \tag{8-3}$$

where, for each j, a_j *is some real number that satisfies* $a_j = O_{c,C,\varepsilon}(1)$.

Furthermore, there exists a natural number s *at most* $d-2$ *such that the system* Ψ *has* $\Omega_{c,C}(1)$*-Cauchy–Schwarz complexity at most* s, *in the sense of Definition 4.6.*

Proof of Proposition 8.2. For ease of notation, let

$$\beta := \tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d).$$

Noting that $\ker L$ is a vector space of dimension $h-m$, define $\{\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(h-m)}\} \subset \mathbb{R}^h$ to be an orthonormal basis for $\ker L$. Then the map $\Phi : \mathbb{R}^{h-m} \to \mathbb{R}^h$, defined by

$$\Phi(\boldsymbol{x}) := \sum_{i=1}^{h-m} x_i \boldsymbol{v}^{(i)}, \tag{8-4}$$

is an injective map that parametrises $\ker L$. (This is reminiscent of Lemma 7.2).

Now, extend the orthonormal basis $\{\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(h-m)}\}$ for $\ker L$ to an orthonormal basis $\{\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(h)}\}$ for $\mathbb{R}^h$. By implementing a change of basis, we may rewrite β as

$$\frac{1}{N^{h-m}} \int_{\boldsymbol{x} \in \mathbb{R}^h} F\left(\sum_{i=1}^{h} x_i \boldsymbol{v}^{(i)}\right) G\left(L\left(\sum_{i=1}^{h} x_i \boldsymbol{v}^{(i)}\right)\right) \left(\prod_{j=1}^{d} g_j\left(\xi_j\left(\Phi(\boldsymbol{x}_1^{h-m}) + \sum_{i=h-m+1}^{h} x_i \boldsymbol{v}^{(i)}\right) + \tilde{\boldsymbol{r}}_j\right)\right) d\boldsymbol{x}, \tag{8-5}$$

using $\boldsymbol{x}_1^{h-m}$ to refer to the vector in $\mathbb{R}^{h-m}$ given by the first the first $h-m$ coordinates of $\boldsymbol{x}$.

We wish to remove the presence of the variables $x_{h-m+1}, \dots, x_h$. To set this up, note that, by the choice of the vectors $\boldsymbol{v}^{(i)}$,

$$G\left(L\left(\sum_{i=1}^{h} x_i \boldsymbol{v}^{(i)}\right)\right) = G\left(L\left(\sum_{i=h-m+1}^{h} x_i \boldsymbol{v}^{(i)}\right)\right).$$

The vector $\sum_{i=h-m+1}^{h} x_i \boldsymbol{v}^{(i)}$ is in $(\ker L)^{\perp}$. Hence, due to the limited support of G, there is a domain D, contained in $[-O_{\varepsilon,c,C}(1), O_{\varepsilon,c,C}(1)]^m$, such that $G\big(L\big(\sum_{i=h-m+1}^{h} x_i \boldsymbol{v}^{(i)}\big)\big)$ is equal to zero unless $(x_{h-m+1}, \dots, x_h)^T \in D$. (This is proved in full in Lemma D.1).

We can use this observation to bound the right-hand side of (8-5). Indeed, we have

$$\beta \ll \operatorname{vol} D \times \sup_{\boldsymbol{x}_{h-m+1}^{h} \in D} \frac{1}{N^{h-m}} \left| \int_{\boldsymbol{x}_1^{h-m} \in \mathbb{R}^{h-m}} F\left(\sum_{i=1}^{h} x_i \boldsymbol{v}^{(i)}\right) G\left(L\left(\sum_{i=h-m+1}^{h} x_i \boldsymbol{v}^{(i)}\right)\right) \right.$$
$$\left. \times \left(\prod_{j=1}^{d} g_j\left(\xi_j\left(\Phi(\boldsymbol{x}_1^{h-m}) + \sum_{i=h-m+1}^{h} x_i \boldsymbol{v}^{(i)}\right) + \tilde{\boldsymbol{r}}_j\right)\right) d\boldsymbol{x}_1^{h-m} \right|. \tag{8-6}$$

So there exists some fixed vector $(x_{h-m+1}, \dots, x_h)^T$ in D such that

$$\beta \ll_{c,C,\varepsilon} \frac{1}{N^{h-m}} \left| \int_{\boldsymbol{x_1^{h-m}} \in \mathbb{R}^{h-m}} F\Big(\sum_{i=1}^{h} x_i \boldsymbol{v^{(i)}}\Big) G\Big(L\Big(\sum_{i=h-m+1}^{h} x_i \boldsymbol{v^{(i)}}\Big)\Big) \times \Big(\prod_{j=1}^{d} g_j\Big(\xi_j\Big(\Phi(\boldsymbol{x_1^{h-m}}) + \sum_{i=h-m+1}^{h} x_i \boldsymbol{v^{(i)}}\Big) + \tilde{\boldsymbol{r}}_j\Big)\Big) d\boldsymbol{x_1^{h-m}} \right|. \quad (8\text{-}7)$$

Define the function $F_1 : \mathbb{R}^{h-m} \to [0, 1]$ by

$$F_1(\boldsymbol{x_1^{h-m}}) := F\Big(\Phi(\boldsymbol{x_1^{h-m}}) + \sum_{i=h-m+1}^{h} x_i \boldsymbol{v^{(i)}}\Big)$$

and for each j at most d, a shift

$$a_j := \xi_j\Big(\sum_{i=h-m+1}^{h} x_i \boldsymbol{v^{(i)}}\Big) + \tilde{\boldsymbol{r}}_j.$$

Then

$$\beta \ll_{c,C,\varepsilon} \left| \frac{1}{N^{h-m}} \int_{\boldsymbol{x} \in \mathbb{R}^{h-m}} F_1(\boldsymbol{x}) \prod_{j=1}^{d} g_j(\xi_j(\Phi(\boldsymbol{x})) + a_j)\, d\boldsymbol{x} \right|, \quad (8\text{-}8)$$

and F_1 and a_j satisfy the conclusions of the proposition.

Finally, since $\operatorname{dist}((\Xi, L), V^*_{\text{degen},2}(m, d, h)) \geqslant c$ and $\|\Xi\|_\infty, \|L\|_\infty \leqslant C$, Lemma 7.2 tells us that $\Xi\Phi : \mathbb{R}^{h-m} \to \mathbb{R}^d$ satisfies $\operatorname{dist}(\Xi\Phi, V_{\text{degen}}(h-m, d)) \gg_{c,C} 1$. (One may consult Definitions 4.4 and 5.4 for the definitions of $V_{\text{degen}}(h-m, d)$ and $V^*_{\text{degen},2}(m, d, h)$). Thus, by Lemma 4.7, there exists some s at most $d-2$ for which $\Xi\Phi$ has $\Omega_{c,C}(1)$-Cauchy–Schwarz complexity at most s.

Writing Ψ for $\Xi\Phi$, the proposition is proved. □

Proposition 8.3 (Cauchy–Schwarz argument). *Let s, d be natural numbers, with $d \geqslant 3$, and let C be a positive constant. Let σ_F be a parameter in the range $0 < \sigma_F < \frac{1}{2}$. Let $(\psi_1, \dots, \psi_d) = \Psi : \mathbb{R}^{s+1} \to \mathbb{R}^d$ be a linear map, and suppose that $\psi_1(\boldsymbol{e_k}) = 1$, for all the standard basis vectors $\boldsymbol{e_k} \in \mathbb{R}^{s+1}$. Suppose that, for all j in the range $2 \leqslant j \leqslant s+1$, there exists some k such that $\psi_j(\boldsymbol{e_k}) = 0$. Let $N \geqslant 1$ be real, and let $g_1, \dots, g_d : [-N, N] \to [-1, 1]$ be arbitrary measurable functions, and, for each j at most d, let a_j be some real number with $|a_j| \leqslant CN$. Let $F : \mathbb{R}^{s+1} \to [0, 1]$ be any Lipschitz function, supported on $[-CN, CN]^{s+1}$ with Lipschitz constant $O(1/\sigma_F N)$. Suppose that $\|g_1\|_{U^{s+1}(\mathbb{R},N)} \leqslant \rho$, for some parameter ρ in the range $0 < \rho \leqslant 1$. Then*

$$\left| \frac{1}{N^{s+1}} \int_{\boldsymbol{w} \in \mathbb{R}^{s+1}} \prod_{j=1}^{d} g_j(\psi_j(\boldsymbol{w}) + a_j) F(\boldsymbol{w})\, d\boldsymbol{w} \right| \ll_C \rho^{-\Omega(1)} \sigma_F^{-1}. \quad (8\text{-}9)$$

We stress again that implied constants may depend on the implicit dimensions (so the $\Omega(1)$ term in (8-9) may depend on s).

Proof. This theorem is very similar to the usual generalised von Neumann theorem (see [Tao 2012, Exercise 1.3.23]), and the proof is very similar too. A few extra technicalities arise from our dealing with the reals rather than with a finite group, but these are easily surmountable.

We begin with some simple reductions. First, we assume that C is large enough in terms of all other $O(1)$ parameters. For notational convenience, we will also allow C to vary from line to line. Next, since $\psi_1(\boldsymbol{w}) = w_1 + w_2 + \cdots + w_{s+1}$, by shifting w_1 we can assume that $a_1 = 0$ in (8-9). Due to the restricted support of F, we may restrict the integral over $\boldsymbol{w}$ to $[-CN, CN]^{s+1}$. By Lemma B.4, for any $Y > 2$ there is a function $\boldsymbol{c}_Y : \mathbb{R}^{s+1} \to \mathbb{C}$ satisfying $\|\boldsymbol{c}\|_\infty \ll 1$ such that we may replace $F(\boldsymbol{w})$ by

$$\int_{\substack{\boldsymbol{\theta}\in\mathbb{R}^{s+1}\\ \|\boldsymbol{\theta}\|_\infty \leqslant Y}} c_Y(\boldsymbol{\theta}) e\left(\frac{\boldsymbol{\theta}\cdot\boldsymbol{w}}{N}\right) d\boldsymbol{\theta} + O_C\left(\frac{\log Y}{\sigma_F Y}\right).$$

We will determine a particularly suitable Y later (which will depend on ρ).

This means that

$$\left|\frac{1}{N^{s+1}}\int_{\boldsymbol{w}\in\mathbb{R}^{s+1}}\prod_{j=1}^{d} g_j(\psi_j(\boldsymbol{w})+a_j)F(\boldsymbol{w})\,d\boldsymbol{w}\right| \ll \int_{\substack{\boldsymbol{\theta}\in\mathbb{R}^{s+1}\\ \|\boldsymbol{\theta}\|_\infty \leqslant Y}} \left|\frac{1}{N^{s+1}}\int^{*}_{\boldsymbol{w}\in\mathbb{R}^{s+1}} e\left(\frac{\boldsymbol{\theta}}{N}\cdot\boldsymbol{w}\right)\left(\prod_{j=1}^{d} g_j(\psi_j(\boldsymbol{w})+a_j)\right) d\boldsymbol{w}\right| d\boldsymbol{\theta} + O_C\left(\frac{\log Y}{\sigma_F Y}\right), \quad (8\text{-}10)$$

where $\int^*$ indicates the limits $\boldsymbol{w} \in [-CN, CN]^{s+1}$. Fix $\boldsymbol{\theta}$. The inner integral of (8-10) will be our primary focus.

Firstly, we wish to "absorb" the exponential phases $e\left(\frac{\boldsymbol{\theta}}{N}\cdot\boldsymbol{w}\right)$. To do this, we write $e\left(\frac{\boldsymbol{\theta}}{N}\cdot\boldsymbol{w}\right)$ as a product of functions $\prod_{k=1}^{s+1} b_k(\boldsymbol{w})$, where, for each k, the function $b_k : \mathbb{R}^{s+1} \to \mathbb{C}$ is bounded in absolute value by 1 and does not depend on the variable w_k. Since $s+1 \geqslant 2$, this is possible. Now write

$$\prod_{j=2}^{d} g_j(\psi_j(\boldsymbol{w})+a_j) = \prod_{k=1}^{s+1} b'_k(\boldsymbol{w}),$$

where each $b'_k : \mathbb{R}^{s+1} \to \mathbb{C}$ is bounded in absolute value by 1 and does not depend on the variable w_k. This is possible since ψ_1 is the only function ψ_j that includes all the variables $w_1, \ldots, w_{s+1}$.

Therefore we may rewrite the inner integral of (8-10) as

$$\frac{1}{N^{s+1}}\int^{*}_{\boldsymbol{w}\in\mathbb{R}^{s+1}} g_1(\psi_1(\boldsymbol{w}))\prod_{k=1}^{s+1} b'_k(\boldsymbol{w})b_k(\boldsymbol{w})\,d\boldsymbol{w}. \quad (8\text{-}11)$$

A brief aside: readers familiar with the arguments of [Green and Tao 2010a, Appendix C] (which motivate the present proof) may note that a different device is used in that paper to absorb the exponential phases. Those authors work in the setting of the finite group $\mathbb{Z}/N\mathbb{Z}$, and there the exponential phases can be absorbed simply by twisting the functions $g_j : \mathbb{Z}/N\mathbb{Z} \to [-1, 1]$ by a suitable linear phase function (witness the discussion surrounding expression (C.7) from [loc. cit.]). The key point there is that, if

the linear form $\boldsymbol{w} \mapsto \boldsymbol{\theta} \cdot \boldsymbol{w}$ fails to be in the set $\mathrm{span}(\psi_j : 1 \leqslant j \leqslant d)$, then a Fourier expansion of g_j demonstrates that a certain expression, analogous to the inner integral of (8-10), is equal to zero. This clean argument is not quite so easy to apply here, as the linear phases are not integrable over all of $\mathbb{R}$, which is why we choose a different approach.

Returning to (8-11), recall that $\psi_1(\boldsymbol{w}) = w_1 + w_2 + \cdots + w_{s+1}$. Therefore, applying the Cauchy–Schwarz inequality in each of the variables w_1 through w_{s+1} in turn, one establishes that the absolute value of expression (8-11) is at most

$$\ll_C \left(\frac{1}{N^{2s+2}} \int^*_{\boldsymbol{w} \in \mathbb{R}^{s+1}} \int^*_{z \in \mathbb{R}^{s+1}} \prod_{\boldsymbol{\alpha} \in \{0,1\}^{s+1}} g_1\Bigg(\sum_{\substack{k \leqslant s+1 \\ \boldsymbol{\alpha}_k = 0}} w_k + \sum_{\substack{k \leqslant s+1 \\ \boldsymbol{\alpha}_k = 1}} z_k \Bigg) d\boldsymbol{w}\, d\boldsymbol{z} \right)^{1/2^{s+1}}. \tag{8-12}$$

This expression may be immediately related to the real Gowers norm as given in Definition A.3, by the change of variables $m_k := z_k - w_k$, for all k at most $s+1$, and $u := w_1 + \cdots + w_{s+1}$. Performing this change of variables shows that(8-12) is

$$\ll \left(\frac{1}{N^{2s+2}} \int_{(u, \boldsymbol{m}, z_2^{s-1}) \in D} \prod_{\boldsymbol{\alpha} \in \{0,1\}^{s+1}} g_1(u + \boldsymbol{\alpha} \cdot \boldsymbol{m})\, du\, d\boldsymbol{m}\, dz_2^{s+1} \right)^{1/2^{s+1}}, \tag{8-13}$$

where D is convex domain contained within $[-CN, CN]^{2s+2}$. It remains to replace D by a Cartesian box.

By Lemma B.2 we may write

$$1_D = F_\sigma + O(G_\sigma),$$

for any σ in the range $0 < \sigma < \frac{1}{2}$, where $F_\sigma, G_\sigma : \mathbb{R}^{2s+2} \to [0, 1]$ are Lipschitz functions supported on $[-CN, CN]^{2s+2}$, with Lipschitz constant $O_C(1/\sigma N)$, such that

$$\int_{\boldsymbol{x}} G_\sigma(\boldsymbol{x})\, d\boldsymbol{x} = O_C(\sigma N^{2s+2}).$$

Then, since $\|g_1\|_\infty \leqslant 1$, we may bound (8-13) above by

$$\left(\frac{1}{N^{2s+2}} \int^*_{u, \boldsymbol{m}, z_2^{s-1}} F_\sigma(u, \boldsymbol{m}, z_2^{s-1}) \prod_{\boldsymbol{\alpha} \in \{0,1\}^{s+1}} g_1(u + \boldsymbol{\alpha} \cdot \boldsymbol{m}) du\, d\boldsymbol{m}\, dz_2^{s-1} + O_C(\sigma) \right)^{1/2^{s+1}}, \tag{8-14}$$

where $\int^*$ now refers to the domain of integration $[-CN, CN]^{2s+2}$.

By applying Lemma B.4 to F_σ, for any $X > 2$ the absolute value of expression (8-14) is

$$\ll_C \Bigg(\Bigg(\frac{1}{N^{2s+2}} \int_{\substack{\boldsymbol{\xi} \in \mathbb{R}^{2s+2} \\ \|\boldsymbol{\xi}\|_\infty \leqslant X}} \left| \int^*_{u, \boldsymbol{m}, z_2^{s-1}} e\left(\frac{\boldsymbol{\xi}}{N} \cdot (u, \boldsymbol{m}, z_2^{s-1}) \right) \prod_{\boldsymbol{\alpha} \in \{0,1\}^{s+1}} g_1(u + \boldsymbol{\alpha} \cdot \boldsymbol{m}) du\, d\boldsymbol{m}\, dz_2^{s-1} \right| d\boldsymbol{\xi} \Bigg)$$
$$+ O(\sigma) + O\left(\frac{\log X}{\sigma X} \right) \Bigg)^{1/2^{s+1}}. \tag{8-15}$$

Integrating over the variables $z_2, \dots, z_{s+1}$, and splitting the exponential phase amongst the different functions, expression (8-15) is

$$\ll_C \left(\left(\frac{1}{N^{s+2}} \int_{\substack{\boldsymbol{\xi} \in \mathbb{R}^{2s+2} \\ \|\boldsymbol{\xi}\|_\infty \leqslant X}} \left| \int_{(u,\boldsymbol{m}) \in [-CN,CN]^{s+2}} \prod_{\boldsymbol{\alpha} \in \{0,1\}^{s+1}} g_{\boldsymbol{\alpha}}(u + \boldsymbol{\alpha} \cdot \boldsymbol{m})\, du\, d\boldsymbol{m} \right| d\boldsymbol{\xi} \right) \right.$$
$$\left. + O_C(\sigma) + O_C\left(\frac{\log X}{\sigma X} \right) \right)^{1/2^{s+1}}, \quad (8\text{-}16)$$

where each function $g_{\boldsymbol{\alpha}}$ is of the form

$$g_{\boldsymbol{\alpha}}(u) := g_1(u) e(k_{\boldsymbol{\alpha}} u)$$

for some real $k_{\boldsymbol{\alpha}}$. Note that $\|g_{\boldsymbol{\alpha}}\|_{U^{s+1}(\mathbb{R},N)} = \|g_1\|_{U^{s+1}(\mathbb{R},N)}$.

Recall that g_1 is supported on $[-2N, 2N]$. Therefore, if $\prod_{\boldsymbol{\alpha} \in \{0,1\}^{s+1}} g_{\boldsymbol{\alpha}}(u + \boldsymbol{\alpha} \cdot \boldsymbol{m}) \neq 0$ then $(u, \boldsymbol{m}) \in [-O(N), O(N)]^{s+2}$. So, if C is large enough in terms of s, we may replace the restriction $(u, \boldsymbol{m}) \in [-CN, CN]^{s+2}$ in (8-16) with the condition $(u, \boldsymbol{m}) \in \mathbb{R}^{s+2}$, without changing the value of (8-16).

Then, by the Gowers–Cauchy–Schwarz inequality (Proposition A.4) and the triangle inequality, (8-16) is

$$\ll_C \left(X^{O(1)} \|g_1\|_{U^{s+1}(\mathbb{R})}^{2^{s+1}} + \sigma + \frac{\log X}{\sigma X} \right)^{1/2^{s+1}}$$
$$\ll_C \left(X^{O(1)} \rho^{2^{s+1}} + \sigma + \frac{\log X}{\sigma X} \right)^{1/2^{s+1}}. \quad (8\text{-}17)$$

Choosing $X = \rho^{-c_1}$, with c_1 suitably small in terms of s, and $\sigma = \rho^{c_1/2}$, expression (8-17) is $O_C(\rho^{\Omega(1)})$.

Putting this estimate into (8-10), we get a bound on (8-10) of

$$\ll_C Y^{O(1)} \rho^{\Omega(1)} + O\left(\frac{\log Y}{\sigma_F Y} \right). \quad (8\text{-}18)$$

Picking $Y = \rho^{-c_1}$, with c_1 suitably small in terms of s, we may ensure that (8-18) is $O_C(\rho^{\Omega(1)} \sigma_F^{-1})$, thus proving the proposition. □

With these propositions in hand, Theorem 8.1 follows quickly.

Proof of Theorem 8.1. Assuming all the hypotheses of Theorem 8.1, apply the result of Proposition 8.2 to $\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)$. Thus

$$|\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)| \ll_{c,C,\varepsilon} \left| \frac{1}{N^{h-m}} \int_{\boldsymbol{x} \in \mathbb{R}^{h-m}} F_1(\boldsymbol{x}) \prod_{j=1}^{d} g_j(\psi_j(\boldsymbol{x}) + a_j)\, d\boldsymbol{x} \right|, \quad (8\text{-}19)$$

where $\Psi : \mathbb{R}^{h-m} \to \mathbb{R}^d$ has $\Omega_{c,C}(1)$-Cauchy–Schwarz complexity at most s, for some s at most $d-2$, $F_1 : \mathbb{R}^{h-m} \to [0,1]$ is a Lipschitz function supported on $[-O_{c,C,\varepsilon}(N), O_{c,C,\varepsilon}(N)]^{h-m}$ with Lipschitz constant $O(1/\sigma_F N)$, and $a_j = O_{c,C,\varepsilon}(1)$. Furthermore $\|\Psi\|_\infty = O_C(1)$.

We apply Proposition 4.8 to Ψ. Therefore, for *any* real numbers $w_1, \dots, w_{s+1}$,

$$|\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)| \ll \left| \frac{1}{N^{h-m}} \int_{\boldsymbol{x} \in \mathbb{R}^{h-m}} F_1\Big(\boldsymbol{x} + \sum_{k=1}^{s+1} w_k \boldsymbol{f_k}\Big) \prod_{j=1}^{d} g_j(\psi_j'(\boldsymbol{x}, \boldsymbol{w}) + a_j)\, d\boldsymbol{x} \right|, \tag{8-20}$$

where

- for each j at most d, $\psi_j' : \mathbb{R}^{h-m} \times \mathbb{R}^{s+1} \to \mathbb{R}$ is a linear form;
- $\psi_1'(\boldsymbol{0}, \boldsymbol{w}) = w_1 + \cdots + w_{s+1}$;
- $\boldsymbol{f_1}, \dots, \boldsymbol{f_{s+1}} \in \mathbb{R}^{h-m}$ are some vectors that satisfy $\|\boldsymbol{f_k}\|_\infty = O_{c,C}(1)$ for each k at most $s+1$;
- the system of forms $(\psi_1', \dots, \psi_d')$ is in normal form with respect to ψ_1'.

We remark that the right-hand side of expression (8-20) is independent of $\boldsymbol{w}$, as it was obtained by applying the change of variables $\boldsymbol{x} \mapsto \boldsymbol{x} + \sum_{k=1}^{s+1} w_k \boldsymbol{f_k}$ to expression (8-19).

Now, let $P : \mathbb{R}^{s+1} \to [0, 1]$ be some Lipschitz function, supported on $[-N, N]^{s+1}$, with Lipschitz constant $O(1/N)$. Also suppose that $P(\boldsymbol{x}) \equiv 1$ if $\|\boldsymbol{x}\|_\infty \leqslant N/2$. Integrating over $\boldsymbol{w}$, we have that $|\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)|$ is

$$\begin{aligned} &\ll_{c,C,\varepsilon} \frac{1}{N^{h-m+s+1}} \int_{\boldsymbol{w} \in \mathbb{R}^{s+1}} P(\boldsymbol{w}) \left| \int_{\boldsymbol{x} \in \mathbb{R}^{h-m}} F_1\Big(\boldsymbol{x} + \sum_{k=1}^{s+1} w_k \boldsymbol{f_k}\Big) \prod_{j=1}^{d} g_j(\psi_j'(\boldsymbol{x}, \boldsymbol{w}) + a_j)\, d\boldsymbol{x} \right| d\boldsymbol{w} \\ &\ll_{c,C,\varepsilon} \left| \frac{1}{N^{h-m+s+1}} \int_{\substack{\boldsymbol{x} \in \mathbb{R}^{h-m} \\ \boldsymbol{w} \in \mathbb{R}^{s+1}}} H(\boldsymbol{x}, \boldsymbol{w}) \prod_{j=1}^{d} g_j(\psi_j'(\boldsymbol{x}, \boldsymbol{w}) + a_j)\, d\boldsymbol{x}\, d\boldsymbol{w} \right|, \end{aligned} \tag{8-21}$$

where the function $H : \mathbb{R}^{h-m+s+1} \to [0, 1]$ is defined by

$$H(\boldsymbol{x}, \boldsymbol{w}) := F_1\Big(\boldsymbol{x} + \sum_{k=1}^{s+1} w_k \boldsymbol{f_k}\Big) P(\boldsymbol{w}).$$

Since the vectors $\boldsymbol{f_k}$ satisfy

$$\|\boldsymbol{f_k}\|_\infty = O_{c,C}(1),$$

H is a Lipschitz function supported on $[-O_{c,C,\varepsilon}(N), O_{c,C,\varepsilon}(N)]^{h-m+s+1}$, with Lipschitz constant $O_{c,C}(1/\sigma_F N)$. Notice in (8-21) that we were able to move the absolute value signs outside the integral, as P is positive and the integral over $\boldsymbol{x}$ is independent of $\boldsymbol{w}$ (so in particular has constant sign).

Fix $\boldsymbol{x}$. Then the integral over $\boldsymbol{w}$ in (8-21) satisfies the hypotheses of Proposition 8.3. Applying Proposition 8.3 to this integral, and then integrating over $\boldsymbol{x}$, one derives

$$|\tilde{T}^{L,\Xi,\tilde{\boldsymbol{r}}}_{F,G,N}(g_1, \dots, g_d)| \ll_{c,C,\varepsilon} \rho^{\Omega(1)} \sigma_F^{-1}.$$

Theorem 8.1 is proved. □

By our long series of reductions, this means that both Theorem 5.6 and Theorem 2.12 are proved. □

9. Constructions

In this section we prove Theorem 2.14, which, we remind the reader, is the partial converse of Theorem 2.12. In other words, we show that L being bounded away from $V^*_{\text{degen}}(m,d)$ is a necessary hypothesis for Theorem 2.12 to be true.

Proof of Theorem 2.14. Recall the hypotheses of Theorem 2.14. In particular, we suppose that

$$\liminf_{N\to\infty} \operatorname{dist}(L, V^*_{\text{degen}}(m,d)) = 0,$$

i.e., we assume that $\operatorname{dist}(L, V^*_{\text{degen}}(m,d)) = \omega(N)^{-1}$, for some function $\omega(N)$ such that

$$\limsup_{N\to\infty} \omega(N) = \infty.$$

Let η be a small positive quantity, picked small enough in terms of c and C, and let N be a natural number that is large enough so that $\omega(N) \geqslant \eta^{-1}$ and $\eta N \geqslant \max(1,\varepsilon)$. All implied constants to follow will be independent of η.

Since F is the indicator function of $[1,N]^d$ and G is the indicator function of $[-\varepsilon,\varepsilon]^m$, one has

$$T^L_{F,G,N}(f_1,\dots,f_d) = \frac{1}{N^{d-m}} \sum_{\substack{\boldsymbol{n}\in[N]^d \\ \|L\boldsymbol{n}\|_\infty \leqslant \varepsilon}} \prod_{j=1}^{d} f_j(n_j).$$

Our aim is to construct functions $f_1,\dots,f_d : [N] \to [-1,1]$ such that

$$\min_j \|f_j\|_{U^{s+1}[N]} \leqslant \rho$$

for some ρ at most 1 and that

$$T^L_{F,G,N}(f_1,\dots,f_d) > H(\rho) + E_\rho(N). \tag{9-1}$$

We begin by observing that the condition $\|L\boldsymbol{n}\|_\infty \leqslant \varepsilon$ implies certain constraints on two of the variables n_i. Indeed, let $L' \in V^*_{\text{degen}}(m,d)$ be such that $\|L-L'\|_\infty = \operatorname{dist}(L, V^*_{\text{degen}}(m,d))$. Write λ_{ij} for the coefficients of L and λ'_{ij} for the coefficients of L'. By reordering columns, without loss of generality we may assume that there exist real numbers $\{a_i\}_{i=1}^m$ not all 0 such that for all j in the range $3 \leqslant j \leqslant d$ we have

$$\sum_{i=1}^m a_i \lambda'_{ij} = 0, \tag{9-2}$$

and further we may assume that for all i we have $\lambda'_{i1} = \lambda_{i1}$ and $\lambda'_{i2} = \lambda_{i2}$ (else $L' \in V^*_{\text{degen}}(m,d)$ is not one of the closest matrices to L). By reordering rows and rescaling, we may assume that a_1 has maximal absolute value amongst all the a_i, and that $|a_1| = 1$.

Define

$$b_1 := \sum_{i=1}^m a_i \lambda_{i1}, \quad b_2 := \sum_{i=1}^m a_i \lambda_{i2},$$

and let $\boldsymbol{n} \in [N]^d$ be some solution to $\|L\boldsymbol{n}\|_\infty \leqslant \varepsilon$. The critical observation is that (9-2), combined with the assumptions on the a_i, implies that

$$|b_1 n_1 + b_2 n_2| \ll \eta N. \tag{9-3}$$

Indeed, for j in the range $3 \leqslant j \leqslant d$ we have

$$\left|\sum_{i=1}^m a_i \lambda_{ij}\right| = \left|\sum_{i=1}^m a_i(\lambda_{ij} - \lambda'_{ij})\right| \ll \eta.$$

Since $\|L\boldsymbol{n}\|_\infty \leqslant \varepsilon$, we certainly have that

$$\left|b_1 n_1 + b_2 n_2 + \sum_{j=3}^d n_j \sum_{i=1}^m a_i \lambda_{ij}\right| \ll \varepsilon,$$

and then (9-3) follows by the triangle inequality and the fact that $\eta N \geqslant \varepsilon$.

The constraint (9-3) will turn out to be enough for the proof. We consider various cases, constructing different counterexample functions f_1 and f_2 based on the size and sign of b_1 and b_2. To facilitate this, we let c_1 be a suitably small positive constant, depending on c and C, but independent of η. All constants C_1 and C_2 to follow will be assumed to satisfy $C_1, C_2 = O_{c,C}(1)$.

Case 1 $|b_1|, |b_2| \leqslant c_1$.

Under the assumptions of Theorem 2.14, this case is actually precluded. Indeed, consider the matrix L'', defined by taking

$$\lambda''_{ij} = \lambda'_{ij}$$

for all pairs $(i.j) \in [m] \times [d]$, except for $(1, 1)$ and $(1, 2)$. In these cases we let

$$\lambda''_{11} = \lambda'_{11} - \frac{b_1}{a_1} \quad \text{and} \quad \lambda''_{12} = \lambda'_{12} - \frac{b_2}{a_1}.$$

Then

$$\sum_{i=1}^m a_i \lambda''_{ij} = 0$$

for all j in the range $1 \leqslant j \leqslant d$. In other words we have shown that $\|L - L''\|_\infty \leqslant \eta + c_1$ for some matrix L'' with rank less than m. Since $\eta + c_1 < c$ (if c_1 is small enough), this implies that $\operatorname{dist}(L, V_{\mathrm{rank}}(m, d)) < c$, which contradicts the assumptions of Theorem 2.14. Therefore this case is indeed precluded.

Case 2 *b_1, b_2 both of the same sign, and $b_1, b_2 \geqslant c_1$.*

In this case, (9-3) implies that $n_1 \leqslant C_1 \eta N$ for some constant C_1.[14] Now, define $f_1 : [N] \to [-1, 1]$ to be the indicator function of the interval $[\lceil C_1 \eta N \rceil, N] \cap \mathbb{N}$. We then have

$$\|f_1 - 1\|_{U^{s+1}[N]} \ll \left(\frac{1}{N^{s+2}} \sum_{x, h_1, \dots, h_{s+1} \ll C_1 \eta N} 1\right)^{1/2^{s+1}} \leqslant C_2 (C_1 \eta)^{(s+2)/2^{s+1}}$$

[14] The same conclusion is true for n_2, but this will not be needed.

for some constant C_2. However, observe that

$$\begin{aligned}|T^L_{F,G,N}(f_1-1,1,\dots,1)| &= |T^L_{F,G,N}(f_1,1,\dots,1)-T^L_{F,G,N}(1,1,\dots,1)|\\ &= |0-T^L_{F,G,N}(1,1,\dots,1)|\\ &\gg_{c,C,\varepsilon} 1\end{aligned}$$

by the hypotheses of Theorem 2.14. If $T^L_{F,G,N}(f_1-1,1,\dots,1)$ did not satisfy (9-1), then

$$1 \ll_{c,C,\varepsilon} H(\rho)+E_\rho(1),$$

where $\rho := C_2(C_1\eta)^{(s+2)/2^{s+1}}$. Picking η small enough, then N large enough, this inequality cannot possibly hold, and we have a contradiction. So $T^L_{F,G,N}(f_1-1,1,\dots,1)$ satisfies (9-1).

Case 3 *b_1, b_2 of opposite signs, and $b_1, b_2 \geqslant c_1$.*

This is the most involved case, although the central idea is very simple. The condition (9-3) confines n_2 to lie within a certain distance of a fixed multiple of n_1. By constructing functions f_1 and f_2 using random choices of blocks of this length, but coupled in such a way that condition (9-3) is very likely to hold, we can guarantee that $T^L_{F,G,N}(f_1-p, f_2-p, 1,\dots,1)$ is bounded away from zero, where p is the probability used to choose the random blocks. However, despite the block construction and the coupling, the functions f_1 and f_2 still individually exhibit enough randomness to conclude that $\|f_1-p\|_{U^{s+1}[N]} = o(1)$ as $N\to\infty$, and the same for f_2.

We now fill in the technical details. Relation (9-3) implies that

$$|b_1n_1+b_2n_2| \leqslant C_1\eta N, \tag{9-4}$$

for some C_1 satisfying $C_1 = O(1)$, and without loss of generality assume that b_1 is positive, b_2 is negative, and $|b_1|$ is at least $|b_2|$. Let C_2 be some parameter, chosen so that $(C_1C_2\eta)^{-1}$ is an integer. Such a C_2 will of course depend on η, but in magnitude we may pick $C_2 \asymp 1$. We consider the real interval $[0,N]$ modulo N, and for $x\in[0,N]$ and i in the range $0\leqslant i\leqslant (C_1C_2\eta)^{-1}-1$ we define the half-open interval modulo N

$$I_{x,i} := [x+iC_1C_2\eta N, x+(i+1)C_1C_2\eta N).$$

This choice guarantees that

$$[0,N] = \bigcup_{i=0}^{(C_1C_2\eta)^{-1}-1} I_{x,i}, \tag{9-5}$$

and the union is disjoint. Now, for δ a small constant to be chosen later,[15] we define

$$I^\delta_{x,i} := \big[x+\big(i+\tfrac12-\delta\big)C_1C_2\eta N, x+\big(i+\tfrac12+\delta\big)C_1C_2\eta N\big).$$

We will use the partition (9-5) to construct a function f_1, using an averaging argument to choose an x so that the $I^\delta_{x,i}$ intervals capture a positive proportion of the solution density of the linear inequality

[15] This δ is unrelated to the notation $\delta = T^L_{F,G,N}(f_1,\dots,f_d)$ used in previous sections.

system. Indeed, for $n_1 \in [N]$ let the weight $u(n_1)$ denote the number of $d-1$-tuples $n_2, \dots, n_d \leqslant N$ that together with n_1 satisfy the inequality $\|L\boldsymbol{n}\|_\infty < \varepsilon$. The weight $u(n_1)$ could be zero, of course. Let

$$E_{x,\delta} := \bigcup_i I^\delta_{x,i}.$$

Then

$$\begin{aligned}\frac{1}{N}\int_0^N \sum_{n\in[N]} u(n)1_{E_{x,\delta}}(n)\,dx &= \frac{1}{N}\sum_{n\in[N]} u(n)\int_0^N 1_{E_{x,\delta}}(n)\,dx\\ &= \sum_{n\in[N]} u(n)2\delta\\ &= 2\delta N^{d-m}T^L_{F,G,N}(1,\dots,1)\end{aligned}$$

Therefore, by the assumptions of Theorem 2.14, we may fix an x such that

$$\sum_{n\in[N]} u(n)1_{E_{x,\delta}}(n) \gg_{c,C} \delta N^{d-m}T^L_{F,G,N}(1,\dots,1). \tag{9-6}$$

Let us finally define the function f_1. Let p be a small positive constant (to be decided later). Fix a value of x such that (9-6) holds. Then we define a random subset $A \subseteq [N]$ by picking all of $I_{x,i} \cap \mathbb{N}$ to be members of A, with probability p, or none of $I_{x,i} \cap \mathbb{N}$ to be members of A, with probability $1-p$. We then make this same choice for each i in the range $0 \leqslant i \leqslant (C_1C_2\eta)^{-1} - 1$, independently. Observe immediately that for each $n \in [N]$ the probability that $n \in A$ is always p (though these events are not always independent). We let $f_1(n)$ be the indicator function $1_A(n)$.

The function f_2 is defined in terms of f_1. Indeed, let

$$J_{x,i} = \frac{b_1}{|b_2|}I_{x,i} \cap (0, N],$$

where the dilation of the interval $I_{x,i}$ is not considered modulo N but rather just as an operator on subsets of $\mathbb{R}$ (see Section 1B for this notation). Since $b_1 \geqslant |b_2|$ we have that these $J_{x,i}$ also form a disjoint partition of $[0, N]$. (NB: If $b_1 > |b_2|$ it may be that certain $J_{x,i}$ are empty, since the dilate of the corresponding $I_{x,i}$ may land entirely outside $[0, N]$). Then let B be the subset of $[N]$ defined so that for each i with $J_{x,i}$ nonempty we have $J_{x,i} \cap \mathbb{N} \subseteq B$ if and only if $I_{x,i} \cap \mathbb{N} \subseteq A$. Note again that for each individual $n \in [N]$ the probability that $n \in B$ is always p. We let $f_2(n)$ be the indicator function $1_B(n)$.

Our first claim is that, if p is small enough in terms of δ,

$$|\mathbb{E}T^L_{F,G,N}(f_1, f_2, 1, \dots, 1) - T^L_{F,G,N}(p, p, 1, \dots, 1)| \gg_{c,C,\varepsilon} \delta^2. \tag{9-7}$$

Indeed, suppose that $I_{x,i}$ is included in the set A, and suppose that $n_1 \in I^\delta_{x,i}$. If $n_2 \in [N]$ satisfies $\left|\frac{b_1}{|b_2|}n_1 - n_2\right| \leqslant \frac{1}{b_2}C_1\eta N$ and if δ is small enough in terms of b_1 and b_2, then $n_2 \in J_{x,i}$.[16] Thus, by the

[16] This fact is the reason why we introduced the parameter δ.

observation (9-4), $n_2 \in B$, for every integer n_2 that is the second coordinate of a solution vector $\boldsymbol{n}$ for which the first coordinate is n_1.[17] Therefore

$$\begin{aligned}
\mathbb{E}T^L_{F,G,N}(f_1, f_2, 1, \dots, 1) &= \frac{1}{N^{d-m}} \sum_{\substack{\boldsymbol{n}\in[N]^d \\ \|L\boldsymbol{n}\|_\infty \leqslant \varepsilon}} \mathbb{P}(n_1 \in A \wedge n_2 \in B) \\
&\geqslant \frac{1}{N^{d-m}} \sum_{\substack{\boldsymbol{n}\in[N]^d \\ \|L\boldsymbol{n}\|_\infty \leqslant \varepsilon}} \mathbb{P}(n_1 \in A \wedge n_1 \in I^\delta_{x,i} \text{ for some } i \wedge n_2 \in B) \\
&\geqslant \frac{1}{N^{d-m}} \sum_{\substack{\boldsymbol{n}\in[N]^d \\ \|L\boldsymbol{n}\|_\infty \leqslant \varepsilon}} \mathbb{P}(n_1 \in A \wedge n_1 \in I^\delta_{x,i} \text{ for some } i) \\
&= \frac{1}{N^{d-m}} \sum_{n_1\in[N]} u(n_1) p 1_{E_{x,\delta}}(n_1) \\
&\geqslant 2\delta p T^L_{F,G,N}(1, \dots, 1),
\end{aligned}$$

where the final line follows from (9-6). On the other hand $T^L_{F,G,N}(p, p, 1, \dots, 1) = p^2 T^L_{F,G,N}(1, \dots, 1)$, and hence

$$\mathbb{E}T^L_{F,G,N}(f_1, f_2, 1, \dots, 1) - T^L_{F,G,N}(p, p, 1, \dots, 1) \geqslant (2\delta p - p^2) T^L_{F,G,N}(1, \dots, 1). \tag{9-8}$$

Picking p small enough in terms of δ, and using the assumption that $T^L_{F,G,N}(1, \dots, 1) = \Omega_{c,C,\varepsilon}(1)$, this proves the relation (9-7).

Our second claim is that

$$\mathbb{E}\|f_1 - p\|_{U^{s+1}[N]}, \mathbb{E}\|f_2 - p\|_{U^{s+1}[N]} \ll \eta^{1/2^{s+1}}. \tag{9-9}$$

We first consider f_1. Then

$$\mathbb{E}\|f_1 - p\|^{2^{s+1}}_{U^{s+1}[N]} \ll \frac{1}{N^{s+2}} \sum_{(x,\boldsymbol{h})\in\mathbb{Z}^{s+2}} \mathbb{E}\Bigg(\prod_{\boldsymbol{\omega}\in\{0,1\}^{s+1}} (f_1 - p1_{[N]})(x + \boldsymbol{h}\cdot\boldsymbol{\omega}) \Bigg).$$

Observe that for fixed $(x, \boldsymbol{h})$ the random variables $(f_1 - p1_{[N]})(x + \boldsymbol{h}\cdot\boldsymbol{\omega})$ each have mean zero and, unless some two of the expressions $x + \boldsymbol{h}\cdot\boldsymbol{\omega}$ lie in the same block I_i, these random variables are independent. Hence, apart from those exceptional cases, we may factor the expectation and conclude that

$$\mathbb{E}\Bigg(\prod_{\boldsymbol{\omega}\in\{0,1\}^{s+1}} (f_1 - p1_{[N]})(x + \boldsymbol{h}\cdot\boldsymbol{\omega}) \Bigg) = \prod_{\boldsymbol{\omega}\in\{0,1\}^{s+1}} \mathbb{E}((f_1 - p1_{[N]})(x + \boldsymbol{h}\cdot\boldsymbol{\omega})) = 0.$$

Therefore,

$$\mathbb{E}\|f_1 - p\|^{2^{s+1}}_{U^{s+1}[N]} \ll \frac{1}{N^{s+2}} \sum_{(x,\boldsymbol{h})\in[-N,N]^{s+2}} 1_R(\boldsymbol{h}) \ll \eta,$$

[17] i.e., a vector $\boldsymbol{n}$ such that $\|L\boldsymbol{n}\|_\infty \leqslant \varepsilon$.

where

$$R = \{\boldsymbol{h} : |\boldsymbol{h} \cdot (\boldsymbol{\omega_1} - \boldsymbol{\omega_2})| \leqslant C_1 C_2 \eta N \text{ for some } \boldsymbol{\omega_1}, \boldsymbol{\omega_2} \in \{0,1\}^{s+1},\ \boldsymbol{\omega_1} \neq \boldsymbol{\omega_2}\}.$$

Thus by Jensen's inequality we have

$$\mathbb{E}\|f_1 - p\|_{U^{s+1}[N]} \ll \eta^{1/2^{s+1}}, \tag{9-10}$$

as claimed in (9-9).

The calculation for f_2 is essentially identical, noting that the length of the blocks $J_{x,i}$ is also $O(\eta N)$.

It is possible that one could finish the argument here by considering a second moment, and choosing some explicit f_1 and f_2. To avoid calculating a second moment, we argue as follows. Suppose for contradiction that there were no functions $f_1, \dots, f_d$ that satisfied (9-1). Then, by (9-7), if we pick p to be small enough in terms of δ we have

$$\begin{aligned}\delta^2 &\ll_{c,C,\varepsilon} |\mathbb{E}T^L_{F,G,N}(f_1, f_2, 1, \dots, 1) - T^L_{F,G,N}(p, p, 1, \dots, 1)| \\ &\ll |\mathbb{E}T^L_{F,G,N}(f_1 - p, f_2, 1, \dots, 1)| + |\mathbb{E}T^L_{F,G,N}(p, f_2 - p, 1, \dots, 1)| \\ &\ll \mathbb{E}(H(\rho_1) + E_{\rho_1}(N)) + \mathbb{E}(H(\rho_2) + E_{\rho_2}(N)),\end{aligned} \tag{9-11}$$

where ρ_1 (resp. ρ_2) is any chosen upper-bound on $\|f_1 - p\|_{U^{s+1}[N]}$ (resp. $\|f_2 - p\|_{U^{s+1}[N]}$). Note that the values ρ_i may be random variables themselves.

We claim that the random variables ρ_1 and ρ_2 may be chosen so that the right-hand side of (9-11) is $\kappa(\eta) + o_\eta(1)$ as $N \to \infty$. To prove this, we make two observations. Note first that by Markov's inequality

$$\mathbb{P}(\|f_1 - p\|_{U^{s+1}[N]} \geqslant \eta^{1/2^{s+2}}) \ll \eta^{1/2^{s+2}}$$

We choose the (random) upper-bound ρ_1 satisfying

$$\rho_1 = \begin{cases} 1 & \text{if } \|f_1 - p\|_{U^{s+1}[N]} \geqslant \eta^{1/2^{s+2}}, \\ \eta^{1/2^{s+2}} & \text{otherwise .} \end{cases}$$

Secondly, we may upper-bound H by a concave envelope, so without loss of generality we may assume that H is concave.

Then by Jensen's inequality,

$$\mathbb{E}(H(\rho_1) + E_{\rho_1}(N)) \ll H(\mathbb{E}\rho_1) + \mathbb{E}(E_{\rho_1}(1)) \ll \kappa(\eta^{1/2^{s+2}}) + o_\eta(1) \ll \kappa(\eta) + o_\eta(1). \tag{9-12}$$

We do the same manipulation for f_2. Combining (9-12) with (9-11) we conclude that

$$\delta^2 \ll_{c,C,\varepsilon} \kappa(\eta) + o_\eta(1). \tag{9-13}$$

The only condition on δ occurred in the proof of (9-7), in which we assumed that δ was small enough in terms of b_1 and b_2. Therefore there exists a suitable δ that satisfies $\delta = \Omega_{c,C}(1)$. Picking such a δ, and then picking η small enough and N large enough, (9-13) is a contradiction. So there must be some functions $f_1, \dots, f_d$ that satisfy (9-1).

Case 4 *Exactly one of* b_1, b_2 *satisfies* $b_i \geqslant c_1$.

Without loss of generality we may assume that $b_1 \geqslant c_1$. But then, as in Case 2, (9-3) implies that $n_1 \leqslant C_1 \eta N$ for some constant C_1. The same construction as in Case 2 then applies.

We have covered all cases, and thus have concluded the proof of Theorem 2.14. □

Appendix A: Gowers norms

There are several existing accounts of the basic theory of Gowers norms — for example in [Green 2007] and [Tao 2012] — and the reader looking for an introduction to the theory in its full generality should certainly consult these references, as well as Appendices B and C of [Green and Tao 2010a]. However, in the interests of making this paper as self-contained as possible, we use this section to pick out the central definitions and notions that are used in the main text.

Definition A.1. Let N be a natural number. For a function $f : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$, and a natural number d, we define the Gowers U^d norm $\|f\|_{U^d(N)}$ to be the unique nonnegative solution to

$$\|f\|_{U^d(N)}^{2^d} = \frac{1}{N^{d+1}} \sum_{x,h_1,\dots,h_d} \prod_{\boldsymbol{\omega}\in\{0,1\}^d} \mathcal{C}^{|\boldsymbol{\omega}|} f(x + \boldsymbol{h}\cdot\boldsymbol{\omega}), \tag{A-1}$$

where $|\boldsymbol{\omega}| = \sum_i \omega_i$, $\boldsymbol{h} = (h_1, \dots, h_d)$, $\mathcal{C}$ is the complex-conjugation operator, and the summation is over $x, h_1, \dots, h_d \in \mathbb{Z}/N\mathbb{Z}$.

For example,

$$\|f\|_{U^1(N)} = \left|\frac{1}{N}\sum_x f(x)\right|,$$

and

$$\|f\|_{U^2(N)} = \left(\frac{1}{N^3}\sum_{x,h_1,h_2} f(x)\overline{f(x+h_1)f(x+h_2)}f(x+h_1+h_2)\right)^{1/4}.$$

It is not immediately obvious that the right-hand side of (A-1) is always a nonnegative real, nor why the U^d norms are genuine norms if $d \geqslant 2$: proofs of both these facts may be found in [Tao and Vu 2006]. An immediate Cauchy–Schwarz argument, which may also be found in [loc. cit.], gives the so-called "nesting property" of Gowers norms, namely the fact that

$$\|f\|_{U^2(N)} \leqslant \|f\|_{U^3(N)} \leqslant \|f\|_{U^4(N)} \leqslant \cdots.$$

The functions in the main text do not have a cyclic group as a domain but rather the interval $[N]$, but the theory may easily be adapted to this case.

Definition A.2. Let N, N' be natural numbers, with $N' \geqslant N$. Identify $[N]$ with a subset of $\mathbb{Z}/N'\mathbb{Z}$ in the natural way, i.e., $[N] = \{1, \dots, N\} \subseteq \{1, \dots, N'\}$, which we then view as $\mathbb{Z}/N'\mathbb{Z}$. For a function $f : [N] \to \mathbb{C}$, and a natural number d, we define the Gowers norm $\|f\|_{U^d[N]}$ to be the unique nonnegative

real solution to the equation

$$\|f\|_{U^d[N]}^{2^d} = \frac{1}{|R|} \sum_{x,h_1,\dots,h_d} \prod_{\boldsymbol{\omega}\in\{0,1\}^d} \mathcal{C}^{|\boldsymbol{\omega}|} f1_{[N]}(x + \boldsymbol{h}\cdot\boldsymbol{\omega}), \tag{A-2}$$

where $f1_{[N]}$ is the extension by zero of f to $\mathbb{Z}/N'\mathbb{Z}$, the summation is over $x, h_1, \dots, h_d \in \mathbb{Z}/N'\mathbb{Z}$, and R is the set

$$R := \{x, h_1, \dots, h_d \in \mathbb{Z}/N'\mathbb{Z} : \text{ for every } \boldsymbol{\omega} \in \{0,1\}^d, x + \boldsymbol{h}\cdot\boldsymbol{\omega} \in [N]\}.$$

One can immediately see that this definition is equivalent to

$$\|f\|_{U^d[N]} = \|f1_{[N]}\|_{U^d(N')}/\|1_{[N]}\|_{U^d(N')},$$

and is also independent of the choice of N' as long as N'/N is large enough (in terms of d). Taking $N' = O(N)$ we have $\|1_{[N]}\|_{U^d(N')} \asymp 1$, and thus $\|f\|_{U^d[N]} \asymp \|f1_{[N]}\|_{U^d(N')}$. (See [Green and Tao 2010a, Lemma B.5] for more detail on this).

We observe that there is only a contribution to the summand in (A-2) when $x \in [N]$ and for every i we have $h_i \in \{-N, -N+1, \dots, N-1, N\}$ modulo N'. Further, it may be easily seen that $|R| \asymp N^{d+1}$. Therefore, choosing N'/N sufficiently large, we conclude that

$$\|f\|_{U^d[N]} \asymp \left(\frac{1}{N^{d+1}} \sum_{x,h_1,\dots,h_d\in\mathbb{Z}} \prod_{\boldsymbol{\omega}\in\{0,1\}^d} \mathcal{C}^{|\boldsymbol{\omega}|} f(x + \boldsymbol{h}\cdot\boldsymbol{\omega})\right)^{1/2^d}. \tag{A-3}$$

The relation (A-3) is implicitly assumed throughout the main text.

In order to succinctly state Theorem 8.1, we had to refer to a Gowers norm $U^d(\mathbb{R})$, which has been used in some recent work on linear patterns in subsets of Euclidean space (see [Cook et al. 2017, Lemma 4.2; Durcik et al. 2018, Proposition 3.3]). This Gowers norm is a less well-studied object, as the theory was originally developed over finite groups. Nevertheless it may be perfectly well defined, and even deep aspects of its inverse theory may be deduced from the corresponding theory of the discrete Gowers norm (see [Tao 2015]).

Definition A.3. Let $f : [0,1] \to \mathbb{C}$ be a bounded measurable function, and let d be a natural number. Then we define the Gowers norm $\|f\|_{U^d(\mathbb{R})}$ to be the unique nonnegative real satisfying

$$\|f\|_{U^d(\mathbb{R})}^{2^d} = \int_{(x,\boldsymbol{h})\in\mathbb{R}^{d+1}} \prod_{\boldsymbol{\omega}\in\{0,1\}^d} \mathcal{C}^{|\boldsymbol{\omega}|} f\left(x + \sum_{i=1}^{d} h_i\omega_i\right) dx\, dh_1 \cdots dh_d \tag{A-4}$$

where $|\boldsymbol{\omega}| = \sum_i \omega_i$, and $\mathcal{C}$ is the complex-conjugation operator.

Let N be a positive real, and let $g : [-N, N] \to \mathbb{C}$ be a measurable function. Define the function $f : [0,1] \to \mathbb{C}$ by $f(x) := g(2Nx - N)$, and then set

$$\|g\|_{U^d(\mathbb{R},N)} := \|f\|_{U^d(\mathbb{R})}.$$

Explicitly, a change of variables shows that

$$\|g\|_{U^d(\mathbb{R},N)}^{2^d} \asymp \frac{1}{N^{d+1}} \int_{(x,\boldsymbol{h})\in\mathbb{R}^{d+1}} \prod_{\boldsymbol{\omega}\in\{0,1\}^d} \mathcal{C}^{|\boldsymbol{\omega}|} g\Big(x + \sum_{i=1}^{d} h_i\omega_i\Big)\, dx\, dh_1 \cdots dh_d. \tag{A-5}$$

We require one further fact about Gowers norms.

Proposition A.4 (Gowers–Cauchy–Schwarz inequality). *Let d be a natural number, and, for each $\boldsymbol{\omega} \in \{0,1\}^d$, let $f_{\boldsymbol{\omega}} : [0,1] \to \mathbb{C}$ be a bounded measurable function. Define the Gowers inner-product*

$$\langle (f_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\{0,1\}^d} \rangle := \int_{(x,\boldsymbol{h})\in\mathbb{R}^{d+1}} \prod_{\boldsymbol{\omega}\in\{0,1\}^d} \mathcal{C}^{|\boldsymbol{\omega}|} f_{\boldsymbol{\omega}}\Big(x + \sum_{i=1}^{d} h_i\omega_i\Big)\, dx\, dh_1 \cdots dh_d.$$

Then

$$|\langle (f_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in\{0,1\}^d} \rangle| \leqslant \prod_{\boldsymbol{\omega}\in\{0,1\}^d} \|f_{\boldsymbol{\omega}}\|_{U^d(\mathbb{R})}.$$

Proof. See [Tao and Vu 2006, Chapter 11] for the proof in the finite group setting. The modification to the setting of the reals is trivial. □

Appendix B: Lipschitz functions

In the body of the paper we made extensive use of properties of Lipschitz functions.

Definition B.1 (Lipschitz functions). We say that a function $F : \mathbb{R}^m \to \mathbb{C}$ is Lipschitz, with Lipschitz constant at most M, if

$$M \geqslant \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^m \\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|F(\boldsymbol{x}) - F(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_\infty}.$$

We say that a function $G : \mathbb{R}^m/\mathbb{Z}^m \to \mathbb{C}$ is Lipschitz, with Lipschitz constant at most M, if

$$M \geqslant \sup_{\substack{\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^m/\mathbb{Z}^m \\ \boldsymbol{x}\neq\boldsymbol{y}}} \frac{|G(\boldsymbol{x}) - G(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^m/\mathbb{Z}^m}}.$$

We record the three properties of Lipschitz functions that we will require.

Lemma B.2. *Let N be a positive real, let m be a natural number, let K be a convex subset of $[-N,N]^m$, and let σ be some parameter in the range $0 < \sigma < \frac{1}{2}$. Then there exist Lipschitz functions $F_\sigma, G_\sigma : \mathbb{R}^m \to [0,1]$ supported on $[-2N,2N]^m$, both with Lipschitz constant at most $O\big(\frac{1}{\sigma N}\big)$, such that*

$$1_K = F_\sigma + O(G_\sigma)$$

and $\int_{\boldsymbol{x}} G_\sigma(\boldsymbol{x})\, d\boldsymbol{x} = O(\sigma N^m)$. Furthermore, $F_\sigma(\boldsymbol{x}) \geqslant 1_K(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^m$, and G is supported on $\{\boldsymbol{x} \in \mathbb{R}^m : \operatorname{dist}(\boldsymbol{x}, \partial(K)) \leqslant \sigma N\}$.

This is [Green and Tao 2010a, Corollary A.3]. It was be used in Lemmas 5.9 and 5.11 to replace sums with sharp cutoffs by sums with Lipschitz cutoffs.

Lemma B.3. *Let X be a positive real, with $X > 2$. Let $F : \mathbb{R}^m/\mathbb{Z}^m \to \mathbb{C}$ be a Lipschitz function such that $\|F\|_\infty \leqslant 1$ and the Lipschitz constant of F is at most M. Then*

$$F(\boldsymbol{x}) = \sum_{\substack{\boldsymbol{k}\in\mathbb{Z}^m \\ \|\boldsymbol{k}\|_\infty \leqslant X}} c_X(\boldsymbol{k}) e(\boldsymbol{k}\cdot\boldsymbol{x}) + O\left(M\frac{\log X}{X}\right) \tag{B-1}$$

for every $\boldsymbol{x} \in \mathbb{R}^m/\mathbb{Z}^m$, for some function $c_X(\boldsymbol{k})$ satisfying $\|c_X(\boldsymbol{k})\|_\infty \ll 1$. (The implied constant in the error term above may depend on the underlying dimensions, as always in this paper).

This is [Green and Tao 2008b, Lemma A.9], and was used in Lemma 3.4 as a way of bounding the number of solutions to a certain inequality.

Lemma B.4. *Let X, N, C be positive reals, with $X > 2$ and $N > 1$. Let $F : \mathbb{R}^m \to \mathbb{C}$ be a Lipschitz function, supported on $[-CN, CN]^m$, such that $\|F\|_\infty \leqslant 1$ and the Lipschitz constant of F is at most M. Then*

$$F(\boldsymbol{x}) = \int_{\substack{\boldsymbol{\xi}\in\mathbb{R}^m \\ \|\boldsymbol{\xi}\|_\infty \leqslant X}} c_X(\boldsymbol{\xi}) e\left(\frac{\boldsymbol{\xi}\cdot\boldsymbol{x}}{N}\right) d\boldsymbol{\xi} + O_C\left(MN\frac{\log X}{X}\right) \tag{B-2}$$

for every $\boldsymbol{x} \in \mathbb{R}^m$, for some function $c_X(\boldsymbol{\xi})$ satisfying $\|c_X(\boldsymbol{\xi})\|_\infty \ll_C 1$.

Lemma B.4 is very similar to Lemma B.3, and may be easily proved by adapting that standard harmonic analysis argument found in [Green and Tao 2008b, Lemma A.9] from $\mathbb{R}^m/\mathbb{Z}^m$ to $\mathbb{R}^m$. For completeness, we sketch the proof.

Sketch of proof. By rescaling the variable $\boldsymbol{x}$ by a factor of N, we reduce to the case where F is supported on $[-C, C]^m$ and has Lipschitz constant at most MN.

Let

$$K_X(\boldsymbol{x}) := \prod_{i=1}^m \frac{1}{X}\left(\frac{\sin(\pi X x_i)}{\pi x_i}\right)^2.$$

Then

$$\widehat{K_X}(\boldsymbol{\xi}) = \prod_{i=1}^m \max\left(1 - \frac{|\xi_i|}{X}, 0\right).$$

We have

$$(F * K_X)(\boldsymbol{x}) = \int_{\substack{\boldsymbol{\xi}\in\mathbb{R}^m \\ \|\boldsymbol{\xi}\|_\infty \leqslant X}} \widehat{F}(\boldsymbol{\xi})\widehat{K_X}(\boldsymbol{\xi}) e(\boldsymbol{\xi}\cdot\boldsymbol{x})\, d\boldsymbol{x},$$

and, since $|\widehat{F}(\boldsymbol{\xi})| \leqslant \|F\|_1 \ll_C 1$, letting $c_X(\boldsymbol{\xi}) := \widehat{F}(\boldsymbol{\xi})\widehat{K_X}(\boldsymbol{\xi})$ gives a main term of the desired form.

It remains to show that

$$\|F - F * K_X\|_\infty \ll_C MN\frac{\log X}{X}.$$

By writing

$$|F(\boldsymbol{x}) - (F * K_X)(\boldsymbol{x})| = \left|\int_{\boldsymbol{y}\in\mathbb{R}^m} (F(\boldsymbol{x}) - F(\boldsymbol{y}))K_X(\boldsymbol{x}-\boldsymbol{y})\, d\boldsymbol{y}\right|,$$

one sees that it suffices to show that

$$\int_{\|z\|_\infty \leqslant 2C} \|z\|_\infty K_X(z)\, dz \ll_C \frac{\log X}{X}.$$

But this bound follows immediately from a dyadic decomposition. □

We used Lemma B.4 extensively in the generalised von Neumann theorem argument in Section 8.

Appendix C: Rank matrix and normal form: proofs

In this appendix we prove the two quantitative statements from earlier in the paper, namely Propositions 3.1 and 4.8.

Proposition C.1. *Let n be a natural number, and let $S = \{f_1, \dots, f_k\}$ be a finite set of continuous functions $f_1, \dots, f_k : \mathbb{R}^n \to \mathbb{R}$. Let*

$$V(S) = \{\boldsymbol{x} \in \mathbb{R}^n : f_i(\boldsymbol{x}) = 0 \text{ for all } i \leqslant k\}.$$

Suppose that $\boldsymbol{x} \in \mathbb{R}^n$ is a point with $\|\boldsymbol{x}\|_\infty \leqslant C$ and with $\operatorname{dist}(\boldsymbol{x}, V(S)) \geqslant c$, for some absolute positive constants c and C. Then, there is some f_j such that $|f_j(\boldsymbol{x})| = \Omega_{c,C,S}(1)$.

Proof. This is nothing more than the fact that every continuous function on a compact set is bounded, applied to the continuous function $\min(1/|f_1|, \dots, 1/|f_k|)$ and the compact set $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_\infty \leqslant C,\ \operatorname{dist}(\boldsymbol{x}, V(S)) \geqslant c\}$. □

From Proposition C.1 it is easy to deduce the existence of rank matrices, namely Proposition 3.1.

Proof of Proposition 3.1. Let k be equal to $\binom{d}{m}$, and identify $\mathbb{R}^{md}$ with the space of m-by-d real matrices. Then let $f_1, \dots, f_k$ be the k polynomials on $\mathbb{R}^{md}$ that are given by the k determinants of m-by-m submatrices of L. One then sees that $V_{\mathrm{rank}}(m, d)$ is exactly the set of common zeros of the functions f_i. This is since row rank equals column rank, and linear independence of columns in a square matrix can be detected by the determinant.

Since we assume that $\|L\|_\infty \leqslant C$ and $\operatorname{dist}(L, V_{\mathrm{rank}}(m, d)) \geqslant c$ we can fruitfully apply Proposition C.1 to deduce that there is some j for which $|f_j(L)| = \Omega_{c,C}(1)$. The matrix M whose determinant corresponds to the polynomial f_j is exactly the claimed rank matrix.

This settles the first part of Proposition 3.1. The second part then follows immediate by the construction of M^{-1} as the adjugate matrix of M divided by $\det M$.

The third part, namely the statement about linear combinations of rows, follows quickly from the others. Indeed, without loss of generality, assume that the rank matrix M is realised by columns 1 through m. Then, the fact that the rows of L are linearly independent means that there are unique real numbers a_i such that $\sum_{i=1}^m a_i \lambda_{ij} = v_j$ for all j in the range $1 \leqslant j \leqslant d$. (Recall that $(\lambda_{ij})_{i \leqslant m, j \leqslant d}$ denotes the coefficients

of L). Restricting to j in the range $1 \leqslant j \leqslant m$, we observe that the a_i are forced to satisfy

$$\begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = (M^T)^{-1} \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix}.$$

Since $\|(M^{-1})^T\|_\infty = \|M^{-1}\|_\infty = O_{c,C}(1)$, we conclude that $a_i = O_{c,C,C_1}(1)$ for all i.

The final part of the proposition is to show that if $\operatorname{dist}(L, V^{\mathrm{unif}}_{\mathrm{rank}}(m,d)) \geqslant c$ then, for each j, there exists a rank matrix of L that does not include the j-th column. But this statement follows immediately from the above, after having deleted the j-th column. □

We now prove Proposition 4.8 on the existence of quantitative normal form parametrisations. We remind the reader that, in the proof, the implied constants may depend on the dimensions of the underlying spaces, namely m and n. For the definition of the variety $V_{\mathcal{P}_i}$, which consists of all systems of linear forms for which the partition $\mathcal{P}_i$ is not "suitable", the reader may consult Definition 4.4. The reader may also find the example that follows the proof to be informative.

Proof of Proposition 4.8. Fix i, and let $\mathcal{P}_i$ be a partition of $[m] \setminus \{i\}$ such that $\operatorname{dist}(\Psi, V_{\mathcal{P}_i}) \geqslant c_1$ (such a $\mathcal{P}_i$ exists by the definition of c_1-Cauchy–Schwarz complexity, i.e., by Definition 4.6). The partition $\mathcal{P}_i$ has $s_i + 1$ parts, for some s_i at most s.

It is clear from Definition 4.6 that we may, without loss of generality, further subdivide the partition and assume that the partition $\mathcal{P}_i$ has exactly $s+1$ parts. Call the parts $\mathcal{C}_1$ through $\mathcal{C}_{s+1}$.

Following Section 4 of [Green and Tao 2010a], for each $k \in [s+1]$ there exists a vector $\boldsymbol{f_k} \in \mathbb{R}^n$ that witnesses the fact that $\operatorname{dist}(\Psi, V_{\mathcal{P}_i}) > 0$, i.e., for which $\psi_i(\boldsymbol{f_k}) = 1$ but $\psi_j(\boldsymbol{f_k}) = 0$ for all $j \in \mathcal{C}_k$. Such a vector can be found using Gaussian elimination, say. Consider the extension

$$\Psi'(\boldsymbol{u}, w_1, \dots, w_{s+1}) := \Psi(\boldsymbol{u} + w_1 \boldsymbol{f_1} + \cdots + w_{s+1} \boldsymbol{f_{s+1}}).$$

Then, if $\Psi' = (\psi'_1, \dots, \psi'_m)$, the form $\psi'_i(\boldsymbol{u}, w_1, \dots, w_{s+1})$ is the only one that uses all of the w_k variables. Furthermore, $\psi'_i(\boldsymbol{0}, \boldsymbol{w}) = w_1 + \cdots + w_{s+1}$. Also, $n' = n+s+1$, which is at most $n+m-1$. So Proposition 4.8 is proved if for each k we can find such a vector $\boldsymbol{f_k}$ that additionally satisfies $\|\boldsymbol{f_k}\|_\infty = O_{c_1,C_1}(1)$.

Consider a fixed k, and let Γ be the set of possible implementations of Gaussian elimination on the set of forms $\psi_i \cup \{\psi_j : j \in \mathcal{C}_k\}$ to find a solution vector $\boldsymbol{f_k}$. If in the course of implementing these algorithms we are given a free choice for a coordinate of $\boldsymbol{f_k}$, we set it to be equal to zero. Note that $|\Gamma| = O(1)$.

Now, for each $\gamma \in \Gamma$, let the rational functions

$$\frac{p_{\gamma,1}(\Psi)}{q_{\gamma,1}(\Psi)}, \dots, \frac{p_{\gamma,n}(\Psi)}{q_{\gamma,n}(\Psi)}$$

be the n rational functions defining the claimed coefficients of $\boldsymbol{f_k}$. One may assume without loss of generality that, for all j, we have $p_{\gamma,j}, q_{\gamma,j} \in \mathbb{Z}[X_1, \dots, X_n]$ with coefficients of size $O(1)$. Now let

$$Q_\gamma := \prod_{j \leqslant n} q_{\gamma,j}.$$

We claim that $V(I) \subseteq V_{\mathcal{P}_i}$, where I is the ideal generated by the set of polynomials $\{Q_\gamma : \gamma \in \Gamma\}$ and $V(I)$ is the affine algebraic variety generated by I. Indeed, if $Q_\gamma(\Psi) = 0$ for all $\gamma \in \Gamma$ then there is no Gaussian elimination implementation that finds a solution $\boldsymbol{f_k}$, and this in turn implies that $\mathcal{P}_i$ is not suitable for Ψ and hence that $\Psi \in V_{\mathcal{P}_i}$.

Since $V(I) \subseteq V_{\mathcal{P}_i}$, the assumptions of Proposition 4.8 imply that $\operatorname{dist}(\Psi, V(I)) \geqslant c_1$. Applying Proposition C.1 to the polynomials $\{Q_\gamma : \gamma \in \Gamma\}$, we conclude that there is some $\gamma \in \Gamma$ such that $|Q_\gamma(\Psi)| = \Omega_{c_1,C_1}(1)$. In particular, we conclude that the solution vector $\boldsymbol{f_k}$ obtained by the implementation γ has coefficients that are $O_{c_1,C_1}(1)$. This concludes the proof of Proposition 4.8. □

Let us illustrate the above proof with an example which we hope will be instructive. Consider $n = 3$, $m = 2$, $i = 1$, and denote

$$\Psi = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}.$$

Then the partition $\mathcal{P}_i$ consists of the singleton $\{2\}$, and suppose one wished to construct a suitable $\boldsymbol{f_1}$ simply by applying Gaussian elimination. Implementing the algorithm a certain way we have

$$\boldsymbol{f_1} = \begin{pmatrix} a_{22}/(a_{11}a_{22} - a_{12}a_{21}) \\ -a_{21}/(a_{11}a_{22} - a_{12}a_{21}) \\ 0 \end{pmatrix}$$

as a solution, in the case where $a_{11}a_{22} - a_{12}a_{21}$ is nonzero. Of course if $a_{11}a_{23} - a_{13}a_{21}$ is nonzero too, we have another solution

$$\boldsymbol{f_1} = \begin{pmatrix} a_{23}/(a_{11}a_{23} - a_{13}a_{21}) \\ 0 \\ -a_{21}/(a_{11}a_{23} - a_{13}a_{21}) \end{pmatrix}.$$

So, if one applied Gaussian elimination idly, one might end up with either of these two solutions. Unfortunately it could be the case that $\operatorname{dist}(\Psi, V_{\mathcal{P}_i}) \geqslant c_1$ whilst one of these determinants, $a_{11}a_{22} - a_{12}a_{21}$ say, was nonzero yet $o(1)$ (as the unseen variable N, on which Ψ will ultimately depend, tends to infinity). In this instance, applying the first implementation of the algorithm would not give a desirable solution vector $\boldsymbol{f_1}$. For this reason we had to apply somewhat indirect arguments in order to find the appropriate vector $\boldsymbol{f_1}$.

It is worth including a brief discussion on why these quantitative subtleties do not arise in the setting of [Green and Tao 2010a]. Indeed, assume that Ψ has rational coefficients of naive height at most C_1 and that $\Psi \notin V_{\mathcal{P}_i}$. Since there are only $O_{C_1}(1)$ many possible choices of Ψ we immediately conclude that $\operatorname{dist}(\Psi, V_{\mathcal{P}_i}) \gg_{C_1} 1$, without needing to assume this as an extra hypothesis. Then *any* implementation of Gaussian elimination succeeds in finding a suitably bounded $\boldsymbol{f_k}$, since one is only ever working with rationals of bounded height.

Appendix D: Additional linear algebra

In this appendix, we collect together the assortment of standard linear algebra lemmas that we used at various points throughout the paper. We also give the linear algebra argument that we used to construct the matrix P during the proof of Lemma 5.10.

This first lemma demonstrates the intuitive fact, that if $L : \mathbb{R}^d \to \mathbb{R}^m$ is a linear map then $L : (\ker L)^\perp \to \mathbb{R}^m$ has bounded inverse.

Lemma D.1. *Let m, d be natural numbers, with $d \geqslant m+1$, and let c, C, l be positive constants. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map, and suppose $\|L\|_\infty \leqslant C$ and $\operatorname{dist}(L, V_{\mathrm{rank}}(m,d)) \geqslant c$. Let K denote $\ker L$. Let R be a convex set contained in $[-l, l]^m$. Then, if $\boldsymbol{v} \in K^\perp$, $L\boldsymbol{v} \in R$ only when $\boldsymbol{v} \in R'$, where R' is some convex region that satisfies $R' \subseteq [-O_{c,C}(l), O_{c,C}(l)]^d$.*

Proof. We choose to prove this statement using the concept of the "rank matrix" introduced earlier. Writing L as a m-by-d matrix with respect to the standard bases, let $\boldsymbol{\lambda_i} \in \mathbb{R}^d$ denote the column vector such that $\boldsymbol{\lambda_i}^T$ is the i-th row of L. Since $\operatorname{dist}(L, V_{\mathrm{rank}}(m,d)) \geqslant c$, the vectors $\boldsymbol{\lambda_i}$ are linearly independent. Moreover, we may extend the set $\{\boldsymbol{\lambda_i} : i \leqslant m\}$ by orthogonal vectors of unit length to form a basis $\{\boldsymbol{\lambda_i} : i \leqslant d\}$ for $\mathbb{R}^d$.

We claim that for all $k \in [d]$ we have

$$\sum_{i=1}^{d} a_{ki} \boldsymbol{\lambda_i} = \boldsymbol{e_k},$$

for some coefficients a_{ki} satisfying $|a_{ki}| = O_{c,C}(1)$, where $\boldsymbol{e_k} \in \mathbb{R}^d$ is the k-th standard basis vector. Indeed, fix k, and note that $\boldsymbol{e_k} = \boldsymbol{x_k} + \boldsymbol{y_k}$, where $\boldsymbol{x_k} \in \operatorname{span}(\boldsymbol{\lambda_i} : i \leqslant m)$ and $\boldsymbol{y_k} \in \operatorname{span}(\boldsymbol{\lambda_i} : m+1 \leqslant i \leqslant d)$. The vectors $\boldsymbol{x_k}$ and $\boldsymbol{y_k}$ are orthogonal by construction, so in particular $\|\boldsymbol{x_k}\|_2^2 + \|\boldsymbol{y_k}\|_2^2 = 1$, and hence $\|\boldsymbol{x_k}\|_\infty, \|\boldsymbol{y_k}\|_\infty \ll 1$. By the third part of Proposition 3.1 applied to $\boldsymbol{x_k}$ we get $|a_{ki}| = O_{c,C}(1)$ when $i \leqslant m$, and the orthonormality of $\{\boldsymbol{\lambda_i} : m+1 \leqslant i \leqslant d\}$ implies that $|a_{ki}| = O(1)$ when i is in the range $m+1 \leqslant i \leqslant d$.

Now notice that $\operatorname{span}(\boldsymbol{\lambda_i} : m+1 \leqslant i \leqslant d)$ is exactly equal to K. Let $\boldsymbol{v} \in K^\perp$, and suppose $L\boldsymbol{v} \in R$. Letting L' be the d-by-d matrix whose rows are $\boldsymbol{\lambda_i}^T$, we have that $L'\boldsymbol{v} = \boldsymbol{w}$ for some vector $\boldsymbol{w}$ satisfying $\|\boldsymbol{w}\|_\infty \ll l$. Premultiplying by the matrix $A = (a_{ki})$, we immediately get $\boldsymbol{v} = A\boldsymbol{w}$, and hence $\|\boldsymbol{v}\|_\infty = O_{c,C}(l)$. The region $R' := (L^{-1}R) \cap K^\perp$ is therefore bounded. R' is clearly convex, and so the lemma is proved. □

The second lemma concerns vectors, with integer coordinates, that lie near to a subspace.

Lemma D.2. *Let h, d be natural numbers, with $h \leqslant d$, and let C, η be positive reals. Let $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ be an injective linear map, with $\|\Xi\|_\infty \leqslant C$. Suppose further that $\Xi(\mathbb{Z}^h) = \mathbb{Z}^d \cap \Xi(\mathbb{R}^h)$. Let $\boldsymbol{n}, \tilde{\boldsymbol{r}} \in \mathbb{Z}^d$. Suppose that*

$$\operatorname{dist}(\boldsymbol{n}, \Xi(\mathbb{R}^h) + \tilde{\boldsymbol{r}}) \leqslant \eta. \tag{D-1}$$

Then, if η is small enough in terms of C, h and d, $\boldsymbol{n} = \Xi(\boldsymbol{m}) + \tilde{\boldsymbol{r}}$, for some unique $\boldsymbol{m} \in \mathbb{Z}^h$.

Proof. By replacing $\boldsymbol{n}$ with $\boldsymbol{n} - \tilde{\boldsymbol{r}}$, we can assume without loss of generality that $\tilde{\boldsymbol{r}} = \boldsymbol{0}$. It will also be enough to show that $\boldsymbol{n} \in \Xi(\mathbb{R}^h)$, as the injectivity of Ξ and the assumption that $\Xi(\mathbb{Z}^h) = \mathbb{Z}^d \cap \Xi(\mathbb{R}^h)$ immediately go on to imply the existence of a unique $\boldsymbol{m}$.

Suppose for contradiction then that $\boldsymbol{n} \notin \Xi(\mathbb{R}^h)$. In matrix form, Ξ is a d-by-h matrix with linearly independent columns, all of whose coefficients are integers with absolute value at most C. We can extend this matrix to a d-by-d matrix $\widetilde{\Xi}$, with linearly independent columns, all of whose coefficients are integers with absolute value at most C. Then $(\widetilde{\Xi})^{-1}$ is a d-by-d matrix with rational coefficients of naive height at most $C^{O(1)}$, and $(\widetilde{\Xi})^{-1}(\Xi(\mathbb{R}^h)) = \mathbb{R}^h \times \{0\}^{d-h}$.

Since $\boldsymbol{n} \notin \Xi(\mathbb{R}^h)$, we have $(\widetilde{\Xi})^{-1}(\boldsymbol{n}) \notin \mathbb{R}^h \times \{0\}^{d-h}$. But $(\widetilde{\Xi})^{-1}(\boldsymbol{n}) \in \frac{1}{K}\mathbb{Z}^d$, for some natural number K satisfying $K = O(C^{O(1)})$. Therefore

$$\operatorname{dist}((\widetilde{\Xi})^{-1}(\boldsymbol{n}), (\widetilde{\Xi})^{-1}(\Xi(\mathbb{R}^h))) \gg C^{-O(1)}.$$

Applying $\widetilde{\Xi}$, we conclude that

$$\operatorname{dist}(\boldsymbol{n}, \Xi(\mathbb{R}^h)) \gg C^{-O(1)},$$

which is a contradiction to (D-1) if η is small enough. □

The construction of the matrix $\widetilde{\Xi}$ in the above proof also has an even more basic consequence, namely that $\Xi^{-1} : \operatorname{im} \Xi \to \mathbb{R}^h$ is bounded.

Lemma D.3. *Let h, d be natural numbers, with $h \leqslant d$, and let C, η be positive reals. Suppose that $\Xi : \mathbb{R}^h \to \mathbb{R}^d$ is an injective linear map, with $\|\Xi\|_\infty \leqslant C$. Suppose further that $\Xi(\mathbb{Z}^h) \subseteq \mathbb{Z}^d \cap \Xi(\mathbb{R}^h)$. Then if $\|\Xi(\boldsymbol{y})\|_\infty \leqslant \eta$, we have $\|\boldsymbol{y}\|_\infty \ll C^{-O(1)}\eta$.*

Proof. Construct the matrix $\widetilde{\Xi}$ as in the previous proof. Then $\|(\widetilde{\Xi})^{-1}(\Xi(\boldsymbol{y}))\|_\infty \ll C^{O(1)}\eta$, by the bound on the size of the coefficients of $\widetilde{\Xi}$. But $(\widetilde{\Xi})^{-1}(\Xi(\boldsymbol{y})) \in \mathbb{R}^d$ is nothing more than the vector $\boldsymbol{y} \in \mathbb{R}^h$ extended by zeros. So $\|\boldsymbol{y}\|_\infty \ll C^{O(1)}\eta$ as claimed. □

Finally, we give the linear algebra argument used to construct the matrix P during the proof of Lemma 5.10.

Lemma D.4. *Let m, d be natural numbers, with $d \geqslant m + 1$. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map with rational dimension u, and let $\Theta : \mathbb{R}^m \to \mathbb{R}^u$ be a rational map for L. Suppose that $\|L\|_\infty \leqslant C$ and $\|\Theta\|_\infty \leqslant C$. Equating L with its matrix, suppose that the first m columns of L form the identity matrix. Let $\{\boldsymbol{a_1}, \dots, \boldsymbol{a_u}\}$ be a basis for the lattice $\Theta L(\mathbb{Z}^d)$ that satisfies $\|\boldsymbol{a_i}\|_\infty = O_C(1)$ for every i. Let $\boldsymbol{x_1}, \dots, \boldsymbol{x_u} \in \mathbb{Z}^d$ be vectors such that, for every i, $\Theta L(\boldsymbol{x_i}) = \boldsymbol{a_i}$ and $\|\boldsymbol{x_i}\|_\infty = O_C(1)$. Then*

$$\mathbb{R}^m = \operatorname{span}(L\boldsymbol{x_i} : i \leqslant u) \oplus \ker \Theta \tag{D-2}$$

and there is an invertible linear map $P : \mathbb{R}^m \to \mathbb{R}^m$ such that

$$P((\operatorname{span}(L\boldsymbol{x_i} : i \leqslant u))) = \mathbb{R}^u \times \{0\}^{m-u}, \quad P(\ker \Theta) = \{0\}^u \times \mathbb{R}^{m-u},$$

and both $\|P\|_\infty = O_C(1)$ and $\|P^{-1}\|_\infty = O_C(1)$.

Note that both $\{\boldsymbol{a_1}, \dots, \boldsymbol{a_u}\}$ and $\boldsymbol{x_1}, \dots, \boldsymbol{x_u} \in \mathbb{Z}^d$ exist by applying Lemma 5.7 to the map $S := \Theta L$.

Proof. The expression (D-2) is immediate from the definitions, so it remains to construct P. We may assume, since the first m columns of L form the identity matrix, that Θ has integer coefficients.

As $\|\Theta\|_\infty = O_C(1)$, we may pick a basis $\{\boldsymbol{y_1}, \dots, \boldsymbol{y_{m-u}}\}$ for $\ker\Theta$ in which $\boldsymbol{y_j} \in \mathbb{Z}^m$ and $\|\boldsymbol{y_j}\|_\infty = O_C(1)$ for all j. Let $\boldsymbol{b_1}, \dots, \boldsymbol{b_m}$ denote the standard basis of $\mathbb{R}^m$, and define P by letting

$$\begin{aligned} P(L\boldsymbol{x_i}) &:= \boldsymbol{b_i}, \qquad 1 \leqslant i \leqslant u, \\ P(\boldsymbol{y_j}) &:= \boldsymbol{b_{j+u}}, \quad 1 \leqslant j \leqslant m-u, \end{aligned} \tag{D-3}$$

and then extending linearly to all of $\mathbb{R}^m$. Clearly $P((\mathrm{span}(L\boldsymbol{x_i} : i \leqslant u))) = \mathbb{R}^u \times \{0\}^{m-u}$ and $P(\ker\Theta) = \{0\}^u \times \mathbb{R}^{m-u}$. It is also immediate that $\|P^{-1}\|_\infty = O_C(1)$, since $\|L\boldsymbol{x_i}\|_\infty = O_C(1)$ and $\|\boldsymbol{y_j}\|_\infty = O_C(1)$ for all i and j. It remains to bound $\|P\|_\infty$. If $L\boldsymbol{x_i}$ were all vectors with integer coordinates then this bound would be immediate as well, as then P^{-1} would have integer coordinates and hence $|\det P^{-1}| \geqslant 1$. As it is, we have to proceed more slowly.

To this end, for a standard basis vector $\boldsymbol{b_k}$ write

$$\boldsymbol{b_k} = \sum_{i=1}^{u} \lambda_i L\boldsymbol{x_i} + \sum_{j=1}^{d-u} \mu_j \boldsymbol{y_j}.$$

It will be enough to show that $|\lambda_i|, |\mu_j| = O_C(1)$ for all i and j. First note that, since the first m columns of L form the identity, $\boldsymbol{b_k} \in L(\mathbb{Z}^d)$. Also $\Theta(\boldsymbol{b_k}) = \sum_{i=1}^u \lambda_i \boldsymbol{a_i}$. So $\boldsymbol{a} := \sum_{i=1}^u \lambda_i \boldsymbol{a_i}$ is an element of $\Theta L(\mathbb{Z}^d)$ that satisfies $\|\boldsymbol{a}\|_\infty = O_C(1)$. Since $\|\boldsymbol{a_i}\|_\infty = O_C(1)$ for every i, and $\{\boldsymbol{a_1}, \dots, \boldsymbol{a_u}\}$ is a basis for the lattice $\Theta L(\mathbb{Z}^d)$, this implies that $|\lambda_i| = O_C(1)$ for every i.

So then $\sum_{j=1}^{d-u} \mu_j \boldsymbol{y_j}$ is a vector in $\ker\Theta$ satisfying $\left\|\sum_{j=1}^{d-u} \mu_j \boldsymbol{y_j}\right\|_\infty = O_C(1)$. Since $\{\boldsymbol{y_1}, \dots, \boldsymbol{y_{m-u}}\}$ is a set of linearly independent vectors, each of which has integer coordinates with absolute value $O_C(1)$, this implies that $|\mu_j| = O_C(1)$ for every j.

Therefore P satisfies the conclusions of the lemma. □

Remark D.5. We note the effects of the above construction in the case when L has algebraic coefficients. We use a rudimentary version of height: if $Q \in \mathbb{Z}[X]$ we define

$$H(Q) := \max(|q_i| : q_i \text{ a coefficient of } Q)$$

to be the *height* of Q, and we say that the height of an algebraic number is the height of its minimal polynomial. (So there are $O_{k,H}(1)$ algebraic numbers of degree at most k and height at most H.) Then, if in the statement of Lemma D.4 all the coefficients of L are algebraic numbers with degree at most k and height at most H, all the coefficients of P are algebraic numbers of degree $O_k(1)$ and height $O_{C,k,H}(1)$.

Appendix E: The approximation function in the algebraic case

We use this final appendix to give the proof of relation (2-3). The following lemma makes this relation quantitatively precise.

Lemma E.1. *Let m, d be natural numbers, with $d \geqslant m+1$, and let c, C be positive constants. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a surjective linear map, and suppose that the matrix of L has algebraic coefficients of algebraic degree at most k and algebraic height at most H (see Remark D.5 for definitions). Suppose that $\|L\|_\infty \leqslant C$, that $\operatorname{dist}(L, V_{\mathrm{rank}}(m,d)) \geqslant c$, and that L has rational complexity at most C. Let τ_1, τ_2 be two parameters in the range $0 < \tau_1, \tau_2 \leqslant 1$. Then*

$$A_L(\tau_1, \tau_2) \gg_{k,H,c,C} \min(\tau_1, \tau_2^{O_k(1)}).$$

Proof. We begin by reducing to the case when L is purely irrational. Indeed, consider Lemma 5.10 and replace L by the map L' (expression (5-10)). By part (9) of Lemma 5.10, $A_{L'}(\tau_1, \tau_2) \ll_{c,C} A_L(\Omega_{c,C}(\tau_1), \Omega_{c,C}(\tau_2))$. Also, using Remark D.5, it follows that L' has algebraic coefficients of algebraic degree at most $O_k(1)$ and algebraic height at most $O_{c,C,k,H}(1)$. So, replacing L with L', without loss of generality we may assume that L is purely irrational.

Suppose for contradiction that for all choices of constants c_1 and C_2, there exist parameters τ_1 and τ_2 such that $A_L(\tau_1, \tau_2) < c_1 \min(\tau_1, \tau_2^{C_2})$, i.e., there exists a map $\alpha \in (\mathbb{R}^m)^*$ and a map $\varphi \in (\mathbb{Z}^d)^T$ such that $\tau_1 \leqslant \|\alpha\|_\infty \leqslant \tau_2^{-1}$ and

$$\|L^*\alpha - \varphi\|_\infty < c_1 \min(\tau_1, \tau_2^{C_2}). \tag{E-1}$$

Fix α and φ so that they satisfy (E-1). We will obtain a contradiction if c_1 is small enough in terms of c, C, k, H, and if and C_2 is large enough in terms of k.

In the first part of the proof, we apply various reductions to enable us to replace α with a map that has integer coordinates with respect to the standard dual basis of $(\mathbb{R}^m)^*$.

Let M be a rank matrix of L (Proposition 3.1), and assume without loss of generality that M consists of the first m columns of L. Then there exists a map $\beta \in (\mathbb{R}^m)^*$, namely $\beta := M^*\alpha$, such that $\tau_1 \ll_{c,C} \|\beta\|_\infty \ll_{c,C} \tau_2^{-1}$ and

$$\|L^*(M^{-1})^*\beta - \varphi\|_\infty < c_1 \min(\tau_1, \tau_2^{C_2}). \tag{E-2}$$

Since the first m columns of $M^{-1}L$ form the identity matrix, (E-2) implies that

$$\operatorname{dist}(\beta, (\mathbb{Z}^m)^T) < c_1 \min(\tau_1, \tau_2^{C_2}). \tag{E-3}$$

We know that $\|\beta\|_\infty = \Omega_{c,C}(\tau_1)$. Also, considering (E-3), by perturbing β by a suitable element $\gamma \in (\mathbb{R}^m)^*$ with $\|\gamma\|_\infty < c_1 \min(\tau_1, \tau_2^{C_2})$ we may obtain a map $\rho \in (\mathbb{Z}^m)^T$. Combining these facts, note how

$$\|\rho\|_\infty \geqslant \|\beta\|_\infty - c_1 \min(\tau_1, \tau_2^{C_2}) \gg_{c,C} \tau_1$$

if c_1 is small enough, and so certainly $\rho \neq 0$.

From (E-2), we therefore conclude that there exists some $\rho \in (\mathbb{Z}^m)^T \setminus \{0\}$, satisfying $\|\rho\|_\infty = O_{c,C}(\tau_2^{-1})$, such that

$$\|L^*(M^{-1})^*\rho - \varphi\|_\infty < c_1 C_3 \tau_2^{C_2} \tag{E-4}$$

where C_3 is some constant that depends on c and C. Referring back to (E-1), we see that we have achieved our goal of replacing α with a map that has integer coefficients.

Expression (E-4) leads to a contradiction. Morally this follows from Liouville's theorem on the diophantine approximation of algebraic numbers, but we could not find exactly the statement we needed in the literature, so we include a short argument here.

Indeed, let $\varphi = (\varphi_1 \ \cdots \ \varphi_d)$ be the representation of φ with respect to the standard dual basis of $(\mathbb{R}^d)^*$ (with analogous notation for $L^*(M^{-1})^*\rho$). Since L is assumed to be purely irrational, so is $M^{-1}L$. Therefore, since $\rho : \mathbb{R}^m \to \mathbb{R}$ is surjective (since it is nonzero), we may pick some coordinate i at most d for which $(L^*(M^{-1})^*\rho)_i - \varphi_i \neq 0$. So there are algebraic numbers $\lambda_1, \dots, \lambda_m$ with algebraic degree $O_k(1)$ and algebraic height $O_{c,C,k,H}(1)$ for which

$$0 < \left| \sum_{j=1}^{m} \lambda_j \rho_j - \varphi_i \right| < c_1 C_3 \tau_2^{C_2}, \tag{E-5}$$

where $(\rho_1 \ \cdots \ \rho_m)$ is the representation of ρ with respect to the standard dual basis. Note that if c_1 is small enough, by (E-5) and the fact that $\|\rho\|_\infty = O_{c,C}(\tau_2^{-1})$ one has $|\varphi_i| = O_{c,C}(\tau_2^{-1})$.

Our aim will be to find a suitable polynomial Q for which $Q\big(\sum_{j\leqslant m} \lambda_j \rho_j\big) = 0$, and then to apply Liouville's original argument.

Assume without loss of generality that each $\lambda_j \rho_j$ is nonzero. For each j at most m, let $Q_j \in \mathbb{Z}[X]$ denote the minimal polynomial of $\lambda_j \rho_j$. Note that the degree of Q_j is $O_k(1)$ (since $\rho_j \in \mathbb{Z}$). By the bounds on the degree and height of λ_j, and since $\|\rho\|_\infty = O_{c,C}(\tau_2^{-1})$, we have $H(Q_j) = O_{c,C,k,H}(\tau_2^{-O_k(1)})$.

By using the standard construction based on resultants (see [Cohen 1993, Section 4.2.1]), this implies that there is a polynomial $Q \in \mathbb{Z}[X]$ with degree $O_k(1)$ such that $Q\big(\sum_{j\leqslant m} \lambda_j \rho_j\big) = 0$ and $H(Q) = O_{c,C,k,H}(\tau_2^{-O_k(1)})$.

Now, it could be that φ_i is a root of Q. If this is the case, we use the factor theorem and Gauss' lemma to replace Q by the integer-coefficient polynomial $Q \cdot (X - \varphi_i)^{-1}$. In this case, $H(Q \cdot (X-\varphi_i)^{-1}) \ll_{c,C,k,H} (\varphi_i + 1)^{O_k(1)} \tau_2^{-O_k(1)}$. By repeating this process as necessary, since $|\varphi_i| = O_{c,C}(\tau_2^{-1})$ we may assume therefore that φ_i is not a root of Q.

This immediately implies a bound on the derivative of Q, namely that, for any θ,

$$|Q'(\theta)| \ll_{c,C,k,H} \tau_2^{-O_k(1)} \sum_{0\leqslant a\leqslant O_k(1)} \theta^a.$$

But then the mean value theorem implies that for some θ in the interval $\big[\sum_j \lambda_j \alpha_j, \varphi_i\big]$ one has

$$1 \leqslant |Q(\varphi_i)| = \left| Q\left(\sum_{j=1}^{m} \lambda_j \rho_j\right) - Q(\varphi_i) \right| \leqslant |Q'(\theta)| \left| \sum_{j=1}^{m} \lambda_j \rho_j - \varphi_i \right| \ll_{c,C,k,H} c_1 C_3 \tau_2^{-O_k(1)} \tau_2^{C_2}.$$

If C_2 is large enough in terms of k, this implies that $c_1 = \Omega_{c,C,k,H}(1)$, which is a contradiction if c_1 is small enough. Therefore the lemma holds. □

Acknowledgements

We would like to thank several anonymous referees, for their very careful reading of earlier versions of this work, and Ben Green, for his advice and comments. We also benefited from conversations with Sam Chow, Trevor Wooley, and Yufei Zhao. Some of the paper was completed while the author was a Programme Associate at the Mathematical Sciences Research Institute in Berkeley, who provided excellent working conditions. During part of the project the author was supported by EPSRC grant no. EP/M50659X/1.

References

[Baker 1967] A. Baker, "On some Diophantine inequalities involving primes", *J. Reine Angew. Math.* **228** (1967), 166–181. MR Zbl

[Baker 1986] R. C. Baker, *Diophantine inequalities*, Lond. Math. Soc. Monogr. (N.S) **1**, Oxford Univ. Press, 1986. MR Zbl

[Cohen 1993] H. Cohen, *A course in computational algebraic number theory*, Grad. Texts in Math. **138**, Springer, 1993. MR Zbl

[Cook et al. 2017] B. Cook, A. Magyar, and M. Pramanik, "A Roth-type theorem for dense subsets of $\mathbb{R}^d$", *Bull. Lond. Math. Soc.* **49**:4 (2017), 676–689. MR Zbl

[Davenport 1963] H. Davenport, *Analytic methods for Diophantine equations and Diophantine inequalities*, Ann Arbor Publ., 1963. MR Zbl

[Davenport and Heilbronn 1946] H. Davenport and H. Heilbronn, "On indefinite quadratic forms in five variables", *J. Lond. Math. Soc.* **21** (1946), 185–193. MR Zbl

[Durcik et al. 2018] P. Durcik, V. Kovač, and L. Rimanić, "On side lengths of corners in positive density subsets of the Euclidean space", *Int. Math. Res. Not.* **2018**:22 (2018), 6844–6869. MR Zbl

[Freeman 2002] D. E. Freeman, "Asymptotic lower bounds and formulas for Diophantine inequalities", pp. 57–74 in *Number theory for the millennium, II* (Urbana, IL, 2000), edited by M. A. Bennett et al., Peters, Natick, MA, 2002. MR Zbl

[Gowers 2001] W. T. Gowers, "A new proof of Szemerédi's theorem", *Geom. Funct. Anal.* **11**:3 (2001), 465–588. MR Zbl

[Gowers and Wolf 2010] W. T. Gowers and J. Wolf, "The true complexity of a system of linear equations", *Proc. Lond. Math. Soc.* (3) **100**:1 (2010), 155–176. MR Zbl

[Green 2007] B. Green, "Montréal notes on quadratic Fourier analysis", pp. 69–102 in *Additive combinatorics* (Montréal, 2006), edited by A. Granville et al., CRM Proc. Lecture Notes **43**, Amer. Math. Soc., Providence, RI, 2007. MR Zbl

[Green and Tao 2008a] B. Green and T. Tao, "The primes contain arbitrarily long arithmetic progressions", *Ann. of Math.* (2) **167**:2 (2008), 481–547. MR Zbl

[Green and Tao 2008b] B. Green and T. Tao, "Quadratic uniformity of the Möbius function", *Ann. Inst. Fourier* (*Grenoble*) **58**:6 (2008), 1863–1935. MR Zbl

[Green and Tao 2010a] B. Green and T. Tao, "Linear equations in primes", *Ann. of Math.* (2) **171**:3 (2010), 1753–1850. MR Zbl

[Green and Tao 2010b] B. Green and T. Tao, "Yet another proof of Szemerédi's theorem", pp. 335–342 in *An irregular mind*, edited by I. Bárány and J. Solymosi, Bolyai Soc. Math. Stud. **21**, Bolyai Math. Soc., Budapest, 2010. MR Zbl

[Green and Tao 2012] B. Green and T. Tao, "The Möbius function is strongly orthogonal to nilsequences", *Ann. of Math.* (2) **175**:2 (2012), 541–566. MR Zbl

[Green et al. 2012] B. Green, T. Tao, and T. Ziegler, "An inverse theorem for the Gowers $U^{s+1}[N]$-norm", *Ann. of Math.* (2) **176**:2 (2012), 1231–1372. MR Zbl

[Margulis 1989] G. A. Margulis, "Discrete subgroups and ergodic theory", pp. 377–398 in *Number theory, trace formulas and discrete groups* (Oslo, 1987), edited by K. E. Aubert et al., Academic Press, Boston, 1989. MR Zbl

[McCutcheon 2005] R. McCutcheon, "FVIP systems and multiple recurrence", *Israel J. Math.* **146** (2005), 157–188. MR Zbl

[Müller 2005] W. Müller, "Systems of quadratic Diophantine inequalities", *J. Théor. Nombres Bordeaux* **17**:1 (2005), 217–236. MR Zbl

[Parsell 2002a] S. T. Parsell, "Irrational linear forms in prime variables", *J. Number Theory* **97**:1 (2002), 144–156. MR Zbl

[Parsell 2002b] S. T. Parsell, "On simultaneous diagonal inequalities, III", *Q. J. Math.* **53**:3 (2002), 347–363. MR Zbl

[Tao 2012] T. Tao, *Higher order Fourier analysis*, Grad. Stud. Math. **142**, Amer. Math. Soc., Providence, RI, 2012. MR Zbl

[Tao 2015] T. Tao, "An inverse theorem for the continuous Gowers uniformity norm", blog post, 2015, Available at https://tinyurl.com/taogowers.

[Tao and Teräväinen 2018] T. Tao and J. Teräväinen, "Odd order cases of the logarithmically averaged Chowla conjecture", *J. Théor. Nombres Bordeaux* **30**:3 (2018), 997–1015. MR Zbl

[Tao and Teräväinen 2019] T. Tao and J. Teräväinen, "The structure of logarithmically averaged correlations of multiplicative functions, with applications to the Chowla and Elliott conjectures", *Duke Math. J.* **168**:11 (2019), 1977–2027. MR Zbl

[Tao and Vu 2006] T. Tao and V. Vu, *Additive combinatorics*, Cambridge Stud. Adv. Math. **105**, Cambridge Univ. Press, 2006. MR Zbl

[Vaughan 1981] R. C. Vaughan, *The Hardy–Littlewood method*, Cambridge Tracts in Math. **80**, Cambridge Univ. Press, 1981. MR Zbl

[Walker 2019] A. Walker, "Linear inequalities in primes", 2019. To appear in *J. Anal. Math.* arXiv

[Wooley 2003] T. D. Wooley, "On Diophantine inequalities: Freeman's asymptotic formulae", pp. art. id. 30 in *Proc. Session in Analytic Number Theory and Diophantine Equations* (Bonn, Germany, 2002), edited by D. R. Heath-Brown and B. Z. Moroz, Bonner Math. Schriften **360**, Univ. Bonn, 2003. MR Zbl

Communicated by Roger Heath-Brown
Received 2019-03-06 Revised 2019-10-30 Accepted 2020-02-06

aledwalker@gmail.com *Centre de Recherches Mathématiques, Montréal QC, Canada*

msp

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1537

Modular invariants for real quadratic fields and Kloosterman sums

Nickolas Andersen and William D. Duke

We investigate the asymptotic distribution of integrals of the j-function that are associated to ideal classes in a real quadratic field. To estimate the error term in our asymptotic formula, we prove a bound for sums of Kloosterman sums of half-integral weight that is uniform in every parameter. To establish this estimate we prove a variant of Kuznetsov's formula where the spectral data is restricted to half-integral weight forms in the Kohnen plus space, and we apply Young's hybrid subconvexity estimates for twisted modular L-functions.

1. Introduction

The relationship between modular forms and quadratic fields is exceedingly rich. For instance, the Hilbert class field of an imaginary quadratic field may be generated by adjoining to the quadratic field a special value of the modular j-function. The connection between class fields of real quadratic fields and invariants of the modular group is much less understood, although there has been striking progress lately by Darmon and Vonk [2017]. Our aim in this paper is to study the asymptotic behavior of certain integrals of the modular j-function that are associated to ideal classes in a real quadratic field. Before turning to this, it is useful to make some definitions and to recall the corresponding problem in the imaginary quadratic case.

Let $\mathbb{K}$ be the quadratic field of discriminant d and let Cl_d^+ denote the narrow class group of $\mathbb{K}$. Let $h(d) = \#\mathrm{Cl}_d^+$ denote the class number. If $d < 0$ then each ideal class $A \in \mathrm{Cl}_d^+$ contains exactly one fractional ideal of the form $z_A\mathbb{Z} + \mathbb{Z}$, where

$$z_A = \frac{-b + i\sqrt{|d|}}{2a}$$

for some relatively prime integers a, b, c with $a > 0$ and $b^2 - 4ac = d$, and where z_A is in the fundamental domain

$$\mathcal{F} := \left\{ z \in \mathcal{H} : -\tfrac{1}{2} < \mathrm{Re}\, z \le \tfrac{1}{2}, |z| \ge 1 \right\}$$

for the action of the modular group $\Gamma_1 = \mathrm{PSL}_2(\mathbb{Z})$. Such z_A are called *reduced*. A beautiful result from the theory of complex multiplication states that the values $j_1(z_A)$, as A runs over ideal classes of

The authors were supported by NSF grant DMS–1701638. Duke was also supported by the Simons Foundation, award number 554649.

MSC2010: primary 11F37; secondary 11L05.

Keywords: Kloosterman sums, real quadratic fields, modular forms.

discriminant d, are conjugate algebraic integers. Here $j_1 = j - 744$ is the normalized modular j-invariant

$$j_1(z) = q^{-1} + 196\,884q + 21\,493\,760q^2 + \cdots,$$

where $q = e(z) = e^{2\pi i z}$. It follows that the trace

$$\mathrm{Tr}_d(j_1) := \frac{1}{\omega_d} \sum_{A \in \mathrm{Cl}_d^+} j_1(z_A), \tag{1-1}$$

where $\omega_{-3} = 3$, $\omega_{-4} = 2$, and $\omega_d = 1$ otherwise, is a rational integer. For example,

$$\mathrm{Tr}_{-3}(j_1) = -248, \quad \mathrm{Tr}_{-4}(j_1) = 492, \quad \mathrm{Tr}_{-7}(j_1) = -4119, \quad \mathrm{Tr}_{-8}(j_1) = 7256.$$

It is natural to ask how these values are distributed as $|d| \to \infty$. As a first approximation, it is not too hard to show that $\mathrm{Tr}_d(j_1) \sim (-1)^d \exp(\pi\sqrt{|d|})$ for large d, but in fact much more is known. In [Bruinier et al. 2006] it was observed, and in [Duke 2006] the second author proved, that

$$\mathrm{Tr}_d(j_1) - \sum_{\mathrm{Im}\, z_A > 1} e(-z_A) \sim -24h(d) \tag{1-2}$$

as $d \to -\infty$ through fundamental discriminants. The value -24 is a suitably defined "average of j_1" over the fundamental domain $\mathcal{F}$; see [Duke 2006].

Now suppose that $\mathbb{K}$ is a real quadratic field, i.e., $d > 0$. Each ideal class $A \in \mathrm{Cl}_d^+$ contains a fractional ideal of the form $w\mathbb{Z} + \mathbb{Z} \in A$ where $w \in \mathbb{K}$ is such that

$$0 < w^\sigma < 1 < w,$$

where σ is the nontrivial Galois automorphism of $\mathbb{K}$. Such w are called *reduced* (in the sense of [Zagier 1975]); unlike in the imaginary quadratic case, a given ideal class may have many reduced representatives. Let $\mathcal{S}_w$ be the oriented hyperbolic geodesic in $\mathcal{H}$ from w to w^σ, and let $\mathcal{C}_A$ be the closed geodesic obtained by projecting $\mathcal{S}_w$ to $\Gamma_1\backslash\mathcal{H}$. The choice of reduced w does not affect $\mathcal{C}_A$. One can view $\mathcal{C}_A$ in $\mathcal{H}$ as the geodesic from some point z_0 on $\mathcal{S}_w$ to $\gamma_w(z_0)$, where γ_w is the hyperbolic element which generates the stabilizer of w in Γ_1. It is well-known that

$$\mathrm{length}(\mathcal{C}_A) = 2\log \varepsilon_d,$$

where ε_d is the fundamental unit of $\mathbb{K}$.

A real quadratic analogue of the trace (1-1) is the sum of integrals

$$\mathrm{Tr}_d(j_1) := \sum_{A \in \mathrm{Cl}_d^+} \int_{\mathcal{C}_A} j_1(z) \frac{|dz|}{y}, \tag{1-3}$$

and one might ask how these invariants are distributed as the discriminant d varies. Numerically, we have

$$\mathrm{Tr}_5(j_1) \approx -11.5417, \quad \mathrm{Tr}_8(j_1) \approx -19.1374, \quad \mathrm{Tr}_{13}(j_1) \approx -23.4094, \quad \mathrm{Tr}_{17}(j_1) \approx -43.9449.$$

Note that these values are quite small even though j_1 grows exponentially in the cusp. It was conjectured in [Duke et al. 2011] that

$$\mathrm{Tr}_d(j_1) \sim -24 \cdot 2 \log \varepsilon_d \, h(d) \tag{1-4}$$

as $d \to \infty$ through fundamental discriminants. This was proved independently in [Duke et al. 2012] (for odd fundamental discriminants, with a power-saving of $d^{-1/5325}$) and in [Masri 2012] (for all fundamental discriminants, with a power-saving of $d^{-1/400}$).

The real quadratic invariants $\mathrm{Tr}_d(j_1)$ were first studied in [Duke et al. 2011] in the context of harmonic Maass forms (nonholomorphic modular forms which are annihilated by the hyperbolic Laplacian). There is a family of harmonic Maass forms $\{f_{d'}\}$ of weight $\frac{1}{2}$, indexed by positive discriminants d', whose Fourier coefficients can be written in terms of the sums (1-3) twisted by genus characters. For each factorization $D = dd'$ of the fundamental discriminant D into fundamental discriminants d, d', there is a real character $\chi_d = \chi_{d'}$ of Cl_D^+ called a genus character. The d-th Fourier coefficient of $f_{d'}$ is given by

$$\mathrm{Tr}_{d,d'}(j_1) := \sum_{A \in \mathrm{Cl}_D^+} \chi_d(A) \int_{\mathcal{C}_A} j_1(z) \frac{|dz|}{y}.$$

In particular, the d-th Fourier coefficient of f_1 is $\mathrm{Tr}_d(j_1)$. The remaining non-square-indexed coefficients can be described in terms of $\mathrm{Tr}_{d,d'}(j_m)$ for $m \geq 1$, where j_m is the unique modular function in $\mathbb{C}[j]$ of the form $j_m = q^{-m} + O(q)$. Our first result concerns the asymptotic distribution of the values of $\mathrm{Tr}_{d,d'}(j_m)$ as any of the parameters d, d', m tends to infinity. We define $\delta_1 = 1$ and $\delta_d = 0$ otherwise, and $\sigma_s(n) = \sum_{\ell | n} \ell^s$ for any $s \in \mathbb{C}$.

Theorem 1.1. *For each positive fundamental discriminant D, let d be any positive fundamental discriminant dividing D. Then for each $m \geq 1$ we have*

$$\sum_{A \in \mathrm{Cl}_D^+} \chi_d(A) \int_{\mathcal{C}_A} j_m(z) \frac{|dz|}{y} = -24 \, \delta_d \, \sigma_1(m) \cdot 2h(D) \log \varepsilon_D + O\big(m^{8/9} D^{13/27} (mD)^{\varepsilon}\big). \tag{1-5}$$

Remarks. In the case $d = 1$, the power-saving of $D^{-1/54}$ in Theorem 1.1 improves on the results of [Masri 2012; Duke et al. 2012]. The generalizations to $d > 1$ and $m > 1$ are new, and the latter confirms the observation in [Duke et al. 2011] that $\mathrm{Tr}_D(j_m) \sim -24\sigma_1(m) \cdot 2 \log \varepsilon_D h(D)$ as $m \to \infty$.

When $D = dd'$ is a factorization of D into negative fundamental discriminants, the left-hand side of (1-5) is identically zero. To see this, let J denote the class of the different $(\sqrt{D})$ of $\mathbb{K}$. The closed geodesic associated to JA^{-1} has the same image in $\Gamma_1 \backslash \mathcal{H}$ as $\mathcal{C}_A$ but with the opposite orientation. Since $\chi_d(J) = \operatorname{sgn} d$, the left-hand side of (1-5) is forced to vanish whenever $d < 0$.

In order to give a better geometric interpretation when $D = dd'$ where d and d' are negative, Imamoğlu, Tóth, and the second author [Duke et al. 2016] recently defined a new invariant $\mathcal{F}_A$, which is a finite area hyperbolic surface with boundary $\mathcal{C}_A$. We briefly describe the construction of $\mathcal{F}_A$; see [Duke et al. 2016]

for details. Let w be one of the reduced quadratic irrationalities associated to A, and let $\gamma_w \in \Gamma_1$ be the hyperbolic element that fixes w and w^σ. Then γ_w can be written as

$$\gamma_w = T^{\lceil w \rceil} S T^{n_1} S T^{n_2} S \cdots T^{n_\ell} S \tag{1-6}$$

for some integers $n_i \geq 2$, where $T = \pm\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $S = \pm\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ are generators of Γ_1. The cycle $(n_1, \dots n_\ell)$ is the period of the minus continued fraction of w, and ℓ is the number of distinct reduced representatives of A. Let $S_k := T^{(n_1+\cdots+n_k)} S T^{-(n_1+\cdots+n_k)}$ and define

$$\Gamma_A := \langle S_1, \dots, S_\ell, T^{(n_1+\cdots+n_\ell)} \rangle.$$

This group is an infinite-index (i.e., thin) subgroup of Γ_1. Let $\mathcal{N}_A$ be the Nielsen region of Γ_A: the smallest nonempty Γ_A-invariant open convex subset of $\mathcal{H}$. Then the surface $\mathcal{F}_A$ is defined as $\Gamma_A \backslash \mathcal{N}_A$. A different choice of reduced w representing A yields a subgroup of Γ_1 conjugate to Γ_A by a translation, so the surface $\mathcal{F}_A$ is uniquely defined by A; see Theorem 1 of [Duke et al. 2016]. In that theorem we also find that the area of $\mathcal{F}_A$ is $\pi\ell$, with ℓ as in (1-6).

Our second result concerns the distribution of sums of the integrals of j_m over the surfaces $\mathcal{F}_A$ as the discriminant varies. The functions j_m grow exponentially in the cusp, so we regularize the integrals as follows. For each $Y \geq 1$, let $\mathcal{F}_{A,Y} = \mathcal{F}_A \cap \{z : \operatorname{Im} z \leq Y\}$. We define

$$\int_{\mathcal{F}_A} j_m(z) \frac{dx\,dy}{y^2} := \lim_{Y \to \infty} \int_{\mathcal{F}_{A,Y}} j_m(z) \frac{dx\,dy}{y^2}. \tag{1-7}$$

These real quadratic invariants are asymptotically related to products of class numbers of imaginary quadratic fields.

Theorem 1.2. *For each positive fundamental discriminant D, let $D = dd'$ be any factorization into negative fundamental discriminants. Then for each $m \geq 1$ we have*

$$\frac{1}{4\pi} \sum_{A \in \mathrm{Cl}_D^+} \chi_d(A) \int_{\mathcal{F}_A} j_m(z) \frac{dx\,dy}{y^2} = -24\sigma_1(m) \frac{h(d)h(d')}{\omega_d \omega_{d'}} + O\big(m^{8/9} D^{13/27} (mD)^{\varepsilon}\big). \tag{1-8}$$

Remark. When $D = dd'$ is a factorization into positive discriminants, the left-hand side of (1-8) is identically zero because $A \mapsto JA^{-1}$ reverses the orientation of the surface $\mathcal{F}_A$.

An interesting special case occurs when $D = 4p$ where $p \equiv 3 \pmod 4$ is a prime. In this case the identity class $I = I_p$ is not equivalent to the class of the different $J = J_p$. The Cohen-Lenstra heuristics predict that approximately 75% of such fields have wide class number one, which would imply that the classes containing I and J are the only ideal classes. If this is the case, then there is a sequence of primes $p \equiv 3 \pmod 4$ for which

$$\int_{\mathcal{F}_{I_p}} j_1(z) \frac{dx\,dy}{y^2} \sim -2\pi h(-p) \quad \text{and} \quad \int_{\mathcal{F}_{J_p}} j_1(z) \frac{dx\,dy}{y^2} \sim 2\pi h(-p).$$

The method used in [Duke 2006] to prove (1-2) and in [Masri 2012] to prove (1-4) involves the equidistribution of CM points and closed geodesics originally developed in [Duke 1988]. By contrast,

here we employ a relation between the invariants in (1-5) and (1-8) and sums of Kloosterman sums; see Section 2. We then estimate the sums of Kloosterman sums directly via a Kuznetsov-type formula.

The Kloosterman sums in question are those which appear in the Fourier coefficients of Poincaré series of half-integral weight in the Kohnen plus space. In weight $k = \lambda + \frac{1}{2}$, the plus space consists of holomorphic or Maass cusp forms whose Fourier coefficients are supported on exponents n such that $(-1)^\lambda n \equiv 0, 1 \pmod 4$. For integers m, n satisfying the plus space condition and c a positive integer divisible by 4 we define

$$S_k^+(m,n,c) := e\Big(-\frac{k}{4}\Big) \sum_{d \bmod c} \Big(\frac{c}{d}\Big) \varepsilon_d^{2k}\, e\Big(\frac{m\bar{d}+nd}{c}\Big) \times \begin{cases} 1 & \text{if } 8 \mid c, \\ 2 & \text{if } 4 \,||\, c, \end{cases} \tag{1-9}$$

where $d\bar{d} \equiv 1 \pmod c$ and $\varepsilon_d = 1$ or i according to $d \equiv 1$ or $3 \pmod 4$, respectively. The Kloosterman sums (1-9) are real-valued and satisfy the relation

$$S_k^+(m,n,c) = S_{-k}^+(-m,-n,c). \tag{1-10}$$

We prove a strong uniform bound for these sums which is of independent interest. We remark that similar (but weaker) estimates are hiding in the background of the methods of [Duke 2006; Masri 2012].

Theorem 1.3. *Let $k = \pm\frac{1}{2} = \lambda + \frac{1}{2}$. Suppose that m, n are positive integers such that $(-1)^\lambda m = v^2 d'$ and $(-1)^\lambda n = w^2 d$, where d, d' are fundamental discriminants not both equal to* 1. *Then*

$$\sum_{4|c\le x} \frac{S_k^+(m,n,c)}{c} \ll \big(x^{1/6} + (dd')^{2/9}(vw)^{1/3}\big)(mnx)^\varepsilon. \tag{1-11}$$

Friedlander, Iwaniec, and the second author [Duke et al. 2012] proved an analogous estimate for smoothed sums of Kloosterman sums on $\Gamma_0(4q)$ with a power saving of $n^{-1/1330}$ when n is squarefree. Individually, the Kloosterman sums satisfy the Weil-type bound

$$|S_k^+(m,n,c)| \le 2\sigma_0(c)\gcd(m,n,c)^{1/2}\sqrt{c}, \tag{1-12}$$

(see, e.g., Lemma 6.1 of [Duke et al. 2012]) so the sum in (1-11) is trivially bounded above by $(mnx)^\varepsilon\sqrt{x}$. Theorem 1.3 should be compared with the bound of [Sarnak and Tsimerman 2009] for the ordinary integral weight Kloosterman sums $S(m,n,c)$ which improves on the pivotal result of [Kuznetsov 1980]. The main result of [Sarnak and Tsimerman 2009] is unconditional and depends on progress toward the Ramanujan conjecture for Maass cusp forms of weight 0. Assuming that conjecture, their theorem states that

$$\sum_{c\le x} \frac{S(m,n,c)}{c} \ll \big(x^{1/6} + (mn)^{1/6}\big)(mnx)^\varepsilon.$$

Our method also yields an exponent of $\frac{1}{6}$ for dd' in (1-11) if we assume the Lindelöf hypothesis for $L\big(\frac{1}{2}, \chi\big)$ and $L\big(\frac{1}{2}, f \times \chi\big)$, where χ is a quadratic Dirichlet character and f is an integral weight cusp form (holomorphic or Maass). Via the correspondence of Waldspurger, the Lindelöf hypothesis for all such $L\big(\frac{1}{2}, f \times \chi\big)$ is equivalent to the Ramanujan conjecture for half-integral weight forms.

Recently Ahlgren and the first author used a similar approach to study the half-integral weight Kloosterman sums associated to the multiplier system for the Dedekind eta function. This was used in [Ahlgren and Andersen 2018] to improve the error bounds of [Lehmer 1939; Folsom and Masri 2010] for the classical formula of Hardy, Ramanujan, and Rademacher for the partition function $p(n)$. In particular, it was shown that the discrepancy between $p(n)$ and the first $O(\sqrt{n})$ terms in the formula is at most $O(n^{-(1/2)-(1/168)+\varepsilon})$.

The proof of Theorem 1.3 hinges on a version of Kuznetsov's formula which relates the Kloosterman sums (1-9) to the coefficients of holomorphic cusp forms, Maass cusp forms, and Eisenstein series of half-integral weight in the plus space. One advantage of the plus space is that the Waldspurger correspondence is completely explicit on that space via [Kohnen and Zagier 1981; Baruch and Mao 2010]; knowledge of the exact proportionality constant in the Waldspurger correspondence is crucial for us. Here we briefly define the relevant quantities and state a special case of our version of the Kuznetsov formula. Let $\mathcal{H}_k^+$ (resp. $\mathcal{V}_k^+$) denote an orthonormal Hecke basis for the plus space of holomorphic (resp. Maass) cusp forms of weight k for $\Gamma_0(4)$. For each $g \in \mathcal{H}_k^+$ (resp. $u_j \in \mathcal{V}_k^+$) let $\rho_g(n)$ (resp. $\rho_j(n)$) denote the suitably normalized n-th Fourier coefficient of g (resp. u_j). For each j, let $\lambda_j = \frac{1}{4} + r_j^2$ denote the Laplace eigenvalue of u_j. The full statement with detailed definitions appears in Section 5 below.

Theorem 1.4. *Let $k = \pm\frac{1}{2} = \lambda + \frac{1}{2}$. Suppose that m, n are positive integers such that $(-1)^\lambda m$ and $(-1)^\lambda n$ are fundamental discriminants. Suppose that $\varphi : [0, \infty) \to \mathbb{R}$ is a smooth test function which satisfies (5-1), and let $\widetilde{\varphi}$ and $\widehat{\varphi}$ denote the integral transforms in (5-2)–(5-3). Then*

$$\begin{aligned}\sum_{4|c>0} \frac{S_k^+(m,n,c)}{c}\varphi\left(\frac{4\pi\sqrt{mn}}{c}\right) &= 6\sqrt{mn}\sum_{u_j\in\mathcal{V}_k^+}\frac{\overline{\rho_j(m)}\rho_j(n)}{\cosh \pi r_j}\widehat{\varphi}(r_j) + \frac{3}{2}\sum_{\ell\equiv k \bmod 2} e\left(\tfrac{1}{4}(\ell-k)\right)\widetilde{\varphi}(\ell)\Gamma(\ell)\sum_{g\in\mathcal{H}_\ell^+}\overline{\rho_g(m)}\rho_g(n)\\ &\quad + \int_{-\infty}^{\infty}\left(\frac{n}{m}\right)^{ir}\frac{L\left(\frac{1}{2}-2ir,\,\chi_{(-1)^\lambda m}\right)L\left(\frac{1}{2}+2ir,\,\chi_{(-1)^\lambda n}\right)}{2\cosh \pi r\left|\Gamma\left(\frac{1}{2}(k+1)+ir\right)\right|^2|\zeta(1+4ir)|^2}\widehat{\varphi}(r)\,dr. \quad (1\text{-}13)\end{aligned}$$

Remark. This version of the Kuznetsov formula for Maass forms in the plus space for $\Gamma_0(4)$ with weight $\pm\frac{1}{2}$ is precisely analogous to the original version of Kuznetsov's formula for the full modular group. To prove it we apply Biró's idea [2000] of taking a linear combination of Proskurin's Kuznetsov-type formula evaluated at various cusp-pairs in order to project the holomorphic and Maass cusp forms into the plus space. The main technical complication arises from the sum of Eisenstein series terms from Proskurin's formula, which we show simplifies to the integral of Dirichlet L-functions in (1-13). The simplicity of that integral is reminiscent of the corresponding term in Kuznetsov's original formula [1980, Theorem 2] for the ordinary weight 0 Kloosterman sums; in that formula, the Eisenstein term is

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\left(\frac{n}{m}\right)^{ir}\frac{\sigma_{2ir}(m)\sigma_{-2ir}(n)}{|\zeta(1+2ir)|^2}\widehat{\varphi}(r)\,dr.$$

Note that if $k = 0$ then $\cosh \pi r \left|\Gamma\left(\frac{1}{2}(k+1) + ir\right)\right|^2 = \pi$.

The most crucial input in the proof of Theorem 1.3 is Young's Weyl-type hybrid subconvexity estimates [2017] for $L\left(\frac{1}{2}, f \times \chi_d\right)$ and $L\left(\frac{1}{2}, \chi_d\right)$ which improve on the groundbreaking results of Conrey and Iwaniec [2000]. Young proved that

$$\sum_f L\left(\tfrac{1}{2}, f \times \chi_d\right)^3 \ll (kd)^{1+\varepsilon} \tag{1-14}$$

for odd fundamental discriminants d, where the sum is over all holomorphic newforms of weight k and level dividing d. In the Appendix we sketch the details required to generalize Young's result to twists by χ_d for even fundamental discriminants d, where the sum is over f of level dividing the squarefree part of d. The uniformity of Young's result in both the level and weight directly influences the quality of the exponents in (1-11). There are corresponding results in [Young 2017] for twisted L-functions of Maass cusp forms and Dirichlet L-functions which we also use in the proof of Theorem 1.3.

Remark. The condition in (1-14) (and our extension in the Appendix) that f have level dividing the squarefree part of d (which is odd unless $d = 4q$ with $q \equiv 2 \pmod 4$) is why we require a Kuznetsov formula that involves only coefficients of cusp forms in the plus space. Under the Shimura correspondence, the plus spaces of half-integral weight forms on $\Gamma_0(4)$ are isomorphic as Hecke modules to spaces of weight 0 cusp forms on $\Gamma_0(1)$, whereas the full spaces on $\Gamma_0(4)$ lift to $\Gamma_0(2)$.

The paper is organized as follows. In Section 2 we use the formulas of [Duke et al. 2016] to relate the geometric invariants to sums of Kloosterman sums, and we apply Theorem 1.3 to prove Theorems 1.1 and 1.2. The remainder of the paper is dedicated to the proof of Theorem 1.3. In Section 3 we give some background on the spectrum of the hyperbolic Laplacian in half-integral weight. In Section 4 we prove general estimates for the mean square of Fourier coefficients of Maass cusp forms of half-integral weight with arbitrary multiplier system. We prove Theorem 1.4 in Section 5 and Theorem 1.3 in Section 6. Finally, the Appendix contains a sketch of the proof of Young's subconvexity result extended to even discriminants.

2. Geometric invariants and Kloosterman sums

In this section we relate the real quadratic invariants to Kloosterman sums and show how Theorems 1.1 and 1.2 follow from Theorem 1.3. Actually, we will prove more general forms of the main theorems which allow for nonfundamental discriminants. It is convenient to use binary quadratic forms

$$Q(x, y) = [a, b, c] = ax^2 + bxy + cy^2$$

in place of ideal classes, as this point of view makes the generalization to arbitrary discriminants straightforward. A discriminant is any integer $D \equiv 0, 1 \pmod 4$. A discriminant D is fundamental if it is either odd and squarefree or if $D/4$ is squarefree and congruent to $2, 3 \pmod 4$. Fix a discriminant $D > 1$ and a factorization $D = dd'$ into positive or negative discriminants d, d' such that d is fundamental. Let $\mathcal{Q}_D$ be the set of all integral binary quadratic forms $[a, b, c]$ with discriminant $b^2 - 4ac = D$. The

modular group Γ_1 acts on $\mathcal{Q}_D$ in the usual way. When D is fundamental all forms in $\mathcal{Q}_D$ are primitive (i.e., $\gcd(a,b,c)=1$) and there is a simple correspondence between $\Gamma_1\backslash\mathcal{Q}_D$ and Cl_D^+ via

$$[a,b,c]\mapsto w\mathbb{Z}+\mathbb{Z}, \qquad \text{where } w=\frac{-b+\sqrt{D}}{2a}, \tag{2-1}$$

assuming $[a,b,c]$ is chosen in its class to have $a>0$. If D is fundamental and if Q corresponds to A via (2-1) then we define $\mathcal{C}_Q:=\mathcal{C}_A$ and $\mathcal{F}_Q:=\mathcal{F}_A$. We extend this to arbitrary discriminants via $\mathcal{C}_{\delta Q}:=\mathcal{C}_Q$ and $\mathcal{F}_{\delta Q}:=\mathcal{F}_Q$. There is a generalized genus character χ_d on $\Gamma_1\backslash\mathcal{Q}_D$ (see [Gross et al. 1987, I.2]) associated to the factorization $D=dd'$ defined by

$$\chi_d(Q)=\begin{cases}\left(\frac{d}{n}\right) & \text{if } (a,b,c,d)=1 \text{ and } Q \text{ represents } n \text{ and } (d,n)=1,\\ 0 & \text{if } (a,b,c,d)>1.\end{cases}$$

If D is fundamental then $\chi_d=\chi_{d'}$ is the usual genus character, and there is exactly one such character for each such factorization.

Recall the integral (1-7). There is an equivalent, and often useful, regularization which does not involve a limit, which we describe here. Following [Duke et al. 2018], for $z,\tau\in\mathcal{H}$ we define

$$K(z,\tau):=\frac{j'(\tau)}{j(z)-j(\tau)},$$

where $j':=(1/(2\pi i))(dj/dz)$. This function transforms on Γ_1 with weight 0 in z and weight 2 in τ. For each indefinite quadratic form Q define

$$\nu_Q(z):=\int_{\mathcal{C}_Q}K(z,\tau)d\tau.$$

As explained in [Duke et al. 2018], for $z\notin C_Q$ the value of $\nu_Q(z)$ is an integer which counts with signs the number of crossings that a path from $i\infty$ to z in $\mathcal{F}$ makes with $\mathcal{C}_Q$. Furthermore, $\nu_Q(z)$ is Γ_1-invariant and is identically zero for $\operatorname{Im} z$ sufficiently large. It follows that the integral

$$\int_{\mathcal{F}}j_m(z)\nu_Q(z)\frac{dx\,dy}{y^2}$$

converges, and it is not difficult to show that this regularization agrees with (1-7).

The following theorem generalizes Theorems 1.1 and 1.2 to more general discriminants.

Theorem 2.1. *For each positive nonsquare discriminant D, let $D=dd'$ be any factorization into discriminants such that d is fundamental. Let m be any positive integer. If d is positive, we have*

$$\sum_{Q\in\Gamma_1\backslash\mathcal{Q}_D}\chi_d(Q)\int_{\mathcal{C}_Q}j_m(z)\frac{|dz|}{y}=-24\,\delta_d\,\sigma_1(m)\cdot 2h(D)\log\varepsilon_D+O\big(m^{8/9}D^{13/27}(mD)^{\varepsilon}\big),$$

while if d is negative, we have

$$\frac{1}{4\pi}\sum_{Q\in\Gamma_1\backslash\mathcal{Q}_D}\chi_d(Q)\int_{\mathcal{F}}j_m(z)\nu_Q(z)\frac{dx\,dy}{y^2}=-24\sigma_1(m)\frac{h(d)h(d')}{\omega_d\omega_{d'}}+O\big(m^{8/9}D^{13/27}(mD)^{\varepsilon}\big).$$

To deduce Theorem 2.1 from Theorem 1.3 we require several results from [Duke et al. 2016, §8–9], which we borrow from freely here. For $m \geq 0$, let $F_{-m}(z, s)$ denote the index $-m$ nonholomorphic Poincaré series and let

$$j_m(z, s) := 2\pi m^{1/2} F_{-m}(z, s) - \frac{2\pi m^{1-s}\sigma_{2s-1}(m)}{\pi^{-(s+\frac{1}{2})}\Gamma\big(s+\frac{1}{2}\big)\zeta(2s-1)} F_0(z, s).$$

For $m \geq 1$ the Fourier expansion of $F_{-m}(z, s)$ shows that it has an analytic continuation to $\mathrm{Re}(s) > \frac{3}{4}$. In particular, $F_{-m}(z, 1)$ is holomorphic as a function of z. Furthermore, $F_0(z, s)$ is the nonholomorphic Eisenstein series of weight $\frac{1}{2}$, and we have

$$\lim_{s\to 1} \frac{2\pi m^{1-s}\sigma_{2s-1}(m)}{\pi^{-(s+\frac{1}{2})}\Gamma\big(s+\frac{1}{2}\big)\zeta(2s-1)} F_0(z, s) = 24\sigma_1(m).$$

A computation then shows that $j_m(z) = j_m(z, 1)$ for $m \geq 1$; see [Duke et al. 2011, (4.11)].

Since the length of $\mathcal{C}_Q$ is $2\log\varepsilon_D$ for every $Q \in \mathcal{Q}_D$, we have

$$\sum_{Q\in\Gamma_1\backslash\mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{C}_Q} \frac{|dz|}{y} = 2\log\varepsilon_D \sum_{Q\in\Gamma_1\backslash\mathcal{Q}_D} \chi_d(Q) = 2\delta_d h(D)\log\varepsilon_D.$$

By Corollary 4 of [Duke et al. 2018], we have (note that $w_d = 2\omega_d$ in that paper)

$$\frac{1}{4\pi} \sum_{Q\in\Gamma_1\backslash\mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{F}} \nu_Q(z) \frac{dx\,dy}{y^2} = \frac{h(d)h(d')}{\omega_d\omega_d'}.$$

So to prove Theorem 2.1 it suffices to show that

$$\sqrt{m} \sum_{Q\in\Gamma_1\backslash\mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{C}_Q} F_{-m}(z, 1) \frac{|dz|}{y} \ll m^{8/9} D^{13/27} (mD)^{\varepsilon} \tag{2-2}$$

and

$$\sqrt{m} \sum_{Q\in\Gamma_1\backslash\mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{F}} F_{-m}(z, 1)\nu_Q(z) \frac{dx\,dy}{y^2} \ll m^{8/9} D^{13/27} (mD)^{\varepsilon}. \tag{2-3}$$

We will prove (2-2)–(2-3) by relating the integrals of $F_{-m}(z, 1)$ to the quadratic Weyl sums

$$T_m(d', d; c) := \sum_{\substack{b \bmod c \\ b^2\equiv D \bmod c}} \chi_d\Big(\Big[\frac{c}{4}, b, \frac{b^2-D}{c}\Big]\Big) e\Big(\frac{2mb}{c}\Big).$$

Here we are still assuming that $D = dd'$ with d fundamental. Note that $T_m(d', d; c) = S_m(d', d; c)$ in the notation of [Duke et al. 2016]; we have changed the notation here to avoid confusion with the Kloosterman sums. The Weyl sums are related to the plus space Kloosterman sums via Kohnen's identity

$$T_m(d, d'; c) = \sum_{n|(m,c/4)} \Big(\frac{d}{n}\Big)\sqrt{\frac{2n}{c}}\, S^+_{1/2}\Big(d', \frac{m^2}{n^2}d; \frac{c}{n}\Big); \tag{2-4}$$

see Lemma 8 of [Duke et al. 2016]. The Weil bound (1-12) for Kloosterman sums shows that

$$T_m(d, d'; c) \ll \gcd(d', m^2 d, c)^{1/2} c^{\varepsilon}.$$

A direct corollary of Theorem 1.3 is the following bound for the Weyl sums.

Theorem 2.2. *Suppose that $D = dd'$ is a positive nonsquare discriminant and that d is a fundamental discriminant. Then for any $m \geq 1$ we have*

$$\sum_{4|c\leq x} \frac{T_m(d,d';c)}{\sqrt{c}} \ll \left(x^{1/6} + D^{2/9}m^{1/3}\right)(mDx)^{\varepsilon}. \tag{2-5}$$

Proof. When d, d' are positive this is immediate from (2-4) and the $k = \frac{1}{2}$ case of Theorem 1.3. When d, d' are negative we apply (1-10) after (2-4). Then the estimate (2-5) follows from the $k = -\frac{1}{2}$ case of Theorem 1.3. □

We are now ready to prove (2-2)–(2-3).

Proof of (2-2). Let $J_\nu(x)$ denote the J-Bessel function

$$J_\nu(2x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+\nu}}{k!\Gamma(\nu+k+1)}. \tag{2-6}$$

By Lemma 4 of [Duke et al. 2016] we have

$$\sum_{Q\in\Gamma\backslash\mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{C}_Q} F_{-m}(z,s)\frac{|dz|}{y} = 2^{s-\frac{1}{2}}\frac{\Gamma\left(\frac{1}{2}s\right)^2}{\Gamma(s)} D^{1/4} \sum_{0<c\equiv 0(4)} \frac{T_m(d',d;c)}{\sqrt{c}} J_{s-\frac{1}{2}}\left(\frac{4\pi m\sqrt{D}}{c}\right) \tag{2-7}$$

for $\mathrm{Re}(s) > 1$. By (2-6) we find that $J_\nu(1/x) \ll x^{-\nu}$ and $[J_\nu(1/x)]' \ll x^{-\nu-1}$ as $x \to \infty$ uniformly for $\nu \in \left[\frac{1}{2}, 1\right]$. Let $a = 4\pi m\sqrt{D}$ and let $N \geq a$. Suppose that $s \in \left[1, \frac{3}{2}\right]$. Then by partial summation and Theorem 2.2 we have

$$\sum_{4|c\geq N} \frac{T_m(d',d;c)}{\sqrt{c}} J_{s-\frac{1}{2}}\left(\frac{4\pi m\sqrt{D}}{c}\right)$$
$$= \lim_{x\to\infty} S(x) J_{s-\frac{1}{2}}\left(\frac{a}{x}\right) - S(N) J_{s-\frac{1}{2}}\left(\frac{a}{N}\right) - \int_N^\infty S(t)\left(J_{s-\frac{1}{2}}\left(\frac{a}{t}\right)\right)' dt \ll_a N^{-\frac{1}{3}+\varepsilon},$$

where $S(x)$ denotes the partial sum on the left-hand side of (2-5). It follows that the sum on the right-hand side of (2-7) converges uniformly for $s \in \left[1, \frac{3}{2}\right]$. Since $J_{\frac{1}{2}}(x) = \sqrt{2/\pi x}\,\sin x$ we conclude that

$$\sqrt{m} \sum_{Q\in\Gamma\backslash\mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{C}_Q} F_{-m}(z,1)\frac{|dz|}{y} = \sum_{0<c\equiv 0(4)} T_m(d',d;c)\sin\left(\frac{4\pi m\sqrt{D}}{c}\right). \tag{2-8}$$

We split the sum at $c = A$ with $A \ll m\sqrt{D}$. Estimating the initial segment $c \leq A$ trivially, we obtain

$$\sum_{c\leq A} T_m(d',d;c)\sin\left(\frac{4\pi m\sqrt{D}}{c}\right) \ll A(mDA)^{\varepsilon}. \tag{2-9}$$

Then by partial summation we have

$$\sum_{c>A} T_m(d', d; c) \sin\left(\frac{4\pi m\sqrt{D}}{c}\right) = -S(A)\sqrt{A}\,\sin\left(\frac{4\pi m\sqrt{D}}{A}\right) - \int_A^\infty S(t)\left(\sqrt{t}\,\sin\left(\frac{4\pi m\sqrt{D}}{t}\right)\right)' dt,$$

where $S(x)$ denotes the partial sum on the left-hand side of (2-5). Since

$$\left(\sqrt{t}\,\sin\left(\frac{4\pi m\sqrt{D}}{t}\right)\right)' \ll \frac{m\sqrt{D}}{t^{3/2}},$$

we conclude that

$$\sum_{c>A} T_m(d', d; c) \sin\left(\frac{4\pi m\sqrt{D}}{c}\right) \ll \left(mD^{1/2}A^{-1/3} + m^{4/3}D^{13/18}A^{-1/2}\right)(mDA)^{\varepsilon}. \tag{2-10}$$

Letting $A = m^a D^b$, we choose $a = \frac{8}{9}$ and $b = \frac{13}{27}$ to balance the exponents in (2-9) and (2-10). This, together with (2-8), yields (2-2). □

Proof of (2-3). Define $\mathcal{F}_Y := \mathcal{F} \cap \{z : \operatorname{Im} z \le Y\}$. Let $Q \in \mathcal{Q}_D$, and let Y be sufficiently large so that $v_Q(z) = 0$ for $\operatorname{Im} z > Y$ and so that the image of $\mathcal{C}_Q$ in $\mathcal{F}$ is contained in $\mathcal{F}_Y$. Then for $\operatorname{Re} s > 1$ we have

$$\int_{\mathcal{F}} F_{-m}(z, s) v_Q(z) \frac{dx\,dy}{y^2} = \int_{\mathcal{C}_Q} \int_{\mathcal{F}_Y} F_{-m}(z, s) K(\tau, z) \frac{dx\,dy}{y^2}\, d\tau.$$

The function $F_{-m}(z, s)$ satisfies the relation

$$\Delta_0 F_{-m}(z, s) := -4y^2 \partial_{\bar{z}} \partial_z F_{-m}(z, s) = s(1-s) F_{-m}(z, s);$$

see Section 8 of [Duke et al. 2016]. So by the proof of Lemma 1 of [Duke et al. 2018] (essentially an application of Stokes' theorem), we find that

$$\frac{s(1-s)}{2} \int_{\mathcal{F}_Y} F_{-m}(z, s) K(\tau, z) \frac{dx\,dy}{y^2} = i\,\partial_\tau F_{-m}(\tau, s).$$

It follows that

$$\frac{s(1-s)}{2} \int_{\mathcal{F}} F_{-m}(z, s) v_Q(z) \frac{dx\,dy}{y^2} = \int_{\mathcal{C}_Q} i\,\partial_z F_{-m}(z, s)\, dz.$$

Differentiating with respect to s and setting $s = 1$ we conclude that

$$\int_{\mathcal{F}} F_{-m}(z, 1) v_Q(z) \frac{dx\,dy}{y^2} = -2\partial_s \int_{\mathcal{C}_Q} i\,\partial_z\, F_{-m}(z, s)\, dz \Big|_{s=1}.$$

By Lemma 5 of [Duke et al. 2016] we have

$$\sum_{Q \in \Gamma\backslash\mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{C}_Q} i\,\partial_z F_{-m}(z, s) dz = 2^{s-\frac{1}{2}} \frac{\Gamma\left(\frac{1}{2}(s+1)\right)^2}{\Gamma(s)} D^{\frac{1}{4}} \sum_{0<c\equiv 0(4)} \frac{T_m(d', d; c)}{\sqrt{c}} J_{s-\frac{1}{2}}\left(\frac{4\pi m\sqrt{D}}{c}\right).$$

A straightforward computation involving (2-6) shows that, uniformly for $s \in \left[1, \frac{3}{2}\right]$, we have

$$\partial_s \left[2^{s-\frac{1}{2}} \frac{\Gamma\left(\frac{1}{2}(s+1)\right)^2}{\Gamma(s)} J_{s-\frac{1}{2}}(x) \right] \ll x^{s-\frac{1}{2}} |\log x| \quad \text{as} \quad x \to 0^+.$$

Thus, an argument involving partial summation, as in the proof of (2-2), shows that we are justified in setting $s = 1$, and we obtain

$$\sqrt{m} \sum_{Q \in \Gamma \backslash \mathcal{Q}_D} \chi_d(Q) \int_{\mathcal{F}} F_{-m}(z, 1) \nu_Q(z) \frac{dx\, dy}{y^2} = -\frac{2}{\pi} \sum_{0 < c \equiv 0(4)} T_m(d', d; c) f\left(\frac{4\pi m \sqrt{D}}{c}\right),$$

where

$$f(x) := \mathrm{Ci}(2x) \sin(x) - \mathrm{Si}(2x) \cos(x) + \log(2) \sin(x)$$

and Ci, Si are the cosine and sine integrals, respectively. The remainder of the proof is quite similar to the proof of (2-2) because we have

$$f(x) \ll \min\{1, x|\log x|\} \quad \text{and} \quad \left(\sqrt{t}\, f\left(\frac{4\pi m \sqrt{D}}{t}\right)\right)' \ll \frac{m\sqrt{D}}{t^{\frac{3}{2}}} (mDt)^{\varepsilon}.$$

We omit the details. □

3. Background

In this section we recall several facts about automorphic functions which transform according to multiplier systems of half-integral weight k, and the spectrum of the hyperbolic Laplacian Δ_k in this setting. For more details see [Duke et al. 2002; Sarnak 1984; Proskurin 2003; Ahlgren and Andersen 2018] along with the original papers of Maass [1949; 1952], Roelcke [1966], and Selberg [1956; 1965].

Let $\Gamma = \Gamma_0(N)$ for some $N \geq 1$, and let k be a real number. We say that $\nu : \Gamma \to \mathbb{C}^\times$ is a multiplier system of weight k if

(i) $|\nu| = 1$,

(ii) $\nu(-I) = e^{-\pi i k}$, and

(iii) $\nu(\gamma_1 \gamma_2) = w(\gamma_1, \gamma_2) \nu(\gamma_1) \nu(\gamma_2)$ for all $\gamma_1, \gamma_2 \in \Gamma$, where

$$w(\gamma_1, \gamma_2) = j(\gamma_2, z)^k j(\gamma_1, \gamma_2 z)^k j(\gamma_1 \gamma_2, z)^{-k},$$

and $j(\gamma, z)$ is the automorphy factor

$$j(\gamma, z) := \frac{cz+d}{|cz+d|} = e^{i \arg(cz+d)}.$$

If ν is a multiplier system of weight k, then $\bar{\nu}$ is a multiplier system of weight $-k$.

The group $\mathrm{SL}_2(\mathbb{R})$ acts on $\mathcal{H}$ via $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) z = (az+b)/(cz+d)$. The cusps of Γ are those points in the extended upper half-plane $\mathcal{H}^*$ which are fixed by parabolic elements of Γ. Given a cusp $\mathfrak{a}$ of Γ let

$\Gamma_{\mathfrak{a}} := \{\gamma \in \Gamma : \gamma\mathfrak{a} = \mathfrak{a}\}$ denote its stabilizer in Γ, and let $\sigma_{\mathfrak{a}}$ denote any element of $\mathrm{SL}_2(\mathbb{R})$ satisfying $\sigma_{\mathfrak{a}}\infty = \mathfrak{a}$ and $\sigma_{\mathfrak{a}}^{-1}\Gamma_{\mathfrak{a}}\sigma_{\mathfrak{a}} = \Gamma_\infty$. Define $\kappa_{\mathfrak{a}} = \kappa_{\nu,\mathfrak{a}}$ by the conditions

$$\nu\left(\sigma_{\mathfrak{a}}\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}\sigma_{\mathfrak{a}}^{-1}\right) = e(-\kappa_{\mathfrak{a}}) \quad \text{and} \quad 0 \le \kappa_{\mathfrak{a}} < 1.$$

We say that $\mathfrak{a}$ is singular with respect to ν if ν is trivial on $\Gamma_{\mathfrak{a}}$, that is, if $\kappa_{\nu,\mathfrak{a}} = 0$. Note that if $\kappa_{\nu,\mathfrak{a}} > 0$ then

$$\kappa_{\bar{\nu},\mathfrak{a}} = 1 - \kappa_{\nu,\mathfrak{a}}.$$

We are primarily interested in the multiplier system ν_θ of weight $\frac{1}{2}$ (and its conjugate $\bar{\nu}_\theta = \nu_\theta^{-1}$ of weight $-\frac{1}{2}$) on $\Gamma_0(4)$ defined by

$$\theta(\gamma z) = \nu_\theta(\gamma)\sqrt{cz+d}\,\theta(z),$$

where

$$\theta(z) := \sum_{n\in\mathbb{Z}} e(n^2 z).$$

Explicitly, we have

$$\nu_\theta\left(\begin{pmatrix}* & *\\ c & d\end{pmatrix}\right) = \left(\frac{c}{d}\right)\varepsilon_d^{-1},$$

where $\left(\frac{\cdot}{\cdot}\right)$ is the extension of the Kronecker symbol given in [Shimura 1973], for example, and

$$\varepsilon_d = \left(\frac{-1}{d}\right)^{\frac{1}{2}} = \begin{cases} 1 & \text{if } d \equiv 1 \pmod 4,\\ i & \text{if } d \equiv 3 \pmod 4.\end{cases}$$

For $\gamma \in \mathrm{SL}_2(\mathbb{R})$ we define the weight k slash operator by

$$f\big|_k\gamma := j(\gamma, z)^{-k} f(\gamma z).$$

The weight k hyperbolic Laplacian

$$\Delta_k := y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) - iky\frac{\partial}{\partial x}$$

commutes with the weight k slash operator for every $\gamma \in \mathrm{SL}_2(\mathbb{R})$. A real analytic function $f : \mathcal{H} \to \mathbb{C}$ is an eigenfunction of Δ_k with eigenvalue λ if

$$\Delta_k f + \lambda f = 0. \tag{3-1}$$

If f satisfies (3-1) then for notational convenience we write

$$\lambda = \tfrac{1}{4} + r^2,$$

and we refer to r as the spectral parameter of f.

A function $f : \mathcal{H} \to \mathbb{C}$ is automorphic of weight k and multiplier ν for Γ if

$$f\big|_k\gamma = \nu(\gamma) f \qquad \text{for all } \gamma \in \Gamma.$$

Let $\mathcal{A}_k(N, \nu)$ denote the space of all such functions. A smooth automorphic function which is also an eigenfunction of Δ_k and which has at most polynomial growth at the cusps of Γ is called a Maass form. We let $\mathcal{A}_k(N, \nu, r)$ denote the vector space of Maass forms with spectral parameter r. Complex conjugation $f \to \bar{f}$ gives a bijection $\mathcal{A}_k(N, \nu, r) \longleftrightarrow \mathcal{A}_{-k}(N, \bar{\nu}, r)$.

If $f \in \mathcal{A}_k(n, \nu, r)$, then $f\big|_k \sigma_{\mathfrak{a}}$ satisfies

$$(f\big|_k \sigma_{\mathfrak{a}})(z+1) = e(-\kappa_{\mathfrak{a}})(f\big|_k \sigma_{\mathfrak{a}})(z).$$

For $n \in \mathbb{Z}$ define

$$n_{\mathfrak{a}} := n - \kappa_{\mathfrak{a}}.$$

Then f has a Fourier expansion at the cusp $\mathfrak{a}$ of the form

$$(f\big|_k \sigma_{\mathfrak{a}})(z) = \rho_{f,\mathfrak{a}}(0) y^{\frac{1}{2}+ir} + \rho'_{f,\mathfrak{a}}(0) y^{\frac{1}{2}-ir} + \sum_{n_{\mathfrak{a}} \neq 0} \rho_{f,\mathfrak{a}}(n) W_{\frac{1}{2}k \operatorname{sgn}(n), ir}(4\pi |n_{\mathfrak{a}}| y) e(n_{\mathfrak{a}} x),$$

where $W_{\kappa,\mu}(y)$ is the W-Whittaker function. When the weight is 0, many authors normalize the Fourier coefficients so that $\rho_{f,\mathfrak{a}}(n)$ is the coefficient of $\sqrt{y}\, K_{ir}(2\pi |n_{\mathfrak{a}}| y)$, where $K_\nu(y)$ is the K-Bessel function. Using the relation

$$W_{0,\mu}(y) = \frac{\sqrt{y}}{\sqrt{\pi}} K_\mu(y/2),$$

we see that this has the effect of multiplying $\rho_{f,\mathfrak{a}}(n)$ by $2|n_{\mathfrak{a}}|^{1/2}$.

Let $\mathcal{L}_k(\nu)$ denote the L^2-space of automorphic functions with respect to the Petersson inner product

$$\langle f, g \rangle := \int_{\Gamma \backslash \mathcal{H}} f(z) \overline{g(z)}\, d\mu, \qquad d\mu := \frac{dx\, dy}{y^2},$$

and let $\mathcal{L}_k(\nu, \lambda)$ denote the λ-eigenspace. The spectrum of Δ_k is real and contained in $[\lambda_0(k), \infty)$, where $\lambda_0(k) := \frac{1}{2}|k|\big(1 - \frac{1}{2}|k|\big)$. The minimal eigenvalue $\lambda_0(k)$ occurs if and only if there is a holomorphic modular form F of weight $|k|$ and multiplier ν, in which case

$$f_0(z) = \begin{cases} y^{k/2} F(z) & \text{if } k \geq 0, \\ y^{-k/2} \bar{F}(z) & \text{if } k < 0, \end{cases}$$

is the corresponding eigenfunction. When $k = \pm\frac{1}{2}$ and $\nu = \nu_\theta^{2k}$, the eigenspace $\mathcal{L}_k\big(\nu, \frac{3}{16}\big)$ is one-dimensional, spanned by $y^{\frac{1}{4}}\theta(z)$ if $k = \frac{1}{2}$ and $y^{-\frac{1}{4}}\bar{\theta}(z)$ if $k = -\frac{1}{2}$.

The spectrum of Δ_k on $\mathcal{L}_k(\nu)$ consists of an absolutely continuous spectrum of multiplicity equal to the number of singular cusps, and a discrete spectrum of finite multiplicity. The Eisenstein series, of which there is one for each singular cusp $\mathfrak{a}$, give rise to the continuous spectrum, which is bounded below by $\frac{1}{4}$. Let $\mathfrak{a}$ be a singular cusp. The Eisenstein series for the cusp $\mathfrak{a}$ is defined by

$$E_{\mathfrak{a}}(z, s) := \sum_{\gamma \in \Gamma_{\mathfrak{a}} \backslash \Gamma_\infty} \bar{\nu}(\gamma) \overline{w}(\sigma_{\mathfrak{a}}^{-1}, \gamma)\, j(\sigma_{\mathfrak{a}}^{-1}\gamma, z)^{-k} \operatorname{Im}(\sigma_{\mathfrak{a}}^{-1}\gamma z)^s.$$

If $\mathfrak{b}$ is any cusp, the Fourier expansion for $E_\mathfrak{a}$ at the cusp $\mathfrak{b}$ is given by

$$j(\sigma_\mathfrak{b}, z)^{-k} E_\mathfrak{a}(z, s) = \delta_{\mathfrak{a}=\mathfrak{b}} y^s + \delta_{\kappa_\mathfrak{b}=0} \phi_{\mathfrak{a}\mathfrak{b}}(0, s) y^{1-s} + \sum_{n_\mathfrak{b} \neq 0} \phi_{\mathfrak{a}\mathfrak{b}}(n, s) W_{\frac{1}{2}k \operatorname{sgn}(n), s-\frac{1}{2}}(4\pi |n_\mathfrak{b}| y) e(n_\mathfrak{b} x),$$

where

$$\phi_{\mathfrak{a}\mathfrak{b}}(n, s) = \begin{cases} \dfrac{e\left(-\frac{1}{4}k\right)\pi^s |n|^{s-1}}{\Gamma\left(s + \frac{1}{2}k \operatorname{sgn}(n)\right)} \displaystyle\sum_{c \in \mathcal{C}(\mathfrak{a}, \mathfrak{b})} \frac{S_{\mathfrak{a}\mathfrak{b}}(0, n, c, \nu)}{c^{2s}} & \text{if } n_\mathfrak{b} \neq 0, \\ \dfrac{e\left(-\frac{1}{4}k\right)\pi 4^{1-s}\Gamma(2s-1)}{\Gamma\left(s + \frac{1}{2}k\right)\Gamma\left(s - \frac{1}{2}k\right)} \displaystyle\sum_{c \in \mathcal{C}(\mathfrak{a}, \mathfrak{b})} \frac{S_{\mathfrak{a}\mathfrak{b}}(0, 0, c, \nu)}{c^{2s}} & \text{if } n_\mathfrak{b} = 0. \end{cases} \tag{3-2}$$

Here $\mathcal{C}(\mathfrak{a}, \mathfrak{b}) = \left\{c > 0 : \left(\begin{smallmatrix} * & * \\ c & * \end{smallmatrix}\right) \in \sigma_\mathfrak{a}^{-1} \Gamma \sigma_\mathfrak{b}\right\}$ is the set of allowed moduli and $S_{\mathfrak{a}\mathfrak{b}}(m, n, c, \nu)$ is the Kloosterman sum (defined for any cusp pair $\mathfrak{a}\mathfrak{b}$)

$$S_{\mathfrak{a}\mathfrak{b}}(m, n, c, \nu) := \sum_{\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_\infty \backslash \sigma_\mathfrak{a}^{-1} \Gamma \sigma_\mathfrak{b} / \Gamma_\infty} \bar{\nu}_{\mathfrak{a}\mathfrak{b}}(\gamma) e\left(\frac{m_\mathfrak{a} a + n_\mathfrak{b} d}{c}\right), \tag{3-3}$$

where

$$\nu_{\mathfrak{a}\mathfrak{b}}(\gamma) = \nu(\sigma_\mathfrak{a} \gamma \sigma_\mathfrak{b}^{-1}) \frac{w(\sigma_\mathfrak{a} \gamma \sigma_\mathfrak{b}^{-1}, \sigma_\mathfrak{b})}{w(\sigma_\mathfrak{a}, \gamma)}.$$

The coefficients $\phi_{\mathfrak{a}\mathfrak{b}}(n, s)$ can be meromorphically continued to the entire s-plane and, in particular, are well-defined on the line $\operatorname{Re}(s) = \frac{1}{2}$. In Section 5 we will evaluate certain linear combinations of the coefficients $\phi_{\mathfrak{a}\mathfrak{b}}\left(n, \frac{1}{2} \pm ir\right)$ in terms of Dirichlet L-functions in the cases $k = \pm\frac{1}{2}$ and $\nu = \nu_\theta^{2k}$.

Let $\mathcal{V}_k(\nu)$ denote the orthogonal complement in $\mathcal{L}_k(\nu)$ of the space generated by Eisenstein series. The spectrum of Δ_k on $\mathcal{V}_k(\nu)$ is countable and of finite multiplicity. The exceptional eigenvalues are those which lie in $\left(\lambda_0(k), \frac{1}{4}\right)$ (conjecturally, the set of exceptional eigenvalues is empty). The subspace $\mathcal{V}_k(\nu)$ consists of functions f which decay exponentially at every cusp; equivalently, the zeroth Fourier coefficient of f at each singular cusp vanishes. Eigenfunctions of Δ_k in $\mathcal{V}_k(\nu)$ are called Maass cusp forms.

Let $\{f_j\}$ be an orthonormal basis of $\mathcal{V}_k(\nu)$, and for each j let $\lambda_j = \frac{1}{4} + r_j^2$ denote the Laplace eigenvalue and $\{\rho_{j,\mathfrak{a}}(n)\}$ the Fourier coefficients. Weyl's law describes the distribution of the spectral parameters r_j. Theorem 2.28 of [Hejhal 1983] shows that

$$\sum_{0 \leq r_j \leq T} 1 - \frac{1}{4\pi} \int_{-T}^{T} \frac{\varphi'}{\varphi}\left(\frac{1}{2} + it\right) dt = \frac{\operatorname{vol}(\Gamma\backslash\mathcal{H})}{4\pi} T^2 - \frac{K_0}{\pi} T \log T + O(T),$$

where $\varphi(s)$ and K_0 are the determinant (see [Hejhal 1983, p. 298]) and dimension (see [Hejhal 1983, p. 281]), respectively, of the scattering matrix $\Phi(s)$ whose entries are given in terms of constant terms of Eisenstein series.

4. An estimate for coefficients of Maass cusp forms

In this section we prove a general theorem which applies to the Fourier coefficients at the cusp $\mathfrak{a}$ of weight $\pm\frac{1}{2}$ Maass cusp forms with multiplier ν for $\Gamma = \Gamma_0(N)$. We assume that the bound

$$\sum_{c>0} \frac{|S_{\mathfrak{a}\mathfrak{a}}(n,n,c,\nu)|}{c^{1+\beta}} \ll_\nu n^\varepsilon \tag{4-1}$$

holds for some $\beta = \beta_{\nu,\mathfrak{a}} \in \left(\frac{1}{2}, 1\right)$. A similar estimate was proved in [Ahlgren and Andersen 2018, Theorem 3.1], but the following theorem improves the bound in the x-aspect when $k = \frac{1}{2}$. The proof given here is also considerably shorter.

Theorem 4.1. *Suppose that $k = \pm\frac{1}{2}$ and that ν is a multiplier system of weight k which satisfies (4-1). Fix an orthonormal basis of cusp forms $\{u_j\}$ for $\mathcal{V}_k(\nu)$. For each j, let $\rho_{j,\mathfrak{a}}(n)$ denote the n-th Fourier coefficient of u_j at $\mathfrak{a}$ and let r_j denote the spectral parameter. Then for all $n \geq 1$ we have*

$$n_{\mathfrak{a}} \sum_{x \leq r_j \leq 2x} |\rho_{j,\mathfrak{a}}(n)|^2 e^{-\pi r_j} \ll x^{-k}\left(x^2 + n^{\beta+\varepsilon} x^{1-2\beta} \log^\beta x\right).$$

We begin with an auxiliary version of Kuznetsov's formula [1980, §5] which is Lemma 3 of [Proskurin 2003] with $m = n$, $t \mapsto 2t$, and $\sigma = 1$; see [Ahlgren and Andersen 2018, Section 3] for justification of the latter. While Proskurin assumed that $k > 0$, this lemma is still valid for $k < 0$ by the same proof, and straightforward modifications give the result for an arbitrary cusp $\mathfrak{a}$.

Lemma 4.2. *With the assumptions of Theorem 4.1, and for any $t \in \mathbb{R}$ we have*

$$\begin{aligned}\frac{2\pi^2 n_{\mathfrak{a}}}{\left|\Gamma\left(1-\frac{1}{2}k+it\right)\right|^2}\Bigg[&\sum_{r_j} \frac{|\rho_{j,\mathfrak{a}}(n)|^2}{\cosh 2\pi r_j + \cosh 2\pi t} \\ &+ \frac{1}{4}\sum_{\mathfrak{c}} \int_{-\infty}^{\infty} \frac{\left|\phi_{\mathfrak{c}\mathfrak{a}}\left(n, \frac{1}{2}+ir\right)\right|^2}{(\cosh 2\pi r + \cosh 2\pi t)\left|\Gamma\left(\frac{1}{2}(k+1)+ir\right)\right|^2}\, dr\Bigg] \\ &= \frac{1}{4\pi} + \frac{2n_{\mathfrak{a}}}{i^{k+1}} \sum_{c>0} \frac{S_{\mathfrak{a}\mathfrak{a}}(n,n,c,\nu)}{c^2} \int_L K_{2it}\left(\frac{4\pi n_{\mathfrak{a}}}{c}\, q\right) q^{k-1}\, dq, \end{aligned} \tag{4-2}$$

where $\sum_{\mathfrak{c}}$ is a sum over singular cusps, and L is the semicircular contour $|q| = 1$ with $\mathrm{Re}(q) > 0$, from $-i$ to i.

To prove Theorem 4.1 we follow the method of Motohashi [2003, Section 2]. We begin by evaluating the integral on the right-hand side of (4-2) via the following lemma. For the remainder of this section we frequently use the notation $\int_{(\xi)}$ to denote $\int_{\xi-i\infty}^{\xi+i\infty}$.

Lemma 4.3. *Let $k = \pm\frac{1}{2}$. Suppose that $a > 0$, $\xi > \frac{1}{2}k$. Then*

$$2\int_L K_{2it}(2aq) q^{k-1}\, dq = \frac{1}{2\pi} \int_{(\xi)} \frac{\sin\left(\pi s - \frac{1}{2}\pi k\right)}{s - \frac{1}{2}k} \Gamma(s+it)\Gamma(s-it) a^{-2s}\, ds.$$

Proof. For any $\xi > 0$ we have the Mellin–Barnes integral representation [DLMF 2010, (10.32.13)]

$$2K_{2it}(2z) = \frac{1}{2\pi i}\int_{(\xi)} \Gamma(s)\Gamma(s-2it)z^{2it-2s}\,ds,$$

which is valid for $|\arg z| < \frac{1}{2}\pi$. It follows that

$$\begin{aligned} 2\int_L K_{2it}(2aq)q^{k-1}\,dq &= \frac{1}{2\pi i}\int_{(\xi)} \Gamma(s)\Gamma(s-2it)a^{2it-2s}\int_L q^{2it-2s+k-1}\,dq\,ds \\ &= \frac{1}{2\pi}\int_{(\xi)} \Gamma(s)\Gamma(s-2it)a^{2it-2s}\frac{\sin\left(\pi\left(s-it-\frac{1}{2}k\right)\right)}{s-it-\frac{1}{2}k}\,ds. \end{aligned}$$

The lemma follows after replacing s by $s+it$. □

Let K be a large positive real number. In (4-2) we multiply by the positive weight

$$e^{-(t/K)^2} - e^{-(2t/K)^2}$$

and integrate on t over $\mathbb{R}$. Applying Lemma 4.3 to the result (and noting that all terms on the left-hand side are positive), we obtain

$$n_{\mathfrak{a}}\sum_{r_j}|a_j(n)|^2 h_K(r_j) \ll K + \sum_{c>0}\frac{|S(n,n,c,v)|}{c}\left|M_k\left(K,\frac{2\pi n_{\mathfrak{a}}}{c}\right)\right|, \tag{4-3}$$

where

$$h_K(r) := \int_{-\infty}^{\infty}\frac{e^{-(t/K)^2} - e^{-(2t/K)^2}}{\left|\Gamma\left(1-\frac{1}{2}k+it\right)\right|^2(\cosh 2\pi r + \cosh 2\pi t)}\,dt \tag{4-4}$$

and

$$M(K,a) = \int_{-\infty}^{\infty}\left(e^{-(t/K)^2} - e^{-(2t/K)^2}\right)\int_{(\xi)}\frac{\sin\left(\pi s - \frac{1}{2}\pi k\right)}{s-\frac{1}{2}k}\Gamma(s+it)\Gamma(s-it)a^{1-2s}\,ds\,dt.$$

We will make use of the following well-known estimate for oscillatory integrals; see, for instance, [Titchmarsh 1951, Chapter IV].

Lemma 4.4. *Suppose that F and G are real-valued functions on $[a,b]$ with F differentiable, such that $G(x)/F'(x)$ is monotonic. If $|F'(x)/G(x)| \geq m > 0$ then*

$$\int_a^b G(x)e(F(x))\,dx \ll \frac{1}{m}.$$

Proposition 4.5. *Let K be a large positive real number. Suppose that $k = \pm\frac{1}{2}$ and let $M(K,a)$ be as above. For $a > 0$ we have*

$$M(K,a) \ll \min\left(1, \frac{a\log K}{K^2}\right). \tag{4-5}$$

Proof. Starting with the integral representation [DLMF 2010, (5.12.1)]

$$\Gamma(s+it)\Gamma(s-it) = \Gamma(2s)\int_0^1 y^{s+it-1}(1-y)^{s-it-1}\,dy,$$

we interchange the order of integration, putting the integral on t inside, and find that the integral on t equals

$$\begin{aligned} T(K,y) &= \int_{-\infty}^{\infty}\left(\frac{y}{1-y}\right)^{it}\left(e^{-(t/K)^2} - e^{-(2t/K)^2}\right)dt \\ &= Ke^{-(1/4)K^2\log^2\left(\frac{y}{1-y}\right)} - \tfrac{1}{2}Ke^{-(1/16)K^2\log^2\left(\frac{y}{1-y}\right)}. \end{aligned}$$

Hence

$$M(K,a) = \int_0^1 \frac{T(K,y)}{y(1-y)}\int_{(\xi)}\frac{\sin\left(\pi s - \frac{1}{2}\pi k\right)}{s-\frac{1}{2}k}\Gamma(2s)[y(1-y)]^s a^{1-2s}\,ds\,dy. \tag{4-6}$$

To evaluate the inner integral, we use that

$$\frac{u^{k-2s}}{s-\frac{1}{2}k} = 2\int_u^{\infty} t^{-2s+k-1}\,dt.$$

Setting $u = a\,[y(1-y)]^{-1/2}$, the integral on s in (4-6) equals

$$2au^{-k}\int_u^{\infty} t^{k-1}\int_{(\xi)}\sin\left(\pi s - \tfrac{1}{2}\pi k\right)\Gamma(2s)t^{-2s}\,ds\,dt = af_k(u),$$

where

$$f_k(u) = \cos\left(\tfrac{1}{2}\pi k\right)u^{-k}\int_u^{\infty} t^{k-1}\sin t\,dt - \sin\left(\tfrac{1}{2}\pi k\right)u^{-k}\int_u^{\infty} t^{k-1}\cos t\,dt.$$

Finally, we set $z = K\log(y/(1-y))$ to obtain

$$M(K,a) = a\int_{-\infty}^{\infty}\left(e^{-z^2/4} - \frac{1}{2}e^{-z^2/16}\right)f_k\left(2a\cosh\left(\frac{z}{2K}\right)\right)dz.$$

We claim that $f_k(u) \ll \min(1, 1/u)$. For $u \geq 1$ this follows from Lemma 4.4. Suppose that $u \leq 1$. In the case $k = -\frac{1}{2}$, we have $f_k(u) \ll 1$ by estimating the integrals trivially. When $k = \frac{1}{2}$ a computation shows that

$$f_{\frac{1}{2}}(u) = \frac{\sqrt{\pi}\,C\left(\sqrt{2u/\pi}\right) - \sqrt{\pi}\,S\left(\sqrt{2u/\pi}\right)}{\sqrt{u}},$$

where $C(x)$ and $S(x)$ are the Fresnel integrals [DLMF 2010, §7.2]. It follows that $f_{\frac{1}{2}}(u) \ll 1$.

From the estimate $f_k(u) \ll \min(1, 1/u)$ it follows that

$$M(K,a) \ll 1.$$

Now suppose that $a \ll K^2$. In this case we add and subtract $f_k(2a)$ from the integrand and notice that

$$\int_{-\infty}^{\infty}\left(e^{-z^2/4} - \tfrac{1}{2}e^{-z^2/16}\right)dz = 0,$$

so

$$M(K,a) \ll a\int_0^\infty e^{-z^2/16}\left|f_k(2a) - f_k\left(2a\cosh\left(\frac{z}{2K}\right)\right)\right| dz.$$

Let $T = c\sqrt{\log K}$ with c a large constant, and let $F(z) = f_k(2a) - f_k(2a\cosh z)$. Then $F(0) = F'(0) = 0$, so for $|z| \le T/K$ we have

$$F(z) \ll z^2 \max_{|w|\le T/K} |F''(w)|. \tag{4-7}$$

Since

$$F''(w) \ll a\cosh w|f_k'(2a\cosh w)| + a^2\sinh^2 w|f_k''(2a\cosh w)|$$

and, by Lemma 4.4,

$$f_k'(u),\ f_k''(u) \ll u^{-1},$$

we conclude that

$$F''(w) \ll a\sinh(T/K)\tanh(T/K) \ll \frac{aT^2}{K^2} \ll T^2. \tag{4-8}$$

By (4-7) and (4-8) we have

$$a\int_0^T e^{-z^2/16}\left|F\left(\frac{z}{2K}\right)\right| dz \ll \frac{aT^2}{K^2}\int_0^\infty z^2 e^{-z^2/16}\,dz \ll \frac{aT^2}{K^2}$$

and by $f_k(u) \ll 1$ we have

$$a\int_T^\infty e^{-z^2/16}\left|f_k(2a) - f_k\left(2a\cosh\left(\frac{z}{2K}\right)\right)\right| dz \ll a\int_T^\infty e^{-z^2/16}\,dz \ll a\,e^{-T^2/16}.$$

With our choice of T this yields (4-5). □

Proof of Theorem 4.1. First note that when $r \sim x$ we have $h_x(r) \gg e^{-\pi r}x^{k-1}$, where $h_x(r)$ is defined in (4-4), so by (4-3) and positivity we have

$$n_{\mathfrak{a}}x^k \sum_{x\le r_j\le 2x} |\rho_{j,\mathfrak{a}}(n)|e^{-\pi r_j} \ll x^2 + x\sum_{c>0}\frac{|S_{\mathfrak{a}\mathfrak{a}}(n,n,c,\nu)|}{c}\left|M_k\left(x,\frac{2\pi n_{\mathfrak{a}}}{c}\right)\right|.$$

Let β be as in (4-1). By Proposition 4.5 we have

$$M_k(x,a) \ll \min\left(1, \frac{a\log x}{x^2}\right) \ll \frac{a^\beta \log^\beta x}{x^{2\beta}},$$

from which it follows that

$$\begin{aligned} x\sum_{c>0}\frac{|S_{\mathfrak{a}\mathfrak{a}}(n,n,c,\nu)|}{c}\left|M_k\left(x,\frac{2\pi n_{\mathfrak{a}}}{c}\right)\right| &\ll n_{\mathfrak{a}}^\beta x^{1-2\beta}\log^\beta x\sum_{c>0}\frac{|S_{\mathfrak{a}\mathfrak{a}}(n,n,c,\nu)|}{c^{1+\beta}} \\ &\ll n_{\mathfrak{a}}^{\beta+\varepsilon}x^{1-2\beta}\log^\beta x. \end{aligned}$$

The theorem follows. □

5. The Kuznetsov formula for Kohnen's plus space

In this section we define the plus spaces of holomorphic and Maass cusp forms, and we prove an analogue of Kuznetsov's formula relating the Kloosterman sums $S_k^+(m,n,c)$ to the Fourier coefficients of such forms. For the remainder of the paper we specialize to the case $\Gamma = \Gamma_0(4)$ with $(k,\nu) = \left(\frac{1}{2}, \nu_\theta\right)$ or $\left(-\frac{1}{2}, \bar{\nu}_\theta\right)$. We will often write $k = \lambda + \frac{1}{2}$, and to simplify notation, we write $\mathcal{V}_k = \mathcal{V}_k(\nu)$ and $\mathcal{S}_\ell = \mathcal{S}_\ell(\nu)$, where $S_\ell(\nu)$ is the space of holomorphic cusp forms of weight ℓ and multiplier ν. We fix once and for all a set of inequivalent representatives for the cusps of Γ, namely ∞, 0, and $\frac{1}{2}$, with associated scaling matrices

$$\sigma_\infty = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \sigma_0 = \begin{pmatrix} 0 & -\frac{1}{2} \\ 2 & 0 \end{pmatrix}, \qquad \sigma_{\frac{1}{2}} = \begin{pmatrix} 1 & -\frac{1}{2} \\ 2 & 0 \end{pmatrix}.$$

Then

$$\kappa_\infty = \kappa_0 = 0 \quad \text{and} \quad \kappa_{\frac{1}{2}} = \frac{(-1)^\lambda 3}{4}.$$

Following Kohnen [1980; 1982] we define an operator L on automorphic functions as follows. If f satisfies $f\big|_k\gamma = \nu(\gamma) f$ for all $\gamma \in \Gamma_0(4)$ then we define

$$Lf := \frac{1}{2(1+i^{2k})} \sum_{w=0}^{3} f\big|_k \begin{pmatrix} 1+w & \frac{1}{4} \\ 4w & 1 \end{pmatrix}.$$

It is not difficult to show that L maps Maass cusp forms to Maass cusp forms. It follows from [Kohnen 1980] (see also [Katok and Sarnak 1993]) that L is self-adjoint, that it commutes with the Hecke operators T_{p^2}, and that it satisfies the equation

$$(L-1)\left(L+\tfrac{1}{2}\right) = 0$$

(Kohnen proves this in the holomorphic case, but the necessary modifications are simple). The space $\mathcal{V}_k$ decomposes as $\mathcal{V}_k = \mathcal{V}_k^+ \oplus \mathcal{V}_k^-$ where $\mathcal{V}_k^+$ is the eigenspace with eigenvalue 1, and $\mathcal{V}_k^-$ is the eigenspace with eigenvalue $-\frac{1}{2}$. For each $f \in \mathcal{V}_k$, we have $f \in \mathcal{V}_k^+$ if and only if $\rho_{f,\infty}(n) = 0$ for $(-1)^\lambda n \equiv 2, 3 \pmod 4$. The following lemma describes the action of L on Fourier expansions.

Lemma 5.1. *Let $k = \pm\frac{1}{2} = \lambda + \frac{1}{2}$ and $\nu = \nu_\theta^{2k}$. Suppose that $f\big|_k\gamma = \nu(\gamma) f$ for all $\gamma \in \Gamma$. For each cusp $\mathfrak{a}$ of Γ write the Fourier expansion of f as*

$$(f\big|_k\sigma_\mathfrak{a})(z) = \sum_{n\in\mathbb{Z}} c_{f,\mathfrak{a}}(n,y) e(n_\mathfrak{a} x).$$

Then

$$c_{Lf,\infty}(n,y) = \begin{cases} \dfrac{1}{2} c_{f,\infty}(n,y) + \dfrac{1}{2(1-i^{2k})} c_{f,\mathfrak{a}}\left(\dfrac{n}{4} + \kappa_\mathfrak{a}, 4y\right) & \text{if } (-1)^\lambda n \equiv 0, 1 \pmod 4, \\ -\dfrac{1}{2} c_{f,\infty}(n,y) & \text{if } (-1)^\lambda n \equiv 2, 3 \pmod 4, \end{cases}$$

where $\mathfrak{a} = 0$ if $n \equiv 0 \pmod 4$ and $\mathfrak{a} = \frac{1}{2}$ if $n \equiv (-1)^\lambda \pmod 4$.

Proof. Let $A_w = \left(\begin{smallmatrix} 1+w & 1/4 \\ 4w & 1 \end{smallmatrix}\right)$. Since $A_2 = \left(\begin{smallmatrix} 3 & -2 \\ 8 & -5 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & 3/4 \\ 0 & 1 \end{smallmatrix}\right)$ and $\nu\left(\left(\begin{smallmatrix} 3 & -2 \\ 8 & -5 \end{smallmatrix}\right)\right) = i^{2k}$ we have

$$\begin{aligned} f\big|_k A_0 + f\big|_k A_2 &= f\big(z+\tfrac{1}{4}\big) + i^{2k} f\big(z+\tfrac{3}{4}\big) \\ &= (1+i^{2k}) \sum_{(-1)^\lambda n \equiv 0,1(4)} c_{f,\infty}(n,y) e(nx) - (1+i^{2k}) \sum_{(-1)^\lambda n \equiv 2,3(4)} c_{f,\infty}(n,y) e(nx). \end{aligned}$$

For $w = 1, 3$ we have $A_1 = \left(\begin{smallmatrix} -1 & 1 \\ -4 & 3 \end{smallmatrix}\right)\sigma_{\frac{1}{2}}\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right)$ and $A_3 = \left(\begin{smallmatrix} -1 & 1 \\ -4 & 3 \end{smallmatrix}\right)\sigma_0\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1/2 \end{smallmatrix}\right)$. Since $\nu\left(\left(\begin{smallmatrix} -1 & 1 \\ -4 & 3 \end{smallmatrix}\right)\right) = i^{2k}$, we have

$$\begin{aligned} f\big|_k A_1 + f\big|_k A_3 &= i^{2k}(f\big|_k \sigma_0)(4z) + i^{2k}(f\big|_k \sigma_{\frac{1}{2}}(4z)) \\ &= i^{2k} \sum_{n \equiv 0(4)} c_{f,0}\big(\tfrac{1}{4}n, 4y\big) e(nx) + i^{2k} \sum_{n \equiv (-1)^\lambda (4)} c_{f,\frac{1}{2}}\big(\tfrac{1}{4}n + \kappa_{\frac{1}{2}}, 4y\big) e(nx). \end{aligned}$$

The lemma follows. □

The analogue of L for holomorphic cusp forms is defined as follows. If for some ℓ we have $F(\gamma z) = \nu(\gamma)(cz+d)^\ell F(z)$ for all $\gamma \in \Gamma_0(4)$ then $f(z) = y^{\ell/2} F(z)$ satisfies $f\big|_\ell \gamma = \nu(\gamma) f$, and we define

$$L^* F := y^{-\ell/2} L f.$$

The plus space $\mathcal{S}_\ell^+$ of holomorphic cusp forms is defined as the subspace of $\mathcal{S}_\ell$ consisting of forms F satisfying $L^* F = F$. If $\rho_{F,\mathfrak{a}}(n)$ is the n-th coefficient of F at the cusp $\mathfrak{a}$ then, in the notation of the previous lemma, we have $\rho_{F,\mathfrak{a}}\big(\frac{1}{4}n + \kappa_\mathfrak{a}, 4y\big) = \frac{1}{2} c_{f,\mathfrak{a}}\big(\frac{1}{4}n + \kappa_\mathfrak{a}, 4y\big)$. Therefore we have the following analogue of Lemma 5.1.

Lemma 5.2. *Let $k = \pm\frac{1}{2} = \lambda + \frac{1}{2}$ and $\nu = \nu_\theta^{2k}$. Suppose that $\ell \equiv k \pmod 2$ and that $F \in \mathcal{S}_\ell(\nu)$. Then*

$$\rho_{L^*F,\infty}(n,y) = \begin{cases} \dfrac{1}{2}\rho_{F,\infty}(n,y) + \dfrac{1}{(1-i^{2k})}\,\rho_{F,\mathfrak{a}}\Big(\dfrac{n}{4} + \kappa_\mathfrak{a}, 4y\Big) & \text{if } (-1)^\lambda n \equiv 0, 1 \pmod 4, \\ -\dfrac{1}{2}\rho_{F,\infty}(n,y) & \text{if } (-1)^\lambda n \equiv 2, 3 \pmod 4, \end{cases}$$

where $\mathfrak{a} = 0$ if $n \equiv 0 \pmod 4$ and $\mathfrak{a} = \frac{1}{2}$ if $n \equiv (-1)^\lambda \pmod 4$.

To state the plus space version of the Kuznetsov trace formula, we first fix some notation. Recall that $S_k^+(m,n,c)$ is the plus space Kloosterman sum

$$S_k^+(m,n,c) = e\Big(-\frac{k}{4}\Big) \sum_{d \bmod c} \left(\frac{c}{d}\right) \varepsilon_d^{2k}\, e\left(\frac{m\bar{d} + nd}{c}\right) \times \begin{cases} 1 & \text{if } 8 \mid c, \\ 2 & \text{if } 4 \,||\, c. \end{cases}$$

Let $\varphi : [0,\infty) \to \mathbb{R}$ be a smooth test function which satisfies

$$\varphi(0) = \varphi'(0) = 0 \quad \text{and} \quad \varphi^{(j)}(x) \ll x^{-2-\varepsilon} \quad \text{for } j = 0, 1, 2, 3. \tag{5-1}$$

Define the integral transforms

$$\widetilde{\varphi}(\ell) := \frac{1}{\pi}\int_0^\infty J_{\ell-1}(x)\varphi(x)\,\frac{dx}{x}, \tag{5-2}$$

$$\widehat{\varphi}(r) := \frac{-i\,\xi_k(r)}{\cosh 2\pi r}\int_0^\infty \Big(\cos\big(\tfrac{1}{2}\pi k + \pi i r\big)J_{2ir}(x) - \cos\big(\tfrac{1}{2}\pi k - \pi i r\big)J_{-2ir}(x)\Big)\phi(x)\frac{dx}{x}, \tag{5-3}$$

where

$$\xi_k(r) := \frac{\pi^2}{\sinh \pi r\,\Gamma\big(\frac{1}{2}(1-k)+ir\big)\Gamma\big(\frac{1}{2}(1-k)-ir\big)} \sim \tfrac{1}{2}\pi r^k \quad \text{as } r\to\infty.$$

Note that $\widehat{\varphi}(r)$ is real-valued when $r \geq 0$ and when $ir \in \big(-\frac{1}{4}, \frac{1}{4}\big)$. If d is a fundamental discriminant, let $\chi_d = \big(\frac{d}{\cdot}\big)$ and let $L(s, \chi_d)$ denote the Dirichlet L-function with Dirichlet series

$$L(s,\chi_d) := \sum_{n=1}^\infty \frac{\chi_d(n)}{n^s}.$$

Finally, we define

$$\mathfrak{S}_d(w,s) = \sum_{\ell \mid w} \mu(\ell)\chi_d(\ell)\frac{\tau_s(w/\ell)}{\sqrt{\ell}},$$

where τ_s is the normalized sum of divisors function

$$\tau_s(\ell) = \sum_{ab=\ell}\left(\frac{a}{b}\right)^s = \frac{\sigma_{2s}(\ell)}{\ell^s}.$$

Theorem 5.3. *Let $\varphi : [0,\infty) \to \mathbb{R}$ be a smooth test function satisfying (5-1). Let $k = \pm\frac{1}{2} = \lambda + \frac{1}{2}$ and $v = v_\theta^{2k}$. Suppose that $m, n \geq 1$ with $(-1)^\lambda m, (-1)^\lambda n \equiv 0, 1 \pmod 4$, and write*

$$(-1)^\lambda m = v^2 d', \quad (-1)^\lambda n = w^2 d, \quad \text{with } d, d' \text{ fundamental discriminants.}$$

Fix an orthonormal basis of Maass cusp forms $\{u_j\} \subset \mathcal{V}_k^+$ with associated spectral parameters r_j and coefficients $\rho_j(n)$. For each $\ell \equiv k \pmod 2$ with $\ell > 2$, fix an orthonormal basis of holomorphic cusp forms $\mathcal{H}_\ell^+ \subset \mathcal{S}_\ell^+$ with normalized coefficients given by

$$g(z) = \sum_{n=1}^\infty (4\pi n)^{(\ell-1)/2}\rho_g(n)e(nz) \quad \text{for } g \in \mathcal{H}_\ell^+. \tag{5-4}$$

Then

$$\begin{aligned}\sum_{0<c\equiv 0(4)} \frac{S_k^+(m,n,c)}{c}\,\varphi\left(\frac{4\pi\sqrt{mn}}{c}\right)&\\ = 6\sqrt{mn}\sum_{j\geq 0}\frac{\overline{\rho_j(m)}\rho_j(n)}{\cosh \pi r_j}\widehat{\varphi}(r_j) + \frac{3}{2}\sum_{\ell\equiv k \bmod 2} e\big(\tfrac{1}{4}(\ell-k)\big)\widetilde{\varphi}(\ell)\Gamma(\ell)\sum_{g\in\mathcal{H}_\ell^+}\overline{\rho_g(m)}\rho_g(n)&\\ + \frac{1}{2}\int_{-\infty}^\infty \left(\frac{d}{d'}\right)^{ir}\frac{L\big(\frac{1}{2}-2ir,\chi_{d'}\big)L\big(\frac{1}{2}+2ir,\chi_d\big)\mathfrak{S}_{d'}(v,2ir)\mathfrak{S}_d(w,2ir)}{|\zeta(1+4ir)|^2\cosh\pi r\,\big|\Gamma\big(\frac{1}{2}(k+1)+ir\big)\big|^2}\,\widehat{\varphi}(r)\,dr.&\end{aligned}$$

Biró [2000, Theorem B] stated a version of Theorem 5.3 for $\Gamma_0(4N)$ in the case $k=\frac{1}{2}$ under the added assumption that $\widetilde{\varphi}(\ell)=0$ for all ℓ. His theorem involves coefficients of half-integral weight Eisenstein series at cusps instead of Dirichlet L-functions.

To prove Theorem 5.3, we start with Proskurin's version [2003] of the Kuznetsov formula which is valid for arbitrary weight k and for the cusp-pair $\infty\infty$. The necessary modifications for an arbitrary cusp-pair are straightforward; see [Deshouillers and Iwaniec 1982] for details in the $k=0$ case. Recall the definitions of the generalized Kloosterman sum $S_{\mathfrak{ab}}(m,n,c,\nu)$ in (3-3) and the Eisenstein series coefficients $\phi_{\mathfrak{ab}}(m,s)$ in (3-2).

Proposition 5.4. *Suppose that φ satisfies (5-1). Suppose that $m,n\geq 1$ and that $k=\pm\frac{1}{2}$. Let $\nu=\nu_\theta^{2k}$ and $\Gamma=\Gamma_0(4)$ and let $\mathfrak{a}$, $\mathfrak{b}$ be cusps of Γ. Let $\{u_j\}$ denote an orthonormal basis of Maass cusp forms of weight k with spectral parameters r_j. For each $2<\ell\equiv k \pmod 2$, let $\mathcal{H}_\ell$ denote an orthonormal basis of holomorphic cusp forms of weight ℓ with coefficients normalized as in (5-4). Then*

$$\begin{aligned}
&e\big(-\tfrac{1}{4}k\big)\sum_{c\in\mathcal{C}(\mathfrak{a},\mathfrak{b})}\frac{S_{\mathfrak{ab}}(m,n,c,\nu)}{c}\varphi\bigg(\frac{4\pi\sqrt{m_{\mathfrak{a}}n_{\mathfrak{b}}}}{c}\bigg)\\
&\quad=4\sqrt{m_{\mathfrak{a}}n_{\mathfrak{b}}}\sum_{j\geq 0}\frac{\overline{\rho_{j\mathfrak{a}}(m)}\rho_{j\mathfrak{b}}(n)}{\cosh\pi r_j}\widehat{\varphi}(r_j)+\sum_{\ell\equiv k \bmod 2}e\big(\tfrac{1}{4}(\ell-k)\big)\widetilde{\varphi}(\ell)\Gamma(\ell)\sum_{g\in\mathcal{H}_\ell}\overline{\rho_{g\mathfrak{a}}(m)}\rho_{g\mathfrak{b}}(n)\\
&\qquad+\sum_{\mathfrak{c}\in\{0,\infty\}}\int_{-\infty}^{\infty}\bigg(\frac{n_{\mathfrak{b}}}{m_{\mathfrak{a}}}\bigg)^{ir}\frac{\overline{\phi_{\mathfrak{ca}}\big(m,\frac{1}{2}+ir\big)}\,\phi_{\mathfrak{cb}}\big(n,\frac{1}{2}+ir\big)}{\cosh\pi r\big|\Gamma\big(\frac{1}{2}(k+1)+ir\big)\big|^2}\widehat{\varphi}(r)\,dr. \qquad (5\text{-}5)
\end{aligned}$$

We will apply (5-5) for the cusp-pairs $\infty\infty$, $\infty 0$, and $\infty\frac{1}{2}$, and take a certain linear combination which annihilates all but the plus space coefficients. The following lemma is essential to make this work.

Lemma 5.5. *Suppose that $4\,||\,c$. Let $k=\pm\frac{1}{2}=\lambda+\frac{1}{2}$ and $\nu=\nu_\theta^{2k}$. Let $\mathfrak{a}=0$ or $\frac{1}{2}$ according to $(-1)^\lambda n\equiv 0,1 \pmod 4$, respectively. Then*

$$S_{\infty\infty}(m,n,c,\nu)=(1+i^{2k})S_{\infty\mathfrak{a}}\big(m,\tfrac{1}{4}n+\kappa_{\mathfrak{a}},\tfrac{1}{2}c,\nu\big).$$

Proof. Since $\overline{S_{\mathfrak{ab}}(m,n,c,\nu)}=S_{\mathfrak{ab}}(-m,-n,c,\bar{\nu})$, it is enough to show that

$$S_{\infty\infty}(m,n,c,\nu_\theta)=(1+i)\times\begin{cases}S_{\infty 0}\big(m,\frac{1}{4}n,\frac{1}{2}c,\nu_\theta\big) & \text{if } n\equiv 0 \pmod 4,\\ S_{\infty\frac{1}{2}}\big(m,\frac{1}{4}(n+3),\frac{1}{2}c,\nu_\theta\big) & \text{if } n\equiv 1 \pmod 4.\end{cases}$$

This is proved in [Biró 2000, Lemma A.7]. Note that Biro chooses different representatives and scaling matrices for the cusps 0 and $\frac{1}{2}$, which has the effect of changing the factor $(1-i)$ to $(1+i)$. □

Proof of Theorem 5.3. Let k, ν, and $\mathfrak{a}$ be as in Lemma 5.5. From that lemma and the definition (1-9) it follows that

$$S_k^+(m,n,c)=e\big(-\tfrac{1}{4}k\big)S_{\infty\infty}(m,n,c,\nu)+\delta_{4||c}\sqrt{2}\,S_{\infty\mathfrak{a}}\big(m,\tfrac{1}{4}n+\kappa_{\mathfrak{a}},\tfrac{1}{2}c,\nu\big).$$

Therefore

$$\sum_{4|c>0} \frac{S_k^+(m,n,c)}{c}\,\varphi\!\left(\frac{4\pi\sqrt{mn}}{c}\right) = e\!\left(-\tfrac{1}{4}k\right)\sum_{4|c>0}\frac{S_{\infty\infty}(m,n,c,\nu)}{c}\varphi\!\left(\frac{4\pi\sqrt{mn}}{c}\right) + \frac{1}{\sqrt{2}}\sum_{2\|c>0}\frac{S_{\infty\mathfrak{a}}\big(m,\,\frac{1}{4}n+\kappa_{\mathfrak{a}},\,c,\,\nu\big)}{c}\,\varphi\!\left(\frac{4\pi\sqrt{m\left(\frac{1}{4}n+\kappa_{\mathfrak{a}}\right)_{\mathfrak{a}}}}{c}\right). \tag{5-6}$$

Note that $\mathcal{C}(\infty,\mathfrak{a}) = \{c\in\mathbb{Z}_+ : c\equiv 2 \pmod 4\}$ for $\mathfrak{a}=0,\frac{1}{2}$. We apply Proposition 5.4 for each of the cusp-pairs $\infty\infty$ and $\infty\mathfrak{a}$ on the right-hand side of (5-6). We fix an orthonormal basis $\{u_j^+\}$ for $\mathcal{V}_k^+$ and we choose an orthonormal basis $\{u_j\}$ for $\mathcal{V}_k$ such that $\{u_j^+\}\subseteq\{u_j\}$. Then we do the same for $\mathcal{H}_k^+\subseteq\mathcal{H}_k$. The Maass form contribution is

$$4\sqrt{mn}\sum_{u_j\in\mathcal{V}_k}\frac{\bar{\rho}_j(m)}{\cosh\pi r_j}\widehat{\varphi}(r_j)\left(\rho_{j,\infty}(n)+\frac{1}{2(1-i^{2k})}\rho_{j,\mathfrak{a}}\big(\tfrac{1}{4}n+\kappa_{\mathfrak{a}}\big)\right).$$

Let $\rho_j^{(L)}$ denote the coefficients of Lu_j. Then by Lemma 5.1 we have

$$\rho_{j,\infty}(n)+\frac{1}{2(1-i^{2k})}\rho_{j,\mathfrak{a}}\big(\tfrac{1}{4}n+\kappa_{\mathfrak{a}}\big) = \tfrac{1}{2}\rho_{j,\infty}(n)+\rho_{j,\infty}^{(L)}(n) = \rho_j(n)\times\begin{cases}\frac{3}{2} & \text{if } u_j\in\mathcal{V}_k^+,\\ 0 & \text{if } u_j\in\mathcal{V}_k^-.\end{cases}$$

We compute the contribution from the holomorphic forms similarly. For the Eisenstein series contribution we apply the following proposition, together with the relation $S_{\mathfrak{a}\mathfrak{b}}(m,n,c,\nu)=\overline{S_{\mathfrak{a}\mathfrak{b}}(-m,-n,c,\bar{\nu})}$. □

Proposition 5.6. *Let $k=\frac{1}{2}$ and $\nu=\nu_\theta$ and suppose that $m,n\equiv 0,1 \pmod 4$. Write $m=v^2d'$ and $n=w^2d$, where d',d are fundamental discriminants. Let $\mathfrak{a}=0$ or $\frac{1}{2}$ according to $n\equiv 0,1 \pmod 4$, respectively. Then*

$$\sum_{\mathfrak{c}\in\{\infty,0\}}\overline{\phi}_{\mathfrak{c}\infty}\big(m,\tfrac{1}{2}+ir\big)\left(\phi_{\mathfrak{c}\infty}\big(n,\tfrac{1}{2}+ir\big)+\frac{1+i}{2\cdot 4^{ir}}\phi_{\mathfrak{c}\mathfrak{a}}\big(\tfrac{1}{4}n+\kappa_{\mathfrak{a}},\tfrac{1}{2}+ir\big)\right) = \frac{L\big(\frac{1}{2}-2ir,\chi_{d'}\big)L\big(\frac{1}{2}+2ir,\chi_d\big)}{2|\zeta(1+4ir)|^2}\left(\frac{v}{w}\right)^{2ir}\mathfrak{S}_{d'}(v,2ir)\mathfrak{S}_d(w,2ir). \tag{5-7}$$

The proof of this proposition is quite technical, and we will proceed in several steps. In order to work in the region of absolute convergence, we will evaluate the sum

$$\sum_{\mathfrak{c}\in\{\infty,0\}}\overline{\phi}_{\mathfrak{c}\infty}(m,s)\Big(\phi_{\mathfrak{c}\infty}(n,s)+\frac{1+i}{4^s}\phi_{\mathfrak{c}\mathfrak{a}}\big(\tfrac{1}{4}n+\kappa_{\mathfrak{a}},s\big)\Big),$$

for $\mathrm{Re}(s)$ sufficiently large. Then, by analytic continuation, we can set $s=\frac{1}{2}+ir$ to obtain (5-7). First, for the term $\mathfrak{c}=\infty$, by Lemma 5.5 we have

$$\phi_{\infty\infty}(n,s)+\frac{1+i}{4^s}\phi_{\infty\mathfrak{a}}\big(\tfrac{1}{4}n+\kappa_{\mathfrak{a}},s\big) = e\big(\tfrac{1}{8}\big)\phi^+(n,s), \tag{5-8}$$

where

$$\phi^+(n,s)=\sum_{4|c>0}\frac{S^+(0,n,c)}{c^{2s}}. \tag{5-9}$$

Here we have written $S^+(m,n,c) = S^+_{1/2}(m,n,c)$ for convenience. The following proposition evaluates $\phi^+(n,s)$. It is proved in [Ibukiyama and Saito 2012] and applied in [Duke et al. 2011, Lemma 4]; here we give an alternative proof which uses Kohnen's identity (2-4).

Proposition 5.7. *Let $w \in \mathbb{Z}_+$ and let d be a fundamental discriminant. Then*

$$\phi^+(w^2 d, s) = 2^{\frac{3}{2}-4s} w^{1-2s} \frac{L\left(2s-\frac{1}{2}, \chi_d\right)}{\zeta(4s-1)} \mathfrak{S}_d(w, 2s-1).$$

Proof. By Möbius inversion, it suffices to prove that

$$\sum_{\ell \mid w} \chi_d(\ell) \ell^{\frac{1}{2}-2s} \phi^+\left(\frac{w^2}{\ell^2} d, s\right) = 2^{\frac{3}{2}-4s} w^{1-2s} \tau_{2s-1}(w) \frac{L\left(2s-\frac{1}{2}, \chi_d\right)}{\zeta(4s-1)}.$$

Writing ϕ^+ as the Dirichlet series (5-9), reversing the order of summation, and applying the identity (2-4), we find that

$$\sum_{\ell \mid w} \chi_d(\ell) \ell^{\frac{1}{2}-2s} \phi^+\left(\frac{w^2}{\ell^2} d, s\right) = \frac{1}{\sqrt{2}} \sum_{4 \mid c > 0} \frac{1}{c^{2s-\frac{1}{2}}} \sum_{\ell \mid (w, \frac{c}{4})} \chi_d(\ell) \sqrt{\frac{2\ell}{c}}\, S^+\left(0, \frac{w^2}{\ell^2} d; \frac{c}{\ell}\right)$$

$$= 2^{\frac{1}{2}-4s} \sum_{c=1}^{\infty} \frac{T_w(0,d;4c)}{c^{2s-\frac{1}{2}}}.$$

To evaluate $T_w(0,d;4c)$ for a given c, we write $4c = tu$, where

$$u = \prod_{p^a \| 4c} p^{\lceil a/2 \rceil} \quad \text{and} \quad t = \prod_{p^a \| 4c} p^{\lfloor a/2 \rfloor}.$$

Then $b^2 \equiv 0 \pmod{4c}$ if and only if $b = xu$ for some x modulo t. For each such b, let $g = \left(x, \frac{1}{2}t\right)$ and choose $\lambda \in \mathbb{Z}$ such that

$$\gamma = \begin{pmatrix} t/2g & x/g \\ \lambda & \frac{1+\lambda x/g}{t/2g} \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}).$$

Then $\gamma[c, b, b^2/4c] = [ug^2/t, 0, 0]$ and $\chi_d([c, b, b^2/4c]) = \chi_d(ug^2/t)$. It follows that

$$T_w(0,d;4c) = 2\chi_d(u/t) \sum_{\substack{x \bmod t/2 \\ (x, t/2, d) = 1}} e\left(\frac{mx}{t/2}\right) =: 2f(c).$$

It is straightforward to verify that $f(c)$ is a multiplicative function and that for each prime p

(i) if $p \mid d$ then

$$f(p^a) = \begin{cases} c_{p^{(a/2)}}(w) & \text{if } a \text{ is even,} \\ 0 & \text{if } a \text{ is odd,} \end{cases}$$

(ii) if $p \nmid d$ then

$$f(p^a) = \chi_d(p)^a \times \begin{cases} p^{\lfloor a/2 \rfloor} & \text{if } p^{\lfloor a/2 \rfloor} \mid w, \\ 0 & \text{otherwise.} \end{cases}$$

Here $c_q(w)$ is the Ramanujan sum which satisfies

$$\frac{w^{1-s}\sigma_{s-1}(w)}{\zeta(s)} = \sum_{q=1}^{\infty} \frac{c_q(w)}{q^s} = \prod_p \sum_{a=0}^{\infty} \frac{c_{p^a}(w)}{p^s}.$$

It follows that

$$\sum_{c=1}^{\infty} \frac{f(c)}{c^{2s-\frac{1}{2}}} = w^{2-4s}\sigma_{4s-2}(w)\frac{L\left(2s-\frac{1}{2}, \chi_d\right)}{\zeta(4s-1)}.$$

The proposition follows. □

Next we evaluate the term in (5-7) corresponding to the cusp $\mathfrak{c} = 0$. The following lemma will be useful.

Lemma 5.8. *Let $k = \frac{1}{2}$ and $v = v_\theta$ and suppose that $n \equiv 0, 1 \pmod 4$. Suppose that $4 \mid c$ and $\mathfrak{a} = 0$ or $2 \,||\, c$ and $\mathfrak{a} = \frac{1}{2}$ according to whether $n \equiv 0$ or $1 \pmod 4$, respectively. Then*

$$S_{0\mathfrak{a}}\left(0, \tfrac{1}{4}n + \kappa_{\mathfrak{a}}, c, v_\theta\right) = \tfrac{1}{4}S_{\infty\infty}(0, n, 4c, v_\theta). \tag{5-10}$$

Proof. For each cusp $\mathfrak{a}$ we have $\left(\frac{1}{4}n + \kappa_{\mathfrak{a}}\right)_{\mathfrak{a}} = \frac{1}{4}n$. Suppose first that $n \equiv 0 \pmod 4$ and $\mathfrak{a} = 0$. A straightforward computation shows that $S_{00}(m, n, c, v) = S_{\infty\infty}(m, n, c, v)$ for all $m, n \in \mathbb{Z}$. From the definition of $S_{\infty\infty}(m, n, c, v)$ it follows that, for $c \equiv 0 \pmod 4$, we have $S_{00}\left(0, \frac{1}{4}n, c, v\right) = \frac{1}{4}S_{\infty\infty}(0, n, 4c, v)$.

Now suppose that $n \equiv 1 \pmod 4$ and $\mathfrak{a} = \frac{1}{2}$. We will prove (5-10) directly from the definition of $S_{0\frac{1}{2}}(m, n, c, v)$. Let $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \sigma_0^{-1}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)\sigma_{\frac{1}{2}}$, where $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \Gamma_0(4)$. Then $2 \,||\, c$ and a, d are odd, so (after shifting by $\left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right)$ on the right) we can assume that $4 \mid b$. Then $\varepsilon_D = \varepsilon_{a+2b} = \varepsilon_a = \varepsilon_d$ since $ad \equiv 1 \pmod 4$. We also have

$$\left(\frac{C}{D}\right) = \left(\frac{-4b}{a+2b}\right) = \left(\frac{2a}{a+2b}\right) = (-1)^{\frac{1}{2}(a-1)}\left(\frac{2}{a}\right)\left(\frac{2b}{a}\right) = \left(\frac{4c}{a}\right) = \left(\frac{4c}{d}\right)$$

since $bc \equiv -1 \pmod a$ and $ad \equiv 1 \pmod{4c}$. It follows that

$$S_{0\mathfrak{a}}\left(0, \frac{n}{4} + \kappa_{\mathfrak{a}}, c, v_\theta\right) = \sum_{d \bmod c} \left(\frac{4c}{d}\right)\varepsilon_d\, e\left(\frac{nd}{4c}\right).$$

Note that replacing d by $d + c$ has no net effect since $\varepsilon_{d+c} = \varepsilon_{-d}$ and $\left(\frac{4c}{d+c}\right) = -\left(\frac{-1}{d}\right)\left(\frac{c}{d}\right)$, so

$$\left(\frac{4c}{d+c}\right)\varepsilon_{d+c}\, e\left(\frac{n(d+c)}{4c}\right) = \left(\frac{4c}{d}\right)e\left(\frac{nd}{4c}\right)\left[-\varepsilon_{-d}\left(\frac{-1}{d}\right)e\left(\frac{n}{4}\right)\right] = \left(\frac{4c}{d}\right)\varepsilon_d\, e\left(\frac{nd}{4c}\right)$$

since $n \equiv 1 \pmod 4$. The relation (5-10) follows. □

Proposition 5.9. *Let $k = \frac{1}{2}$ and $v = v_\theta$ and suppose that $n \equiv 0, 1 \pmod 4$. Write $n = w^2 d$ with d a fundamental discriminant. Let $\mathfrak{a} = 0$ or $\frac{1}{2}$ according to $n \equiv 0, 1 \pmod 4$, respectively. Then*

$$\phi_{0\infty}(n, s) + \frac{1+i}{4^s}\,\phi_{0\mathfrak{a}}\left(\frac{n}{4} + \kappa_{\mathfrak{a}}, s\right) = \frac{i}{4^s}w^{1-2s}\frac{L\left(2s-\frac{1}{2}, \chi_d\right)}{\zeta(4s-1)}\mathfrak{S}_d(w, 2s-1). \tag{5-11}$$

Proof. We will prove that

$$\phi_{0\infty}(n,s)+\frac{1+i}{4^s}\,\phi_{0\mathfrak{a}}\Big(\frac{n}{4}+\kappa_{\mathfrak{a}},s\Big)=i\cdot 2^{2s-\frac{3}{2}}\phi^{+}(n,s);$$

then (5-11) will follow from Proposition 5.7. A straightforward computation gives the relation

$$S_{0\infty}(m,n,c,\nu_\theta)=iS_{\infty 0}(m,n,c,\nu_\theta).$$

This, together with Lemma 5.5, shows that

$$\phi_{0\infty}(n,s)=\frac{2^{2s}}{1-i}\sum_{4||c>0}\frac{S_{\infty\infty}(0,n,c,\nu_\theta)}{c^{2s}}. \tag{5-12}$$

Next, by Lemma 5.8 we find that

$$\frac{1+i}{4^s}\,\phi_{0\mathfrak{a}}\Big(\frac{n}{4}+\kappa_{\mathfrak{a}},s\Big)=\frac{2^{2s}}{2(1-i)}\sum\frac{S_{\infty\infty}(0,n,c,\nu_\theta)}{c^{2s}}, \tag{5-13}$$

where the sum is over $c\equiv 0 \pmod{16}$ if $\mathfrak{a}=0$, or $c\equiv 8 \pmod{16}$ if $\mathfrak{a}=\frac{1}{2}$. We claim that we can let the sum run over all $c\equiv 0 \pmod 8$ in either case. Equivalently,

$$S_{\infty\infty}(0,n,c,\nu_\theta)=0 \quad \text{when} \quad \begin{cases} c\equiv 8 \pmod{16} & \text{if } n\equiv 0 \pmod 4,\\ c\equiv 0 \pmod{16} & \text{if } n\equiv 1 \pmod 4. \end{cases} \tag{5-14}$$

To see this, we decompose the Kloosterman sum as follows (see Lemma 1 of [Sturm 1980]): if $c=2^t c'$ with c' odd, then

$$S_{\infty\infty}(0,n,c,\nu_\theta)=\varepsilon_{c'}^{-1}G(n,c')\sum_{r \bmod 2^t}\Big(\frac{2^t}{r}\Big)\varepsilon_r e\Big(\frac{nr}{2^t}\Big),$$

where $G(n,c')$ is a Gauss sum. In the case $n\equiv 0 \pmod 4$ and $c\equiv 8 \pmod{16}$ it is easy to see that

$$\sum_{r \bmod 8}\Big(\frac{8}{r}\Big)\varepsilon_r e\Big(\frac{nr}{8}\Big)=0.$$

If $n\equiv 1 \pmod 4$ then, by replacing r by $r+2^{t-2}$, we see that

$$\sum_{r \bmod 2^t}\Big(\frac{2^t}{r}\Big)\varepsilon_r e\Big(\frac{nr}{2^t}\Big)=e\Big(\frac{n}{4}\Big)\sum_{r \bmod 2^t}\Big(\frac{2^t}{r}\Big)\varepsilon_r e\Big(\frac{nr}{2^t}\Big)$$

as long as $t\geq 4$, from which it follows that the sum modulo 2^t is zero.

By (5-12), (5-13), and (5-14), we conclude that

$$\begin{aligned}\phi_{0\infty}(n,s)+\frac{1+i}{4^s}\,\phi_{0\mathfrak{a}}\Big(\frac{n}{4}+\kappa_{\mathfrak{a}},s\Big)&=\frac{2^{2s-1}}{1-i}\Big(2\sum_{4||c>0}\frac{S_{\infty\infty}(0,n,c,\nu_\theta)}{c^{2s}}+\sum_{8|c>0}\frac{S_{\infty\infty}(0,n,c,\nu_\theta)}{c^{2s}}\Big)\\&=i\cdot 2^{2s-\frac{3}{2}}\phi^{+}(n,s),\end{aligned}$$

which completes the proof of the proposition. □

Proof of Proposition 5.6. By (5-8) and Propositions 5.7 and 5.9 we have

$$\sum_{\mathfrak{c}\in\{\infty,0\}} \overline{\phi}_{\mathfrak{c}\infty}(m,s)\Big(\phi_{\mathfrak{c}\infty}(n,s)+\frac{1+i}{4^s}\phi_{\mathfrak{c}\mathfrak{a}}\Big(\frac{n}{4}+\kappa_{\mathfrak{a}},s\Big)\Big)$$
$$=\Big(e\big(\tfrac{1}{8}\big)2^{\frac{3}{2}-4s}\overline{\phi}_{\infty\infty}(m,s)+i\cdot 2^{-2s}\overline{\phi}_{0\infty}(m,s)\Big)w^{1-2s}\frac{L\big(2s-\frac{1}{2},\chi_d\big)}{\zeta(4s-1)}\mathfrak{S}(w^2d,2s-1).$$

Then by (5-12) we have (writing $s=\sigma+ir$)

$$\begin{aligned}
&e\big(\tfrac{1}{8}\big)2^{\frac{3}{2}-4s}\overline{\phi}_{\infty\infty}(m,s)+i\cdot 2^{-2s}\overline{\phi}_{0\infty}(m,s)\\
&\quad=(1+i)2^{1-4s}\sum_{4|c>0}\frac{S_{\infty\infty}(0,m,c,\nu_\theta)}{c^{2\bar{s}}}+\frac{4^{\bar{s}-s}}{1-i}\sum_{4||c>0}\frac{S_{\infty\infty}(0,m,c,\nu_\theta)}{c^{2\bar{s}}}\\
&\quad=\frac{2^{-4ir}}{1-i}\big(\phi^+(m,\bar{s})+(4^{1-2\sigma}-1)\phi_{\infty\infty}(m,\bar{s})\big).
\end{aligned}$$

The proposition follows after applying Proposition 5.7 and setting $s=\frac{1}{2}+ir$, noting that the factor $4^{1-2\sigma}-1$ in the second term vanishes. □

6. Proof of Theorem 1.3

Let $a=4\pi\sqrt{mn}$ and $x>0$ and let $x^{\frac{1}{3}}\ll T\ll x^{\frac{2}{3}}$ be a free parameter to be chosen later. We choose a test function $\varphi=\varphi_{a,x,T}:[0,\infty)\to[0,1]$ satisfying

(i) $\varphi(t)=1$ for $\dfrac{a}{2x}\le t\le\dfrac{a}{x}$,

(ii) $\varphi(t)=0$ for $t\le\dfrac{a}{2x+2T}$ and $t\ge\dfrac{a}{x-T}$,

(iii) $\varphi'(t)\ll\Big(\dfrac{a}{x-T}-\dfrac{a}{x}\Big)^{-1}\ll\dfrac{x^2}{aT}$, and

(iv) φ and φ' are piecewise monotonic on a fixed number of intervals (whose number is independent of a,x,T).

We apply the plus space Kuznetsov formula in Theorem 5.3 with this test function and we estimate each of the terms on the right-hand side.

We begin by estimating the contribution from the holomorphic cusp forms

$$\mathcal{K}^h:=\sum_{\ell\equiv k \bmod 2}e\big(\tfrac{1}{4}(\ell-k)\big)\widetilde{\varphi}(\ell)\Gamma(\ell)\sum_{g\in\mathcal{H}_\ell^+}\overline{\rho_g(m)}\rho_g(n). \tag{6-1}$$

Since the operator L commutes with the Hecke operators we may assume that the orthonormal basis $\mathcal{H}_\ell^+$ is also a basis consisting of Hecke eigenforms. We will estimate $\mathcal{K}^h$ by applying the Kohnen–Zagier formula [1981] and Young's hybrid subconvexity bound [2017]. Let $g\in\mathcal{H}_\ell^+$ and recall that the coefficients of g are normalized so that

$$g(z)=\sum_{n=1}^{\infty}(4\pi n)^{\frac{1}{2}(\ell-1)}\rho_g(n)e(nz).$$

Since we are working in the plus space, the Shimura correspondence is an isomorphism between $\mathcal{S}_\ell^+(\nu)$ and the space $\mathcal{S}_{2\ell-1}$ of (even) weight $2\ell-1$ cusp forms on Γ_1. So g lifts to a unique normalized $f \in S_{2\ell-1}$ with Fourier expansion

$$f(z) = \sum_{n=1}^{\infty} n^{\ell-1} a_f(n) e(nz), \qquad \text{where } a_f(1) = 1.$$

The coefficients ρ_g and a_f are related via

$$\rho_g(v^2|d|) = \rho_g(|d|) \sum_{u|v} \mu(u) \Big(\frac{d}{u}\Big) u^{-\frac{1}{2}} a_f(v/u),$$

where d is a fundamental discriminant with $(-1)^\lambda d > 0$. Using Deligne's bound $|a_f(n)| \le \sigma_0(n)$, it follows that

$$|\rho_g(v^2|d|)| \le |\rho_g(|d|)| \sigma_0^2(v). \tag{6-2}$$

Suppose that g is normalized so that $\langle g, g\rangle = 1$. If d is a fundamental discriminant satisfying $(-1)^\lambda d > 0$ then the Kohnen–Zagier formula [1981, Theorem 1] can be written as

$$\Gamma(\ell)|\rho_g(|d|)|^2 = 4\pi \frac{\Gamma(2\ell-1)}{(4\pi)^{2\ell-1}\langle f, f\rangle} L\big(\tfrac{1}{2}, f \times \chi_d\big),$$

where $L(s, f \times \chi_d)$ is the twisted L-function with Dirichlet series

$$L(s, f \times \chi_d) = \sum_{m=1}^{\infty} \frac{a_f(m)\chi_d(m)}{m^s}. \tag{6-3}$$

By a result of Hoffstein and Lockhart (see [Hoffstein and Lockhart 1994, Corollary 0.3] and the second remark that follows it, and note that their normalization differs from ours) we have the bound

$$\frac{\Gamma(2\ell-1)}{(4\pi)^{2\ell-1}\langle f, f\rangle} \ll \ell^\varepsilon,$$

so we conclude that

$$\Gamma(\ell)|\rho_g(|d|)|^2 \ll L\big(\tfrac{1}{2}, f \times \chi_d\big)\ell^\varepsilon.$$

Let $\mathcal{H}_{2\ell-1}$ be the image in $\mathcal{S}_{2\ell-1}$ of the Shimura lift of $\mathcal{H}_\ell^+(\nu)$. Young's hybrid subconvexity bound [2017, Theorem 1.1] yields

$$\sum_{f \in \mathcal{H}_{2\ell-1}} L\big(\tfrac{1}{2}, f \times \chi_d\big)^3 \ll (\ell d)^{1+\varepsilon}$$

for odd fundamental d. See the Appendix for the case of even fundamental discriminants d. Applying Hölder's inequality in the case $\frac{1}{6} + \frac{1}{6} + \frac{2}{3} = 1$, together with the fact that $\#\mathcal{H}_{2\ell-1} \asymp \ell$, we obtain the following theorem for d, d' fundamental discriminants. It is extended to all m, n using (6-2).

Theorem 6.1. *Let $\ell \equiv k \pmod 2$ with $k = \pm\frac{1}{2} = \lambda + \frac{1}{2}$ and suppose that $\mathcal{H}_\ell^+$ is an orthonormal basis for $\mathcal{S}_\ell^+$ consisting of Hecke eigenforms. Suppose that m, n are integers with $(-1)^\lambda m, (-1)^\lambda n > 0$, and write $(-1)^\lambda m = v^2 d'$ and $(-1)^\lambda n = w^2 d$ with d, d' fundamental discriminants. Then*

$$\Gamma(\ell) \sum_{g \in \mathcal{H}_\ell^+} |\rho_g(|m|)\rho_g(|n|)| \ll \ell |dd'|^{\frac{1}{6}+\varepsilon} (vw)^\varepsilon.$$

Applying Theorem 6.1 to the sum (6-1) we find that

$$\mathcal{K}^h \ll |dd'|^{\frac{1}{6}+\varepsilon} (vw)^\varepsilon \sum_{\ell \equiv k(2)} \ell\, \widetilde{\varphi}(\ell).$$

The latter sum was estimated in [Sarnak and Tsimerman 2009] (see the discussion following (50); see also Lemma 5.1 of [Dunn 2018]) where the authors found that $\sum_\ell \ell\, \widetilde{\varphi}(\ell) \ll \sqrt{mn}/x$. We conclude that

$$\mathcal{K}^h \ll \frac{vw|dd'|^{\frac{2}{3}}}{x} (mn)^\varepsilon. \tag{6-4}$$

Next, we estimate the contribution from the Maass cusp forms

$$\mathcal{K}^m := \sqrt{mn} \sum_{j \geq 0} \frac{\overline{\rho_j(m)}\rho_j(n)}{\cosh \pi r_j} \widehat{\varphi}(r_j).$$

We follow the same general idea as in the holomorphic case, but instead of the Kohnen–Zagier formula we apply a formula of Baruch and Mao [2010]. As in the holomorphic case, we may assume that the orthonormal basis $\{u_j\}$ of $\mathcal{V}_k^+$ consists of eigenforms for the Hecke operators. Suppose that $u_j \in \mathcal{V}_k^+$ has spectral parameter r_j. The lowest eigenvalue is $\lambda_0 = \frac{3}{16}$ which corresponds to $u_0 = y^{1/4}\theta(z)$ or its conjugate. Since the coefficients $\rho_0(n)$ are supported on squares and since m, n are not both squares, we find that the term in $\mathcal{K}^m$ corresponding to $j = 0$ does not appear. In what follows we assume that $j \geq 1$.

Theorem 1.2 of [Baruch and Mao 2010] shows that there is a unique normalized Maass cusp form v_j of weight 0 with spectral parameter $2r_j$ which is even if $k = \frac{1}{2}$ and odd if $k = -\frac{1}{2}$, and such that the Hecke eigenvalues of u_j and v_j agree. Since there are no exceptional eigenvalues for weight 0 on $\mathrm{SL}_2(\mathbb{Z})$ this lift implies that there are no exceptional eigenvalues in weights $\pm\frac{1}{2}$ in the plus space. It follows that $r_j \geq 0$ for each $j \geq 1$ (in fact $r_1 \approx 1.5$). If $a_j(n)$ is the n-th coefficient of v_j (with respect to the Whittaker function, not the K-Bessel function) then for d a fundamental discriminant we have

$$w\, \rho_j(dw^2) = \rho_j(d) \sum_{\ell | w} \ell^{-1} \mu(\ell) \chi_d(\ell) a_j(w/\ell).$$

Let θ denote an admissible exponent toward the Ramanujan conjecture in weight 0; we have $\theta \leq \frac{7}{64}$ by work of Kim and Sarnak [2003]. Then $a_j(w) \ll w^{\theta+\varepsilon}$ since v_j is normalized so that $a_j(1) = 1$. Hence

$$w|\rho_j(dw^2)| \ll w^{\theta+\varepsilon} |\rho_j(d)|.$$

Suppose that d is a fundamental discriminant and that $\langle u_j, u_j\rangle = 1$. Then Theorem 1.4 of [Baruch and Mao 2010] implies that

$$|\rho_j(d)|^2 = \frac{L\left(\frac{1}{2}, v_j \times \chi_d\right)}{\pi|d|\langle v_j, v_j\rangle}\left|\Gamma\left(\frac{1-k\operatorname{sgn} d}{2} - ir_j\right)\right|^2,$$

where $L\left(\frac{1}{2}, v_j \times \chi_d\right)$ is defined in a similar way as (6-3). Hoffstein and Lockhart [1994, Corollary 0.3] proved that $\langle v_j, v_j\rangle^{-1} \ll (1+r_j)^{\varepsilon} e^{2\pi r_j}$ (again, note that the Fourier coefficients are normalized differently in that paper). It follows that

$$|d| \sum_{r_j \le x} \frac{|\rho_j(d)|^2}{\cosh \pi r_j} \ll \sum_{2r_j \le 2x} (1+r_j)^{-k\operatorname{sgn}(d)+\varepsilon} L\left(\tfrac{1}{2}, v_j \times \chi_d\right).$$

Young's subconvexity result [2017, Theorem 1.1] in this case shows that

$$\sum_{T \le r_j \le T+1} L\left(\tfrac{1}{2}, v_j \times \chi_d\right)^3 \ll (|d|(1+T))^{1+\varepsilon}.$$

After applying Hölder's inequality as above, we obtain the following.

Theorem 6.2. *Let $k = \pm\frac{1}{2} = \lambda + \frac{1}{2}$. Suppose that $\{u_j\}$ is an orthonormal basis for $\mathcal{V}_k^+$ consisting of Hecke eigenforms with spectral parameters r_j and coefficients ρ_j. Suppose that m, n are integers with $(-1)^\lambda m, (-1)^\lambda n > 0$, and write $(-1)^\lambda m = v^2 d'$ and $(-1)^\lambda n = w^2 d$ with d, d' fundamental discriminants not both equal to 1. Then*

$$\sqrt{|mn|} \sum_{r_j \le x} \frac{|\rho_j(m)\rho_j(n)|}{\cosh \pi r_j} \ll |dd'|^{\frac{1}{6}} (vw)^{\theta} x^{2-\frac{1}{2}k(\operatorname{sgn} m + \operatorname{sgn} n)} (mnx)^{\varepsilon}.$$

To estimate $\mathcal{K}^m$ we consider the dyadic sums

$$\mathcal{K}^m(A) := \sqrt{mn} \sum_{A \le r_j < 2A} \frac{\overline{\rho_j(m)}\rho_j(n)}{\cosh \pi r_j} \widehat{\varphi}(r_j)$$

for $A \ge 1$. Theorem 6.2 gives one estimate for the coefficients $|\rho_j(m)\rho_j(n)|$. Applying Cauchy–Schwarz and Theorem 4.1 with $\beta = \frac{1}{2} + \varepsilon$ we obtain a second estimate:

$$\sqrt{mn} \sum_{r_j \le A} \frac{|\rho_j(m)\rho_j(n)|}{\cosh \pi r_j} \ll A^{-k}\left(A^2 + (m+n)^{\frac{1}{4}} A + (mn)^{\frac{1}{4}}\right)(mnA)^{\varepsilon}.$$

These theorems together imply that

$$\sqrt{mn} \sum_{A \le r_j < 2A} \frac{|\rho_j(m)\rho_j(n)|}{\cosh \pi r_j} \ll A^{-k} \min\left((dd')^{\frac{1}{6}} (vw)^{\theta} A^2, A^2 + (m+n)^{\frac{1}{4}} A + (mn)^{\frac{1}{4}}\right)(mnA)^{\varepsilon}.$$

The following lemma gives an estimate for $\widehat{\varphi}(r)$.

Lemma 6.3. *If $r \geq 1$ then with $\varphi = \varphi_{a,x,T}$ as above we have*

$$\widehat{\varphi}(r) \ll r^k \min\Big(r^{-\frac{3}{2}}, r^{-\frac{5}{2}}\frac{x}{T}\Big).$$

If $|r| \leq 1$ then $\widehat{\varphi}(r) \ll |r|^{-2}$.

Proof. Recall that

$$\widehat{\varphi}(r) = \frac{-i\,\xi_k(r)}{\cosh 2\pi r}\int_0^\infty \big(\cos\big(\tfrac{1}{2}\pi k + \pi i r\big)J_{2ir}(x) - \cos\big(\tfrac{1}{2}\pi k - \pi i r\big)J_{-2ir}(x)\big)\varphi(x)\,\frac{dx}{x},$$

where $\xi_k(r) \asymp r^k$ as $r \to \infty$. Sarnak and Tsimerman [2009, (47)–(48)] proved that

$$e^{-\pi|r|}\int_0^\infty J_{2ir}(x)\varphi(x)\frac{dx}{x} \ll \min\Big(|r|^{-\frac{3}{2}}, |r|^{-\frac{5}{2}}\frac{x}{T}\Big)$$

for $|r| \geq 1$. The first statement of the lemma follows. The second is similar, using [Sarnak and Tsimerman 2009, (43)]. □

Since $\min(x, y) \ll x^a y^{1-a}$ for any $a \in [0, 1]$, we have

$$\mathcal{K}^m(A) \ll \min\Big(1, \frac{x}{AT}\Big)\big(\sqrt{A} + (dd')^{\frac{1}{12}}(vw)^{\frac{\theta}{2}}(m+n)^{\frac{1}{8}} + (dd')^{\frac{3}{16}}(vw)^{\frac{1}{8}+\frac{3}{4}\theta}\big)(mnA)^{\varepsilon},$$

where we used $a = \frac{1}{2}$ in the second term and $a = \frac{3}{4}$ in the third term. Summing over A we conclude that

$$\mathcal{K}^m \ll \left(\sqrt{\frac{x}{T}} + (dd')^{\frac{1}{12}}(vw)^{\frac{\theta}{2}}(m+n)^{\frac{1}{8}} + (dd')^{\frac{3}{16}}(vw)^{\frac{1}{8}+\frac{3}{4}\theta}\right)(mnx)^{\varepsilon}. \tag{6-5}$$

We turn to the estimate of the integral

$$\mathcal{K}^e := \int_{\mathbb{R}}\Big(\frac{d}{d'}\Big)^{ir}\frac{L\big(\frac{1}{2}-2ir, \chi_{d'}\big)L\big(\frac{1}{2}+2ir, \chi_d\big)\mathfrak{S}_{d'}(v, 2ir)\mathfrak{S}_d(w, 2ir)}{|\zeta(1+4ir)|^2\cosh\pi r\big|\Gamma\big(\frac{1}{2}(k+1)+ir\big)\big|^2}\widehat{\varphi}(r)\,dr.$$

By symmetry it suffices to estimate the integrals $\mathcal{K}_0^e = \int_0^1$ and $\mathcal{K}_1^e = \int_1^\infty$. Estimating the divisor sums trivially we find that

$$|\mathfrak{S}_d(w, s)| \leq \sigma_0(w)^2.$$

For $|r| \leq 1$ we have $|\zeta(1+4ir)|^2 \gg r^{-2}$ and $\cosh\pi r\big|\Gamma\big(\frac{1}{2}(k+1)+ir\big)\big|^2 \gg 1$, so by Lemma 6.3 we have the estimate

$$\mathcal{K}_0^e \ll (vw)^{\varepsilon}\int_0^1 \big|L\big(\tfrac{1}{2}-2ir, \chi_{d'}\big)L\big(\tfrac{1}{2}+2ir, \chi_d\big)\big|\,dr.$$

Since $\cosh\pi r\big|\Gamma\big(\frac{1}{2}(k+1)+ir\big)\big|^2 \sim \pi r^k$ for large r and since $|\zeta(1+4ir)|^{-1} \ll r^{\varepsilon}$ for all r we have by Lemma 6.3 that

$$\mathcal{K}_1^e \ll (vw)^{\varepsilon}\int_1^\infty \big|L\big(\tfrac{1}{2}-2ir, \chi_{d'}\big)L\big(\tfrac{1}{2}+2ir, \chi_d\big)\big|\,\frac{dr}{r^{3/2-\varepsilon}}.$$

We multiply each Dirichlet L-function by $r^{-3/8}$ and the last factor by $r^{3/4}$, then apply Hölder's inequality in the case $\frac{1}{6}+\frac{1}{6}+\frac{2}{3}=1$. We obtain

$$\mathcal{K}_1^e \ll (vw)^{\varepsilon}\bigg(\int_1^{\infty}\big|L\big(\tfrac{1}{2}+ir,\chi_{d'}\big)\big|^6\,\frac{dr}{r^{9/4}}\bigg)^{\frac{1}{6}}\bigg(\int_1^{\infty}\big|L\big(\tfrac{1}{2}+ir,\chi_{d}\big)\big|^6\,\frac{dr}{r^{9/4}}\bigg)^{\frac{1}{6}}\bigg(\int_1^{\infty}\frac{dr}{r^{9/8-\varepsilon}}\bigg)^{\frac{2}{3}}. \tag{6-6}$$

Young [2017] proved that

$$\int_T^{T+1}\big|L\big(\tfrac{1}{2}+ir,\chi_d\big)\big|^6\,dr \ll (|d|(1+T))^{1+\varepsilon},$$

from which it follows that $K_0^e \ll (vw)^{\varepsilon}|dd'|^{\frac{1}{6}+\varepsilon}$ and

$$\int_1^{\infty}\big|L\big(\tfrac{1}{2}+ir,\chi_d\big)\big|^6\,\frac{dr}{r^{9/4}} \le \sum_{T=1}^{\infty}\frac{1}{T^{9/4}}\int_T^{T+1}\big|L\big(\tfrac{1}{2}+ir,\chi_d\big)\big|^6\,dr \ll |d|^{1+\varepsilon}.$$

This, together with (6-6) proves that

$$\mathcal{K}^e \ll (vw)^{\varepsilon}|dd'|^{\frac{1}{6}+\varepsilon}. \tag{6-7}$$

Putting (6-4), (6-5), and (6-7) together, we find that

$$\sum_{4|c>0}\frac{S_k^+(m,n,c)}{c}\varphi\bigg(\frac{4\pi\sqrt{mn}}{c}\bigg) \ll \bigg(\sqrt{\frac{x}{T}}+\frac{vw|dd'|^{\frac{2}{3}}}{x}+(dd')^{\frac{1}{12}}(vw)^{\frac{\theta}{2}}(m+n)^{\frac{1}{8}}+(dd')^{\frac{3}{16}}(vw)^{\frac{1}{8}+\frac{3}{4}\theta}\bigg)(mnx)^{\varepsilon}.$$

To unsmooth the sum of Kloosterman sums, we argue as in [Sarnak and Tsimerman 2009; Ahlgren and Andersen 2018] to obtain

$$\sum_{4|c>0}\frac{S_k^+(m,n,c)}{c}\varphi\bigg(\frac{4\pi\sqrt{mn}}{c}\bigg)-\sum_{x\le c<2x}\frac{S_k^+(m,n,c)}{c} \ll \frac{T\log x}{\sqrt{x}}(mn)^{\varepsilon}.$$

Choosing $T=x^{\frac{2}{3}}$ and using that $m+n\le mn$ we obtain

$$\sum_{x\le c<2x}\frac{S_k^+(m,n,c)}{c} \ll \bigg(x^{\frac{1}{6}}+\frac{vw|dd'|^{\frac{2}{3}}}{x}+(dd')^{\frac{5}{24}}(vw)^{\frac{1}{4}+\frac{\theta}{2}}\bigg)(mnx)^{\varepsilon}. \tag{6-8}$$

To prove (1-11) we sum the initial segment $c\le(dd')^a(vw)^b$ and apply the Weil bound (1-12), then sum the dyadic pieces for $c\ge(dd')^a(vw)^b$ using (6-8). To balance the resulting terms we take $a=\frac{4}{9}$ and $b=\frac{2}{3}$, which gives the bound

$$\sum_{c\le x}\frac{S_k^+(m,n,c)}{c} \ll \big(x^{\frac{1}{6}}+(dd')^{\frac{2}{9}}(vw)^{\frac{1}{3}}\big)(mnx)^{\varepsilon}.$$

This completes the proof. □

Appendix: Young's theorem for even discriminants

Let D be a fundamental discriminant. Then $|D| = q$ or $4q$, where q is squarefree (but not necessarily odd). For a positive even integer k, let $\mathcal{B}_k(q)$ denote the set of weight k holomorphic Hecke newforms of level dividing q. Our goal in this appendix is to prove the following generalization of Young's hybrid subconvexity result [2017].

Theorem A.1. *Notation as above, we have*

$$\sum_{f \in B_k(q)} L\left(\tfrac{1}{2}, f \times \chi_D\right)^3 \ll (k|D|)^{1+\varepsilon}.$$

A corresponding generalization also holds for Maass cusp forms and Eisenstein series; for simplicity we only deal with the holomorphic case here.

For ease of comparison with [Young 2017], we have adopted the notation of that paper for this section only. We will indicate the changes that need to be made and refer the reader to [Conrey and Iwaniec 2000; Young 2017] for the remaining details. Starting in Section 4 of [Young 2017], our goal is to show that

$$\sum_{k \equiv a(4)} w\left(\frac{k-1-2T}{\Delta}\right) \sum_{f \in B_k(q)} \omega_f^* L\left(\tfrac{1}{2}, f \times \chi_D\right)^3 \ll \Delta (T|D|)^{1+\varepsilon},$$

where w is a smooth nonnegative function with support in $\left[\frac{1}{2}, 3\right]$ which equals 1 on the interval $[1, 2]$, and a is determined by $i^k = \chi_D(-1)$. Here ω_f^* is a Petersson weight satisfying $\omega_f^* \gg (kq)^{-\varepsilon}$. Applying the approximate functional equation and the Petersson formula as in Section 5 of [Young 2017], we find that it suffices to show the following.

Proposition A.2. *For $i = 1, 2, 3$, let w_i be a smooth weight function supported on $x \asymp N_i$, with $1 \ll N_i \ll (qT)^{1+\varepsilon}$ and with $w_i^{(k)} \ll N_i^{-k}$. Then*

$$\sum_{n_1, n_2, n_3} w_1(n_1) w_2(n_2) w_3(n_3) \chi_D(n_1 n_2 n_3) \sum_{c \equiv 0(q)} \frac{S(n_1 n_2, n_3; c)}{c} B\left(\frac{4\pi\sqrt{n_1 n_2 n_3}}{c}\right) \ll (N_1 N_2 N_3)^{\frac{1}{2}} \Delta T (qT)^{\varepsilon},$$

where $S(m, n; c)$ is the ordinary Kloosterman sum,

$$B(x) = B^{holo}(x) = \sum_{k \equiv a(4)} (k-1) w\left(\frac{k-1-2T}{\Delta}\right) J_{k-1}(x)$$

and $J_{k-1}(x)$ is the J-Bessel function.

With w_1, w_2, and w_3 as in Proposition A.2, let

$$S(N_1, N_2, N_3; C; B) = \sum_{\substack{c \asymp C \\ c \equiv 0(q)}} \mathcal{S}(N_1, N_2, N_3; c),$$

where

$$\mathcal{S}(N_1, N_2, N_3; c) = \sum_{n_1, n_2, n_3} \chi_D(n_1 n_2 n_3) S(n_1 n_2, n_3; c) w_1(n_1) w_2(n_2) w_3(n_3) B\left(\frac{4\pi\sqrt{n_1 n_2 n_3}}{c}\right).$$

We now follow Section 8 of [Young 2017], where the main difference is that we must keep track of the dependence on $\operatorname{lcm}(c, |D|)$, which we write as cs, with $s \in \{1, 2, 4, 8\}$. Applying Poisson summation modulo c to the sum over the lattice $\mathbb{Z}^3$ we find that

$$\mathcal{S}(N_1, N_2, N_3; c) = \sum_{m_1,m_2,m_3} G(m_1, m_2, m_3; c)\, K(m_1, m_2, m_3; c),$$

where

$$G(m_1, m_2, m_3; c) = \frac{1}{(cs)^3} \sum_{a_1,a_2,a_3 \bmod c} \chi_D(a_1a_2a_3)\, S(a_1a_2, a_3; c)\, e\left(\frac{a_1m_1 + a_2m_2 + a_3m_3}{cs}\right)$$

and

$$K(m_1, m_2, m_3; c) = \int_{\mathbb{R}^3} w_1(t_1)w_2(t_2)w_3(t_3) B\left(\frac{4\pi\sqrt{t_1t_2t_3}}{c}\right) e\left(\frac{-m_1t_1 - m_2t_2 - m_3t_3}{cs}\right) dt_1\, dt_2\, dt_3.$$

The analysis of the analytic piece $K(m_1, m_2, m_3; c)$ is almost exactly the same as in [loc. cit., Section 8]; simply replace t_i by $t_i/s^{2/3}$ and apply [loc. cit., Lemma 8.1]. The only difference is that the phase

$$e\left(\frac{-m_1m_2m_3}{c}\right)$$

in [loc. cit., (8.4)] is replaced by

$$e\left(\frac{-m_1m_2m_3}{s^3c}\right). \tag{A-1}$$

For the remainder of this section we will focus on the arithmetic piece $G(m_1, m_2, m_3; c)$. We begin by fixing notation. Let $D = tq'$, where t and q' are fundamental discriminants with $t \mid 2^\infty$ and q' odd, so that $\chi_D = \chi_t\chi_{q'}$. With $q \mid c$ and $cs = \operatorname{lcm}(c, D)$ as before, we have $s = t/(c, t)$. Finally, write $c = c_oc_e$, with $c_e \mid 2^\infty$ and c_o odd. Then cs factors as $cs = c_0 \cdot sc_e$ into odd and even parts. From the twisted multiplicativity of the Kloosterman sums, a straightforward computation gives the factorization

$$G(m_1, m_2, m_3; c) = G(m_1, m_2, \bar{c}_om_3; c_e)G(m_1, m_2, \bar{c}_e\bar{s}^3m_3; c_o), \tag{A-2}$$

where we choose the inverse $\bar{c}_o$ such that

$$c_o\bar{c}_o \equiv 1 \pmod{s^3c_e}.$$

The second term on the right-hand side of (A-2) was evaluated in Lemma 10.2 of [Conrey and Iwaniec 2000], which we record here in the following lemma (see also (9.2) of [Young 2017]). Note that Young's definition of $G(m_1, m_2, m_3; c)$, which we are using here, is slightly different from Conrey and Iwaniec's definition. Let $R_k(m) = S(0, m; k)$ denote the Ramanujan sum and let

$$H(w; q) = \sum_{u,v(q)} \chi_q(uv(u + 1)(v + 1))\, e\left(\frac{(uv - 1)w}{q}\right).$$

Lemma A.3. *Let $c_o = qr$ with c_o odd and q squarefree. Suppose m_1, m_2, m_3 are integers with*

$$(m_3, r) = 1 \quad \text{and} \quad (m_1 m_2, q, r) = 1. \tag{A-3}$$

Then we have

$$e\left(\frac{-m_1 m_2 m_3}{c_o}\right) G(m_1, m_2, m_3; c_o) = \frac{\chi_{k\ell}(-1)h}{rq^2\varphi(k)} R_k(m_1) R_k(m_2) R_k(m_3) H(\overline{rhk} m_1 m_2 m_3; \ell),$$

where $h = (r, q)$, $k = (m_1 m_2 m_3, q)$, and $\ell = q/hk$. If the coprimality conditions above are not satisfied, then $G(m_1, m_2, m_3; c_o)$ vanishes.

Petrow and Young [2019, Lemma 9.4] evaluated $G(m_1, m_2, m_3; c_e)$ when c_e is a power of 2.

Lemma A.4. *Suppose that $c_e \mid 2^\infty$ and factor m_i into even and odd parts as $m_i = m_i^e m_i^o$. Then*

$$e\left(\frac{-m_1 m_2 m_3}{s^3 c_e}\right) G(m_1, m_2, m_3; c_e) = \frac{s^3 c_e^2}{t} \sum_{\Delta \mid 64} \frac{1}{\varphi(\Delta)} \sum_{\chi \bmod \Delta} \mathfrak{g}_\chi \chi(m_1^o m_2^o m_3^o),$$

where $\mathfrak{g}_\chi$ depends on $m_1^e, m_2^e, m_3^e, t, c_e, \chi$ and is bounded by an absolute constant.

Note that the phase terms in Lemmas A.3 and A.4 combine to give

$$e\left(\frac{m_1 m_2 m_3}{s^3 c}\right),$$

which exactly matches the phase term (A-1) coming from $K(m_1, m_2, m_3; c)$.

The last result we require is the following analogue of Lemma 9.3 of [Young 2017]. The remainder of the proof of A.2 follows the proof of Proposition 7.3 of [Young 2017].

Lemma A.5. *Let $c = q'r$ with q' odd and squarefree. Let α_{m_1}, β_{m_2}, and γ_{m_3} be sequences of complex numbers satisfying $\alpha_{m_1} = \alpha_{m_1^e}\alpha_{m_1^o}$, $\beta_{m_2} = \beta_{m_2^e}\beta_{m_2^o}$, $\gamma_{m_3} = \gamma_{m_3^e}\gamma_{m_3^o}$, and $|\alpha_{m_1^e}| = |\beta_{m_2^e}| = |\gamma_{m_3^e}| = 1$, and let δ_r be an arbitrary sequence of complex numbers. Then for $U \geq 1$ we have*

$$\begin{aligned}\int_{|u|\leq U} \Bigg| \sum_{\substack{m_1,m_2,m_3 \\ m_i \asymp M_i}} \sum_{r \asymp R} \alpha_{m_1}\beta_{m_2}\gamma_{m_3}\delta_r G(m_1, m_2, m_3; c) e\left(\frac{-m_1 m_2 m_3}{s^3 c}\right)\left(\frac{m_1 m_2 m_3}{c}\right)^{iu} \Bigg| du \\ \ll \frac{q^{\frac{1}{2}+\varepsilon}}{Rq^2}(qU + M_1 M_2)^{\frac{1}{2}}(qU + M_3 R)^{\frac{1}{2}} \Bigg(\sum_{d, m_1, m_2, m_3, r} d^{1+\varepsilon} |\alpha_{m_1}\beta_{m_2}\gamma_{dm_3}\delta_{dr}|^2 \Bigg)^{\frac{1}{2}}. \end{aligned} \tag{A-4}$$

Remark. As in [Young 2017, Lemma 9.3], when $\gamma_{m_3}, \delta_r \ll 1$ the sum over d does not change the bound which arises from $d = 1$.

Proof. Using Lemmas A.3 and A.4, the left-hand side of (A-4) is

$$\begin{aligned}\ll \sum_{hk\ell = q'} \frac{h}{Rq^2\varphi(k)} \sum_{\Delta \mid 64} \frac{1}{\varphi(\Delta)} \int_{|u|\leq U} \Bigg| \sideset{}{^*}\sum_{m_1,m_2,m_3} \sideset{}{^*}\sum_{r \asymp R} \alpha_{m_1}\beta_{m_2}\gamma_{m_3}\delta_r \mathfrak{g}_\chi \chi(m_1^o m_2^o m_3^o) \\ \times R_k(m_1) R_k(m_2) R_k(m_3) H(\overline{rhk} m_1 m_2 m_3; \ell) \left(\frac{m_1^o m_2^o m_3^o}{q'}\right)^{iu} \Bigg| du,\end{aligned}$$

where the star indicates that the sum is restricted by the coprimality conditions (A-3). Using that $R_k(m_i) = R_k(m_i^e)R_k(m_i^o)$ and $|R_k(m)| \le (k, m)$ we bound the above by

$$\ll \sum_{hk\ell=q'} \frac{h}{Rq^2\varphi(k)} \sum_{\Delta|64} \frac{1}{\varphi(\Delta)} \sum_{\substack{j_1,j_2,j_3 \\ j_i \ll \log_2(M_i)}} \int_{|u|\le U} \Bigg| \sum_{\substack{m_1^o,m_2^o,m_3^o \\ m_i^o \asymp M_i/2^{j_i}, m_i^e = 2^{j_i}}}^{*} \sum_{r\asymp R}^{*} \alpha_{m_1^o}\beta_{m_2^o}\gamma_{m_3^o}\delta_r \chi(m_1^o m_2^o m_3^o)$$

$$\times R_k(m_1^o)R_k(m_2^o)R_k(m_3^o)H(\overline{rhk}bm_1^o m_2^o m_3^o;\ell)\left(\frac{m_1^o m_2^o m_3^o}{q'}\right)^{iu} \Bigg| du,$$

where $b = m_1^e m_2^e m_3^e$. Now following the proof of Lemma 9.3 of [Young 2017] almost exactly, we obtain the desired bound. □

Acknowledgement

The authors thank the referee for a thorough and careful reading of an earlier version of the manuscript, as well as helpful comments and suggestions.

References

[Ahlgren and Andersen 2018] S. Ahlgren and N. Andersen, "Kloosterman sums and Maass cusp forms of half integral weight for the modular group", *Int. Math. Res. Not.* **2018**:2 (2018), 492–570. MR Zbl

[Baruch and Mao 2010] E. M. Baruch and Z. Mao, "A generalized Kohnen–Zagier formula for Maass forms", *J. Lond. Math. Soc.* (2) **82**:1 (2010), 1–16. MR Zbl

[Biró 2000] A. Biró, "Cycle integrals of Maass forms of weight 0 and Fourier coefficients of Maass forms of weight 1/2", *Acta Arith.* **94**:2 (2000), 103–152. MR Zbl

[Bruinier et al. 2006] J. H. Bruinier, P. Jenkins, and K. Ono, "Hilbert class polynomials and traces of singular moduli", *Math. Ann.* **334**:2 (2006), 373–393. MR Zbl

[Conrey and Iwaniec 2000] J. B. Conrey and H. Iwaniec, "The cubic moment of central values of automorphic L-functions", *Ann. of Math.* (2) **151**:3 (2000), 1175–1216. MR Zbl

[Darmon and Vonk 2017] H. Darmon and J. Vonk, "Singular moduli for real quadratic fields: a rigid analytic approach", preprint, 2017, available at https://tinyurl.com/vonksing.

[Deshouillers and Iwaniec 1982] J.-M. Deshouillers and H. Iwaniec, "Kloosterman sums and Fourier coefficients of cusp forms", *Invent. Math.* **70**:2 (1982), 219–288. MR Zbl

[DLMF 2010] F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark (editors), "NIST digital library of mathematical functions", electronic reference, Nat. Inst. Standards Tech., 2010, available at http://dlmf.nist.gov.

[Duke 1988] W. Duke, "Hyperbolic distribution problems and half-integral weight Maass forms", *Invent. Math.* **92**:1 (1988), 73–90. MR Zbl

[Duke 2006] W. Duke, "Modular functions and the uniform distribution of CM points", *Math. Ann.* **334**:2 (2006), 241–252. MR Zbl

[Duke et al. 2002] W. Duke, J. B. Friedlander, and H. Iwaniec, "The subconvexity problem for Artin L-functions", *Invent. Math.* **149**:3 (2002), 489–577. MR Zbl

[Duke et al. 2011] W. Duke, O. Imamoḡlu, and A. Tóth, "Cycle integrals of the j-function and mock modular forms", *Ann. of Math.* (2) **173**:2 (2011), 947–981. MR Zbl

[Duke et al. 2012] W. Duke, J. B. Friedlander, and H. Iwaniec, "Weyl sums for quadratic roots", *Int. Math. Res. Not.* **2012**:11 (2012), 2493–2549. Correction in **2012**:11 (2012), 2646–2648. MR Zbl

[Duke et al. 2016] W. Duke, O. Imamoğlu, and A. Tóth, "Geometric invariants for real quadratic fields", *Ann. of Math.* (2) **184**:3 (2016), 949–990. MR Zbl

[Duke et al. 2018] W. Duke, O. Imamoğlu, and A. Tóth, "Kronecker's first limit formula, revisited", *Res. Math. Sci.* **5**:2 (2018), art. id. 20. MR Zbl

[Dunn 2018] A. Dunn, "Uniform bounds for sums of Kloosterman sums of half integral weight", *Res. Number Theory* **4**:4 (2018), art. id. 45. MR Zbl

[Folsom and Masri 2010] A. Folsom and R. Masri, "Equidistribution of Heegner points and the partition function", *Math. Ann.* **348**:2 (2010), 289–317. MR Zbl

[Gross et al. 1987] B. Gross, W. Kohnen, and D. Zagier, "Heegner points and derivatives of L-series, II", *Math. Ann.* **278**:1-4 (1987), 497–562. MR Zbl

[Hejhal 1983] D. A. Hejhal, *The Selberg trace formula for* $\mathrm{PSL}(2,\mathbb{R})$*, II*, Lecture Notes in Math. **1001**, Springer, 1983. MR Zbl

[Hoffstein and Lockhart 1994] J. Hoffstein and P. Lockhart, "Coefficients of Maass forms and the Siegel zero", *Ann. of Math.* (2) **140**:1 (1994), 161–181. MR Zbl

[Ibukiyama and Saito 2012] T. Ibukiyama and H. Saito, "On zeta functions associated to symmetric matrices, II: Functional equations and special values", *Nagoya Math. J.* **208** (2012), 265–316. MR Zbl

[Katok and Sarnak 1993] S. Katok and P. Sarnak, "Heegner points, cycles and Maass forms", *Israel J. Math.* **84**:1-2 (1993), 193–227. MR Zbl

[Kim and Sarnak 2003] H. H. Kim and P. Sarnak, "Refined estimates towards the Ramanujan and Selberg conjectures", (2003). Appendix to H. H. Kim, "Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2", J. Amer. Math. Soc.**16**:1 (2003), 139–183. MR Zbl

[Kohnen 1980] W. Kohnen, "Modular forms of half-integral weight on $\Gamma_0(4)$", *Math. Ann.* **248**:3 (1980), 249–266. MR Zbl

[Kohnen 1982] W. Kohnen, "Newforms of half-integral weight", *J. Reine Angew. Math.* **333** (1982), 32–72. MR Zbl

[Kohnen and Zagier 1981] W. Kohnen and D. Zagier, "Values of L-series of modular forms at the center of the critical strip", *Invent. Math.* **64**:2 (1981), 175–198. MR Zbl

[Kuznetsov 1980] N. V. Kuznetsov, "The Petersson conjecture for cusp forms of weight zero and the Linnik conjecture: sums of Kloosterman sums", *Mat. Sb.* (*N.S.*) **111(153)**:3 (1980), 334–383. In Russian; translated in *Math. USSR-Sb.* **39**:3 (1981), 299–342. MR

[Lehmer 1939] D. H. Lehmer, "On the remainders and convergence of the series for the partition function", *Trans. Amer. Math. Soc.* **46** (1939), 362–373. MR Zbl

[Maass 1949] H. Maass, "Über eine neue Art von nichtanalytischen automorphen Funktionen und die Bestimmung Dirichletscher Reihen durch Funktionalgleichungen", *Math. Ann.* **121** (1949), 141–183. MR Zbl

[Maass 1952] H. Maass, "Die Differentialgleichungen in der Theorie der elliptischen Modulfunktionen", *Math. Ann.* **125** (1952), 235–263. MR Zbl

[Masri 2012] R. Masri, "The asymptotic distribution of traces of cycle integrals of the j-function", *Duke Math. J.* **161**:10 (2012), 1971–2000. MR Zbl

[Motohashi 2003] Y. Motohashi, "A functional equation for the spectral fourth moment of modular Hecke L-functions", in *Proceedings of the Session in Analytic Number Theory and Diophantine Equations* (Bonn, Germany, 2002), edited by D. R. Heath-Brown and B. Z. Moroz, Bonner Math. Schriften **360**, Univ. Bonn, 2003. MR Zbl

[Petrow and Young 2019] I. Petrow and M. P. Young, "A generalized cubic moment and the Petersson formula for newforms", *Math. Ann.* **373**:1-2 (2019), 287–353. MR Zbl

[Proskurin 2003] N. V. Proskurin, "On general Kloosterman sums", *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* **302**:19 (2003), 107–134. In Russian; translated in *J. Math. Sci. (N.Y.)* **129**:3 (2005), 3874–3889. MR Zbl

[Roelcke 1966] W. Roelcke, "Das Eigenwertproblem der automorphen Formen in der hyperbolischen Ebene, II", *Math. Ann.* **168** (1966), 261–324. MR Zbl

[Sarnak 1984] P. Sarnak, "Additive number theory and Maass forms", pp. 286–309 in *Number theory* (New York, 1982), edited by D. V. Chudnovsky et al., Lecture Notes in Math. **1052**, Springer, 1984. MR Zbl

[Sarnak and Tsimerman 2009] P. Sarnak and J. Tsimerman, "On Linnik and Selberg's conjecture about sums of Kloosterman sums", pp. 619–635 in *Algebra, arithmetic, and geometry: in honor of Yu. I. Manin, II*, edited by Y. Tschinkel and Y. Zarhin, Progr. Math. **270**, Birkhäuser, Boston, 2009. MR Zbl

[Selberg 1956] A. Selberg, "Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series", *J. Indian Math. Soc.* (*N.S.*) **20** (1956), 47–87. MR Zbl

[Selberg 1965] A. Selberg, "On the estimation of Fourier coefficients of modular forms", pp. 1–15 in *Theory of numbers* (Pasadena, CA, 1963), edited by A. L. Whiteman, Proc. Sympos. Pure Math. **8**, Amer. Math. Soc., Providence, RI, 1965. MR Zbl

[Shimura 1973] G. Shimura, "On modular forms of half integral weight", *Ann. of Math.* (2) **97** (1973), 440–481. MR Zbl

[Sturm 1980] J. Sturm, "Special values of zeta functions, and Eisenstein series of half integral weight", *Amer. J. Math.* **102**:2 (1980), 219–240. MR Zbl

[Titchmarsh 1951] E. C. Titchmarsh, *The theory of the Riemann zeta-function*, Clarendon, Oxford, 1951. MR Zbl

[Young 2017] M. P. Young, "Weyl-type hybrid subconvexity bounds for twisted L-functions and Heegner points on shrinking sets", *J. Eur. Math. Soc.* **19**:5 (2017), 1545–1576. MR Zbl

[Zagier 1975] D. Zagier, "A Kronecker limit formula for real quadratic fields", *Math. Ann.* **213** (1975), 153–184. MR Zbl

Communicated by Roger Heath-Brown
Received 2019-05-01 Revised 2019-12-09 Accepted 2020-02-06

nandersen@math.ucla.edu *Brigham Young University, Provo, UT, United States*

wdduke@ucla.edu *UCLA, Los Angeles, CA, United States*

msp

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1577

Generically free representations I: Large representations

Skip Garibaldi and Robert Guralnick

This paper concerns a faithful representation V of a simple linear algebraic group G. Under mild assumptions, we show that if V is large enough, then the Lie algebra of G acts generically freely on V. That is, the stabilizer in $\mathrm{Lie}(G)$ of a generic vector in V is zero. The bound on $\dim V$ grows like $(\mathrm{rank}\, G)^2$ and holds with only mild hypotheses on the characteristic of the underlying field. The proof relies on results on generation of Lie algebras by conjugates of an element that may be of independent interest. We use the bound in subsequent works to determine which irreducible faithful representations are generically free, with no hypothesis on the characteristic of the field. This in turn has applications to the question of which representations have a stabilizer in general position.

Let V be a faithful representation of a simple linear algebraic group G over an algebraically closed field k. In the special case $k = \mathbb{C}$, there is a striking dichotomy between the properties of irreducible representations V whose dimension is small (say, $\leq \dim G$) versus those whose dimension is large, see [Andreev et al. 1967; Elashvili 1972; Popov 1987] for original results and [Popov and Vinberg 1994, §8.7] for a

MSC2010: primary 20G05; secondary 17B10.
Keywords: generically free, virtually free, Lie algebra, representation, generic stabilizer.

G classical				G exceptional		
type of G	char k	$b(G)$	Reference	type of G	char k	$b(G)$
A_ℓ	$\neq 2$	$2.25(\ell+1)^2$	Corollary 6.5	G_2	$\neq 3$	48
A_ℓ	$=2$	$2\ell^2+4\ell$	Corollary 7.2	F_4	$\neq 2$	240
B_ℓ ($\ell \geq 3$)	$\neq 2$	$8\ell^2$	Corollary 9.2	E_6	any	360
C_ℓ ($\ell \geq 2$)	$\neq 2$	$6\ell^2$	Corollary 8.3	E_7	any	630
D_ℓ ($\ell \geq 4$)	$\neq 2$	$2(2\ell-1)^2$	Corollary 9.2	E_8	any	1200
D_ℓ ($\ell \geq 4$)	$=2$	$4\ell^2$	Corollary 10.6			

Table 1. Bound $b(G)$ appearing in Theorem A. The reference for the exceptional types is Proposition 11.4.

survey and bibliography. For example, if $\dim V < \dim G$, then trivially the stabilizer G_v of a vector $v \in V$ is not 1. On the other hand (and nontrivially), for $\dim V$ hardly bigger than $\dim G$, the stabilizer $G_v(k)$ for generic $v \in V$ is 1; in this case one says that V is *generically free* or G acts *generically freely* on V. This property has taken on increased importance recently due to applications in Galois cohomology and essential dimension, see [Reichstein 2010; Merkurjev 2013] for the theory and [Brosnan et al. 2010; Garibaldi and Guralnick 2017; Karpenko 2010; Lötscher et al. 2013; Lötscher 2013] for specific applications.

With applications in mind, it is desirable to extend the results on generically free representations to all fields. The paper [Guralnick and Lawther 2019] showed that, for k of any characteristic and V irreducible, $\dim V > \dim G$ if and only if the stabilizer $G_v(k)$ of a generic $v \in V$ is finite. (This was previously known when $\operatorname{char} k = 0$ [Andreev et al. 1967].) Moreover, except for the cases in Table 5, when $G_v(k)$ is finite it is 1, i.e., the group scheme G_v is infinitesimal. For applications, it is helpful to know if G_v is not just infinitesimal but is the *trivial* group scheme. In this paper, we prove the following general bound:

Theorem A. *Let G be a simple linear algebraic group over a field k such that* $\operatorname{char} k$ *is not special for G. If $\rho\colon G \to \mathrm{GL}(V)$ is a representation of G such that V has a G-subquotient X with $X^{[\mathfrak{g},\mathfrak{g}]} = 0$ and* $\dim X > b(G)$ *for $b(G)$ as in Table 1, then* $\mathrm{Lie}(G)_v = \ker \mathrm{d}\rho$ *for generic $v \in V$.*

Of course, $\mathrm{Lie}(G)_v \supseteq \ker \mathrm{d}\rho$, so equality means that $\mathrm{Lie}(G)_v$ is as small as possible. In this case, we write that $\mathrm{Lie}(G)$ acts *virtually freely* on V. This notion is the natural generalization of "generically freely" to allow for the possibility that G does not act faithfully. We actually prove a somewhat stronger statement than Theorem A; see Theorem 12.2 below.

Note that $\ker \mathrm{d}\rho$ can be read off the weights of V. If $\ker \mathrm{d}\rho$ is a proper ideal in $\mathrm{Lie}(G)$, then (as $\operatorname{char} k$ is assumed not special) it is contained in the center of $\mathrm{Lie}(G)$, i.e., $\mathrm{Lie}(Z(G))$. The restrictions of ρ to $Z(G)$ and of $\mathrm{d}\rho$ to $\mathrm{Lie}(Z(G))$ are determined by the equivalence classes of the weights of V modulo the root lattice.

If we restrict our focus to representations V that are irreducible and whose highest weight is restricted, Theorem A quickly settles whether V is virtually free for all but finitely many types of G:

Corollary B. *Suppose G has type A_ℓ for some $\ell > 15$; type B_ℓ, C_ℓ, or D_ℓ with $\ell > 11$; or exceptional type, over an algebraically closed field k such that* $\operatorname{char} k$ *is not special for G. For $\rho\colon G \to \mathrm{GL}(V)$ an irreducible representation of G whose highest weight is restricted,* $\mathrm{Lie}(G)_v = \ker \mathrm{d}\rho$ *for generic $v \in V$ if and only if* $\dim V > \dim G$.

This is proved in Section 13.

Note that the bound $b(G)$ from Theorem A holds for most k and is $\Theta(\dim G) = \Theta((\operatorname{rank} G)^2)$ in big-O notation, meaning that it grows like $(\operatorname{rank} G)^2$. In the special case $\operatorname{char} k = 0$ one can find a similar result in [Andreev and Popov 1971] where the bound is $\Theta((\operatorname{rank} G)^3)$, which was used in the (existing) proof of the characteristic 0 version of the results of Section 15. The fact that the exponent in our result is 2 (and not 3) is leveraged in two ways: (1) the restricted irreducible representations not covered by Theorem A and Corollary B are among those enumerated in [Lübeck 2001a] and (2) it encompasses all but a very small number of tensor decomposable irreducible representations. We settle these cases in a separate paper, [Garibaldi and Guralnick 2020a], because the arguments are rather different and more computational. Fields with $\operatorname{char} k$ special are treated in [Garibaldi and Guralnick 2020b], which also includes an example to show that the conclusion of Theorem A does not hold for such k and an extension of Corollary B (stated here as Theorem D). Combining the results of these two papers with [Guralnick and Lawther 2019], we get descriptions of the stabilizer G_v as a group scheme when V is irreducible, which we announce in Section 15. This paper contains the main part of the proof of the results in Section 15 for Lie algebras.

Remarks on the proof. Corollary B may be compared to the main result of Guerreiro's thesis [1997], which classifies the irreducible G-modules that are also $\mathrm{Lie}(G)$-irreducible such that the kernel of $\mathrm{d}\rho$ is contained in the center of $\mathrm{Lie}(G)$ with somewhat weaker bounds on $\dim V$. (See also [Auld 2001; Garibaldi and Guralnick 2017] for other results on specific representations.) Our methods are different in the sense that Guerreiro relied on computations with the weights of V, whereas we largely work with the natural module. We do refer to Guerreiro's thesis in the proof of Corollary B to handle a few specific representations.

The change in perspective that leads to our stronger results in fewer pages is the replacement of the popular inequality (1.3), which involves the action on the specific representation V, with (1.4), which only involves the dimension of V and properties of the adjoint representation $\mathrm{Lie}(G)$. Thus our proof of Theorem A depends in only a small way on V, providing a dramatic simplification. Furthermore we prove new bounds on the number of conjugates $e(x)$ of a given noncentral element $x \in \mathrm{Lie}(G)$ that suffice to generate a Lie subalgebra containing the derived subalgebra (with special care being needed in small characteristic; see, for example, Theorem 5.8); these results should be of independent interest. Our bounds depend upon the conjugacy class and give upper bounds for the dimension of fixed spaces for elements in the class. As a special case, we extend the main result of [Cohen et al. 2001]; see Proposition 14.1. We note that some generation bounds are known in the setting of groups; see for example [Gordeev and Saxl 2002a; Gordeev and Saxl 2002b] or [Guralnick and Saxl 2003].

We also prove a result that is of independent interest. We show in Theorem 5.8 that the only proper irreducible Lie subalgebras of $\mathfrak{sl}_n$ containing a maximal toral subalgebra occur in characteristic 2 and any such is conjugate to the Lie algebra of symmetric matrices of trace 0.

Notation. For convenience of exposition, we will assume in most of the rest of the paper that k is algebraically closed of characteristic $p \neq 0$. This is only for convenience, as our results for p prime immediately imply the corresponding results for characteristic zero: simply lift the representation from characteristic 0 to $\mathbb{Z}$ and reduce modulo a sufficiently large prime.

Let G be an affine group scheme of finite type over k. If G is additionally smooth, then we say that G is an *algebraic group*. An algebraic group G is *simple* if its radical is trivial (i.e., it is semisimple), it is not equal to 1, and its root system is irreducible. For example, SL_n is simple for every $n \geq 2$.

We say that $\operatorname{char} k$ is *special* for G if $\operatorname{char} k = p \neq 0$ and the Dynkin diagram of G has a p-valent bond, i.e., if $\operatorname{char} k = 2$ and G has type B_n or C_n for $n \geq 2$ or type F_4, or if $\operatorname{char} k = 3$ and G has type G_2. (Equivalently, these are the cases where G has a very special isogeny.) This definition is as in [Steinberg 1963, §10; Seitz 1987, p. 15; Premet 1997]; in an alternative history, these primes might have been called "extremely bad" because they are a subset of the very bad primes — the lone difference is that for G of type G_2, the prime 2 is very bad but not special.

A dominant weight λ is *restricted* if, when we write

$$\lambda = \sum c_\omega \omega,$$

where ω varies over the fundamental dominant weights, we have $0 \leq c_\omega < p$ for all ω.

If G acts on a variety X, the stabilizer G_x of an element $x \in X(k)$ is a subgroup-scheme of G with R-points

$$G_x(R) = \{g \in G(R) \mid gx = x\}$$

for every k-algebra R. A statement "for generic x" means that there is a dense open subset U of X such that the property holds for all $x \in U$.

If $\operatorname{Lie}(G) = 0$ then G is finite and étale. If additionally $G(k) = 1$, then G is the trivial group scheme $\operatorname{Spec} k$. (Note, however, that when k has characteristic $p \neq 0$, the subgroup-scheme μ_p of μ_{p^2} has the same Lie algebra and k-points. So it is not generally possible to distinguish closed subgroup-schemes by comparing their k-points and Lie algebras.)

We write $\mathfrak{g}$ for $\operatorname{Lie}(G)$ and similarly $\mathfrak{spin}_n$ for $\operatorname{Lie}(\mathrm{Spin}_n)$, etc. We put $\mathfrak{z}(\mathfrak{g})$ for the center of $\mathfrak{g}$; it is the Lie algebra of the (scheme-theoretic) center of G. As $\operatorname{char} k = p$, the Frobenius automorphism of k induces a "p-mapping" $x \mapsto x^{[p]}$ on $\mathfrak{g}$. When G is a subgroup-scheme of GL_n and $x \in \mathfrak{g}$, the element $x^{[p]}$ is the p-th power of x with respect to the typical, associative multiplication for n-by-n matrices; see [Demazure and Gabriel 1970, §II.7, p. 274]. An element $x \in \mathfrak{g}$ is *nilpotent* if $x^{[p]^n} = 0$ for some $n > 0$, *toral* if $x^{[p]} = x$, and *semisimple* if x is contained in the Lie p-subalgebra of $\mathfrak{g}$ generated by $x^{[p]}$, i.e., is in the subspace spanned by $x^{[p]}, x^{[p]^2}, \ldots$, cf. [Strade and Farnsteiner 1988, §2.3].

1. Key inequalities

Inequalities. Put $\mathfrak{g} := \mathrm{Lie}(G)$ and choose a representation $\rho\colon G \to \mathrm{GL}(V)$. For $x \in \mathfrak{g}$, put

$$V^x := \{v \in V \mid \mathrm{d}\rho(x)v = 0\}$$

and x^G for the G-conjugacy class $\mathrm{Ad}(G)x$ of x.

Lemma 1.1. *For $x \in \mathfrak{g}$,*

$$x^G \cap \mathfrak{g}_v = \varnothing \quad \textit{for generic } v \in V \tag{1.2}$$

is implied by:

$$\dim x^G + \dim V^x < \dim V, \tag{1.3}$$

which is implied by:

$$\begin{aligned}&\textit{There exist } e > 0 \textit{ and } x_1, \dots x_e \in x^G \textit{ such that the subalgebra } \mathfrak{s} \textit{ of } \mathfrak{g}\\ &\textit{generated by } x_1, \dots, x_e \textit{ has } V^{\mathfrak{s}} = 0 \textit{ and } e \cdot \dim x^G < \dim V.\end{aligned} \tag{1.4}$$

In many uses of (1.4), one takes $\mathfrak{s}$ to be $\mathfrak{g}$ or $[\mathfrak{g}, \mathfrak{g}]$.

Proof. Suppose (1.3) holds and let $v \in V$. Put

$$V(x) := \{v \in V \mid \text{there is } g \in G(k) \text{ such that } xgv = 0\} = \bigcup_{y \in x^G} V^y.$$

Define $\alpha\colon G \times V^x \to V$ by $\alpha(g, w) = gw$, so the image of α is precisely $V(x)$. The fiber over gw contains (gc^{-1}, cw) for $\mathrm{Ad}(c)$ fixing x, and so $\dim V(x) \le \dim x^G + \dim V^x$. Then (1.3) implies $\overline{V(x)}$ is a proper subvariety of V, whence (1.2). (This observation is essentially in [Andreev and Popov 1971, Lemma 4; Guerreiro 1997, §3.3; Garibaldi and Guralnick 2017, Lemma 2.6], for example, but we have repackaged it here for the convenience of the reader.)

Now assume (1.4). Iterating the formula $\dim(U \cap U') \ge \dim U + \dim U' - \dim V$ for subspaces U, U' of V gives

$$\dim\Big(\bigcap\nolimits_{i=1}^{e} V^{x_i}\Big) \ge \Big(\sum\nolimits_{i=1}^{e} \dim V^{x_i}\Big) - (e-1)\dim V. \tag{1.5}$$

As $\mathrm{d}\rho$ is G-equivariant (and not just a representation of $\mathfrak{g}$), we have $\dim V^{x_i} = \dim V^x$. The left side of (1.5) is zero by hypothesis; hence $\dim V^x \le (1 - 1/e)\dim V$. Now $\dim x^G < (1/e)\dim V$ implies (1.3). □

We will verify (1.3) in many cases, compare Theorem 12.2. To do so, we actually prove (1.4), where the inequality only involves V through the term $\dim V$. This allows us to focus on the element x and its action on the natural module rather than attempting to analyze V^x directly, for which it is natural to require some hypothesis on the structure of V beyond simply a bound on the dimension, such as that V is irreducible as is assumed in [Guerreiro 1997]. When verifying (1.4), one finds that, roughly speaking, when $\dim x^G$ is small, e is large and vice versa. Therefore, at least for the classical groups, we take some care to bound the product $e \cdot \dim x^G$ instead of bounding each term independently.

Comparing subalgebras. We will use the following, which is a small variation on [Garibaldi and Guralnick 2017, Lemma 2.6].

Lemma 1.6. *Suppose G is semisimple over an algebraically closed field k of characteristic $p > 0$, and let $\mathfrak{h}$ be a subspace of $\mathfrak{g}$.*

(1) *If (1.2) holds for every toral or nilpotent $x \in \mathfrak{g} \setminus \mathfrak{h}$, then $\mathfrak{g}_v \subseteq \mathfrak{h}$ for generic $v \in V$.*

(2) *If $\mathfrak{h}$ consists of semisimple elements and (1.2) holds for every $x \in \mathfrak{g} \setminus \mathfrak{h}$ with $x^{[p]} \in \{0, x\}$, then $\mathfrak{g}_v \subseteq \mathfrak{h}$ for generic v in V.*

Proof. For (1), as G is semisimple, there are only finitely many G-conjugacy classes in $\mathfrak{g}$. Therefore, by hypothesis there is a dense open subset U of V such that $x^G \cap \mathfrak{g}_v = \varnothing$ for all $v \in U$ and all toral or nilpotent $x \in \mathfrak{g} \setminus \mathfrak{h}$.

Fix $v \in U$. As k is algebraically closed, every $y \in \mathfrak{g}_v$ can be written as a linear combination of toral and nilpotent elements in $\mathfrak{g}_v$ [Strade and Farnsteiner 1988, p. 82, Theorem 2.3.6(2)], which must belong to $\mathfrak{h}$, so $\mathfrak{g}_v \subseteq \mathfrak{h}$.

Claim (2) now follows as in the proof of [Garibaldi and Guralnick 2017, Lemma 2.6(3)]. □

Often we apply the preceding lemma with $\mathfrak{h} = \mathfrak{z}(\mathfrak{g})$, the Lie algebra of the center $Z(G)$. For G reductive, $Z(G)$ is a diagonalizable group scheme [SGA 3_{III} 2011, XXII.4.1.6], so $Z(G)_v = \ker(\rho|_{Z(G)})$. We immediately obtain:

Lemma 1.7. *Suppose G is reductive. If $\mathfrak{g}_v \subseteq \mathfrak{z}(\mathfrak{g})$ for generic $v \in V$, then $\mathfrak{g}$ acts virtually freely on V.* □

Examples.

Example 1.8 (SL_2). Recall that an irreducible representation $\rho\colon \mathrm{SL}_2 \to \mathrm{GL}(V)$ of SL_2 is specified by its highest weight w, a nonnegative integer. Let $\operatorname{char} k =: p \neq 0$. We claim:

(i) *If $\operatorname{char} k$ divides w (e.g., if $w = 0$), then $\mathrm{d}\rho(\mathfrak{sl}_2) = 0$.*

(ii) *If (a) $w = 1$ or (b) $\operatorname{char} k \neq 2$ and $w = 2$, then $\mathfrak{sl}_2$ does not act virtually freely on V.*

(iii) *If $w = p^e + 1$ for some $e > 0$, then $\mathfrak{sl}_2$ acts virtually freely on V but (1.3) fails for some noncentral $x \in \mathfrak{sl}_2$ with $x^{[p]} \in \{0, x\}$.*

(iv) *Otherwise, (1.3) holds for noncentral $x \in \mathfrak{sl}_2$ with $x^{[p]} \in \{0, x\}$, and in particular $\mathfrak{sl}_2$ acts virtually freely on V.*

To see this, write $w = \sum_{i \geq 0} w_i p^i$, where $0 \leq w_i < p$. By Steinberg, V is isomorphic (as an SL_2-module) to $\bigotimes_i L(\omega_i)^{[p]^i}$, where the exponent $[p]^i$ denotes the i-th Frobenius twist, and the irreducible module $L(w_i)$ with highest weight w_i is also the Weyl module with highest weight w_i by [Winter 1977], of dimension $w_i + 1$. Thus, as a representation of $\mathfrak{sl}_2$, V is isomorphic to a direct sum of $c := \prod_{i>0}(w_i + 1)$ copies of $L(w_0)$. This proves (i), so we suppose for the remainder of the proof that $w_0 > 0$.

As in the previous paragraph, $L(1)$ is the natural representation (with generic stabilizer a maximal nilpotent subalgebra) and $L(2)$ (when $p \neq 2$) is the adjoint action on $\mathfrak{sl}_2$ (with generic stabilizer a Cartan subalgebra). This verifies (ii).

We investigate now (1.3). For x nonzero nilpotent or noncentral toral, we have $\dim(x^{\mathrm{SL}_2}) = 2$. For x nonzero nilpotent, $L(w_0)^x$ is conjugate to the highest weight line. If $x^{[p]} = x$, then up to conjugacy x is diagonal with entries $(a, -a)$ for some $a \in \mathbb{F}_p$; as x is noncentral, $p \neq 2$ and $\dim L(w_0)^x = 0$ or 1 depending on whether w_0 is odd or even. Assembling these, we find $\dim(x^{\mathrm{SL}_2}) + \dim L(w)^x \leq 2 + c$ with equality for x nonzero nilpotent, whereas $\dim L(w) = cw_0 + c$. We divide the remaining cases via the product cw_0, where we have already treated the case (ii) where $c = 1$ and $w_0 = 1$ or 2.

Suppose $c = 2$ and $w_0 = 1$, so we are in case (iii). The action of $\mathfrak{sl}_2$ on V via $\mathrm{d}\rho$ is the same as the action of $\mathfrak{sl}_2$ on two copies of the natural module, equivalently, on 2-by-2 matrices by left multiplication. A generic matrix v is invertible, so $(\mathfrak{sl}_2)_v = 0$. Yet we have verified in the previous paragraph that (1.3) fails for x nonzero nilpotent, proving (iii).

The case (iv) is where $cw_0 > 2$, where we have verified (1.3), completing the proof of the claim.

As a corollary, we find: $\mathfrak{sl}_2$ *fails to act virtually freely on* V *if and only if (a)* $w = 1$ *or (b)* $\operatorname{char} k \neq 2$ *and* $w = 2$. Moreover, when $\mathfrak{sl}_2$ acts faithfully on V (i.e, w_0 is odd), we have: $\mathfrak{sl}_2$ *fails to act generically freely on* V *if and only if* $w = 1$ *if and only if* $\dim V \leq \dim \mathrm{SL}_2$.

Example 1.9. Let $x \in \mathfrak{g}$. *If* $\dim x^G + \dim(V^*)^x < \dim(V^*)$, *then* (1.3) *holds for* x. This is obvious, because $\mathrm{d}\rho(x)$ and $-\mathrm{d}\rho(x)^\top$ have the same rank.

2. Interlude: semisimplification

For Theorem A, we consider representations V of G that need not be semisimple. For each chain of submodules $0 =: V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots \subseteq V_n := V$ of G, we can construct the G-module $V' := \bigoplus_{i=1}^n V_i/V_{i-1}$. For example, if each V_i/V_{i-1} is an irreducible (a.k.a. simple) G-module then V' is the *semisimplification* of V. In this section, we discuss to what extent results for V correspond to results for V_i/V_{i-1} and for V', using the notation of this paragraph and writing $\rho\colon G \to \mathrm{GL}(V)$ and $\rho'\colon G \to \mathrm{GL}(V')$ for the actions.

From the subquotient to V.

Example 2.1. Suppose that for some $x \in \mathfrak{g}$ and some $1 \leq i \leq n$, we have

$$\dim x^G + \dim(V_i/V_{i-1})^x < \dim(V_i/V_{i-1}).$$

We claim that (1.3) *holds for* x. By induction it suffices to consider the case $i = 2$ and a chain $V_1 \subseteq V_2 \subseteq V$.

Suppose first that $V_1 = 0$. Then $\dim x^G + \dim V^x \leq \dim x^G + \dim V_2^x + \dim V/V_2$, whence the claim. Now suppose that $V_2 = V$, so $(V_2/V_1)^*$ is a submodule of V^*; the claim follows by Example 1.9. Combining these two cases gives the full claim.

There is an analogous statement about the dimension of generic stabilizers.

Example 2.2. For each $1 \le i \le n$ and generic $v \in V$ and generic $w \in V_i/V_{i-1}$, we claim that $\dim \mathfrak{g}_v \le \dim \mathfrak{g}_w$. Take $w \in V_i/V_{i-1}$ to be the image of a generic $\widehat{w} \in V_i$. Then $\dim \mathfrak{g}_v \le \dim \mathfrak{g}_{\widehat{w}}$ by upper semicontinuity of dimension and clearly $\dim \mathfrak{g}_{\widehat{w}} \le \dim \mathfrak{g}_w$.

From V' to V.

Example 2.3. When checking the inequality (1.3), it suffices to do it for V'. More precisely, for $x \in \mathfrak{g}$, we have: *If* $\dim x^G + \dim(V')^x < \dim V'$, *then* $\dim x^G + \dim V^x < \dim V$. This is obvious because $\dim V^x \le \sum \dim(V_i/V_{i-1})^x$.

The following strengthens Example 2.2.

Proposition 2.4. *For generic $v \in V$ and $v' \in V'$, we have* $\dim \mathfrak{g}_v \le \dim \mathfrak{g}_{v'}$.

Proof. By induction on the number n of summands in V', we may assume that $V' = W \oplus V/W$ for some $\mathfrak{g}$-submodule W of V.

Suppose first that $\dim V/W = 1$. Pick $v \in V$ with nonzero image $\bar{v} \in V/W$. Put $\mathfrak{t} := \{x \in \mathfrak{g} \mid \mathrm{d}\rho(x)v \in W\}$, a subalgebra of $\mathfrak{g}$ sometimes called the transporter of v in W. A generic vector $v' \in V'$ is of the form $w \oplus c\bar{v}$ for $w \in W$ and $c \in k^\times$. Evidently, $\mathfrak{g}_{v'} = \mathfrak{t}_w$. By upper semicontinuity of dimension, $\dim \mathfrak{t}_{v_0} \le \dim \mathfrak{t}_w$ for generic $v_0 \in V$. On the other hand, writing $v_0 = w_0 + \lambda v$ for $\lambda \in k^\times$ and $w_0 \in W$, for $x \in \mathfrak{g}_{v_0}$ we find $\mathrm{d}\rho(x)v = -\frac{1}{\lambda}\mathrm{d}\rho(x)w_0 \in W$, so $\mathfrak{g}_{v_0} = \mathfrak{t}_{v_0}$, proving the claim.

In the general case, pick a splitting $\phi\colon V/W \hookrightarrow V$ and so identify V with V' as vector spaces. We may intersect open sets defining generic elements in V and V' and so assume the two notions agree under this identification. Let $v := w + \phi(\bar{v})$ be a generic vector in V, where $w \in W$ and $\bar{v} \in V/W$ is the image of v; $v' := w \oplus \bar{v}$ is a generic vector in V'. Defining $\mathfrak{t}$ as in the previous paragraph, we have $\mathfrak{g}_v, \mathfrak{g}_{v'} \subseteq \mathfrak{t}$. Replacing $\mathfrak{g}$, V, V' with $\mathfrak{t}$, $W + kv$, $W \oplus k\bar{v}$ and referring to the previous paragraph gives the claim. □

If $\mathfrak{g}$ acts generically freely on V' (i.e., $\mathfrak{g}_{v'} = 0$), then the proposition says that $\mathfrak{g}$ acts generically freely on V. This immediately gives the following statement about group schemes:

Corollary 2.5. *If $G_{v'}$ is finite étale for generic $v' \in V'$, then G_v is finite étale for generic $v \in V$.* □

While generic freeness of V' implies generic freeness of V for the action by the Lie algebra $\mathfrak{g}$, it does not do so for the action by the algebraic group G, as the following example shows.

Example 2.6. Take $G = \mathbb{G}_a$ acting on $V = \mathbb{A}^3$ via

$$\rho(r) := \begin{pmatrix} 1 & r & r^p \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Let $V_2 \subset V$ be the subspace of vectors whose bottom entry is zero. Then G acts on V_2 via $r \mapsto \left(\begin{smallmatrix} 1 & r \\ 0 & 1 \end{smallmatrix}\right)$ and in particular a generic $v_2 \in V_2$ has $G_{v_2} = 1$. On the other hand, a generic vector $v := \left(\begin{smallmatrix} x \\ y \\ z \end{smallmatrix}\right)$ in V has G_v the étale subgroup with points $\{r \mid ry + r^p z = 0\}$, i.e., the kernel of the homomorphism $zF + y\,\mathrm{Id}\colon \mathbb{G}_a \to \mathbb{G}_a$ for F the Frobenius map.

Direct sums have better properties with respect to calculating generic stabilizers; see for example [Popov 1988, Proposition 8; Lötscher 2015, Lemma 2.15].

3. Lemmas on the structure of $\mathfrak{g}$

When $\operatorname{char} k$ is not zero (more precisely, not very good), then it may happen that $\mathfrak{g}$ depends not just on the isogeny class of G, but may depend on G up to isomorphism. Moreover, $\mathfrak{g}$ need not be perfect even when G is simple. In this section we record for later use some facts that do hold in this level of generality.

Lemma 3.1. *Let G be a simple algebraic group over k such that $(G, \operatorname{char} k) \neq (\mathrm{Sp}_{2n}, 2)$ for all $n \geq 1$. Put $\pi\colon \widetilde{G} \to G$ for the simply connected cover of G and $\tilde{\mathfrak{g}} := \operatorname{Lie}(\widetilde{G})$. Then*:

(1) $[\mathfrak{g}, \mathfrak{g}] = \mathrm{d}\pi(\tilde{\mathfrak{g}})$.

(2) *If V is an irreducible representation of G whose highest weight is restricted, then $V^{[\mathfrak{g},\mathfrak{g}]} = 0$.*

Proof. The map $\mathrm{d}\pi$ restricts to an isomorphism $\tilde{\mathfrak{g}}_\alpha \xrightarrow{\sim} \mathfrak{g}_\alpha$ for each root α, and in particular $\mathrm{d}\pi(\tilde{\mathfrak{g}}) \supseteq \langle \mathfrak{g}_\alpha \rangle$, an ideal in $\mathfrak{g}$. As $\mathfrak{g}/\langle \mathfrak{g}_\alpha \rangle$ is abelian, $\langle \mathfrak{g}_\alpha \rangle \supseteq [\mathfrak{g}, \mathfrak{g}]$.

Conversely, $[\tilde{\mathfrak{g}}, \tilde{\mathfrak{g}}] = \tilde{\mathfrak{g}}$; see [Premet 1997, Lemma 2.3(ii)] if $\operatorname{char} k$ is not special and [Hogeweij 1978, 6.13] in general. So $\mathrm{d}\pi(\tilde{\mathfrak{g}}) = \mathrm{d}\pi([\tilde{\mathfrak{g}}, \tilde{\mathfrak{g}}]) \subseteq [\mathfrak{g}, \mathfrak{g}]$.

To see (2), write the highest weight λ of V as a sum of fundamental dominant weights $\lambda = \sum c_i \omega_i$. As λ is restricted, there is some $c_i \in \mathbb{Z}$ whose image in k is not zero. Put α for the simple root such that $\langle \omega_i, \alpha^\vee \rangle = 1$. Writing $x_\alpha, x_{-\alpha}$ for basis elements of the root subalgebras for $\pm\alpha$ and v for a highest weight vector in V, we have $x_\alpha x_{-\alpha} v = \langle \lambda, \alpha^\vee \rangle v \neq 0$ as in the proof of [Steinberg 1963, Lemma 4.3(a)], so $V^{\mathrm{d}\pi(\tilde{\mathfrak{g}})} = V^{[\mathfrak{g},\mathfrak{g}]}$ is a proper submodule of V, and hence is zero. □

Corollary 3.2. *Let G be a simple algebraic group over k and put $\pi\colon \widetilde{G} \to G$ for the simply connected cover. If $\rho\colon G \to \mathrm{GL}(V)$ is a representation such that $\mathrm{d}\rho\,\mathrm{d}\pi = 0$ (i.e., $\tilde{\mathfrak{g}}$ acts trivially on V), then $\mathfrak{g}$ acts virtually freely on V.*

Proof. If G is simply connected — i.e., $G = \widetilde{G}$ (for example, if $G = \mathrm{Sp}_{2n}$) — then $\mathrm{d}\rho = 0$ and this is trivial. So assume G is not simply connected and apply Lemma 3.1. There is a torus T in G such that $\mathfrak{g} = [\mathfrak{g}, \mathfrak{g}] + \mathfrak{t}$ as a vector space. In particular, the images of $\mathfrak{g}$ and $\mathfrak{t}$ in $\mathfrak{gl}(V)$ are the same. The image consists of simultaneously diagonalizable matrices, so $\mathfrak{t}$ acts virtually freely, ergo the same is true for $\mathfrak{g}$. □

Example 3.3 (PGL_2). Let $\rho : \mathrm{PGL}_2 \to \mathrm{GL}(V)$ be an irreducible representation. The composition $\mathrm{SL}_2 \to \mathrm{PGL}_2 \xrightarrow{\rho} \mathrm{GL}(V)$ is an irreducible representation $L(w)$ of SL_2 as in Example 1.8 with w even. We claim that $\mathfrak{pgl}_2$ *fails to act virtually freely on V if and only* $\operatorname{char} k \neq 2$ *and* $w = 2$.

If $\operatorname{char} k \neq 2$, the induced map $\mathfrak{sl}_2 \to \mathfrak{pgl}_2$ is an isomorphism and the claim follows from Example 1.8.

If $\operatorname{char} k = 2$, then, as w is even, the representation of SL_2 is isomorphic to the Frobenius twist $L(w/2)^{[2]}$ and $\mathfrak{sl}_2$ acts trivially (and ρ is not faithful). By Corollary 3.2, the action of $\mathfrak{pgl}_2$ is virtually free, verifying the claim.

Suppose now additionally that $\mathrm{d}\rho$ is faithful, whence $\operatorname{char} k \neq 2$. Applying the above, we find that $\mathfrak{pgl}_2$ *fails to act generically freely if and only if $w = 2$ if and only if* $\dim V \leq \dim \mathrm{PGL}_2$.

Example 3.4 (adjoint representation). Let G be a simple algebraic group and put $L(\tilde{\alpha})$ for the irreducible representation with highest weight the highest root $\tilde{\alpha}$. It is a composition factor of the adjoint module.

If $\operatorname{char} k = 2$ and G has type C_n for $n \geq 1$ (including the cases $C_1 = A_1$ and $C_2 = B_2$), then $\mathfrak{g}$ acts virtually freely on $L(\tilde{\alpha})$. In case G is simply connected, $L(\tilde{\alpha})$ is a Frobenius twist of the natural representation of dimension $2n$ (since $\tilde{\alpha}$ is divisible by 2 in the weight lattice), so $\mathfrak{g}$ acts as zero (and, in particular, virtually freely); compare Example 1.8 (i) for the case $n = 1$. If G is adjoint, then we apply Corollary 3.2.

Now suppose that $\operatorname{char} k$ is not special for G and $(\text{type } G, \operatorname{char} k) \neq (A_1, 2)$. Put $\pi : \widetilde{G} \to G$ for the simply connected cover of G. The hypotheses give that $L(\tilde{\alpha}) \cong \tilde{\mathfrak{g}}/\mathfrak{z}(\tilde{\mathfrak{g}})$ as G-modules and that Cartan subalgebras of $\tilde{\mathfrak{g}}$ and $\mathfrak{g}$ are Lie algebras of maximal tori. It follows, then, that there is an open subset U of $\tilde{\mathfrak{g}}$ that meets $\operatorname{Lie}(\widetilde{T})$ for every maximal torus $\widetilde{T}$ of $\widetilde{G}$ such that for $a \in U$ the subalgebra $\operatorname{Nil}(a, \tilde{\mathfrak{g}}) := \bigcup_{m>0} \ker(\operatorname{ad} a)^m$ has minimal dimension (i.e., a is regular in the sense of [SGA 3_{II} 1970, §XIII.4]). Pick $a \in U \cap \widetilde{T}$, put $\bar{a} \in L(\tilde{\alpha})$ for the image of a, and set

$$\tilde{\mathfrak{g}}_{\bar{a}} = \{x \in \tilde{\mathfrak{g}} \mid \operatorname{ad}(x)\bar{a} \in \mathfrak{z}(\tilde{\mathfrak{g}})\}.$$

Then

$$\operatorname{Lie}(\widetilde{T}) \subseteq \tilde{\mathfrak{g}}_{\bar{a}} \subseteq \operatorname{Nil}(a, \tilde{\mathfrak{g}}) = \operatorname{Lie}(\widetilde{T}),$$

where the last equality is by [SGA 3_{II} 1970, Corollary XIII.5.4]. The image T of $\widetilde{T}$ in G is a maximal torus that fixes $\bar{a}$, so $\mathfrak{g}_{\bar{a}}$ is generated by $\operatorname{Lie}(T)$ and the root subgroups it contains. But any such root subgroup would be the image of the corresponding root subgroup of $\tilde{\mathfrak{g}}$, which does not stabilize $\bar{a}$, and therefore $\mathfrak{g}_{\bar{a}} = \operatorname{Lie}(T)$. In particular, $\mathfrak{g}$ does not act virtually freely on $L(\tilde{\alpha})$.

Lemma 3.5. *Suppose G is a simple algebraic group such that* $\operatorname{char} k$ *is not special and* $(G, \operatorname{char} k) \neq (\mathrm{SL}_2, 2)$. *If $\mathfrak{s}$ is a subalgebra of $\mathfrak{g}$ such that $\mathfrak{s} + \mathfrak{z}(\mathfrak{g}) \supseteq [\mathfrak{g}, \mathfrak{g}]$, then $\mathfrak{s} \supseteq [\mathfrak{g}, \mathfrak{g}]$.*

For the excluded case where $G = \mathrm{SL}_2$ and $\operatorname{char} k = 2$, we have that $\mathfrak{z}(\mathfrak{g}) = [\mathfrak{g}, \mathfrak{g}]$ is the Lie algebra of every maximal torus.

Proof. We may assume that $\mathfrak{z}(\mathfrak{g}) \neq 0$, and in particular the center of G is not étale and G does not have type A_1.

If G is equal to its simply connected cover $\widetilde{G}$, then for each $g \in G(k)$, there is $z_g \in \mathfrak{z}(\mathfrak{g})$ such that $z_g + gx_{\tilde{\alpha}} \in \mathfrak{s}$, where $\tilde{\alpha}$ denotes the highest root. Thus, $\mathfrak{s}$ contains $[z_g + gx_{\tilde{\alpha}}, z_{g'} + g'x_{\tilde{\alpha}}] = [gx_{\tilde{\alpha}}, g'x_{\tilde{\alpha}}]$ for all $g, g' \in G(k)$; hence $\mathfrak{s} = \mathfrak{g}$ by [Premet 1997, Lemma 2.3 (ii)].

Suppose now that $G \neq \widetilde{G}$. We may replace $\mathfrak{s}$ with $\mathfrak{s} \cap [\mathfrak{g}, \mathfrak{g}]$ and so assume $\mathfrak{s} \subseteq [\mathfrak{g}, \mathfrak{g}]$. Put $\tilde{\mathfrak{s}}$ for the inverse image $\mathrm{d}\pi^{-1}(\mathfrak{s})$ of $\mathfrak{s}$ in $\tilde{\mathfrak{g}}$ and $q : G \to \overline{G}$ for the natural map to the adjoint group. The kernels of $\mathrm{d}q\,\mathrm{d}\pi$ and $\mathrm{d}\pi$ are the centers of $\tilde{\mathfrak{g}}$ and $\mathfrak{g}$ respectively, so $\mathrm{d}q\,\mathrm{d}\pi(\tilde{s}) = \mathrm{d}q(\mathfrak{s}) \supseteq \mathrm{d}q([\mathfrak{g}, \mathfrak{g}])$ by hypothesis, which equals $\mathrm{d}q\,\mathrm{d}\pi(\tilde{\mathfrak{g}})$ by Lemma 3.1 (1). We are done by the case where G is simply connected. □

4. Deforming semisimple elements to nilpotent elements

For $x \in \mathfrak{g}$, we use the shorthand $x^{\mathbb{G}_m G}$ for the orbit of x under the subgroup of $\mathrm{GL}(\mathfrak{g})$ generated by $\mathbb{G}_m$ and $\mathrm{Ad}(G)$. For y in the closure of $x^{\mathbb{G}_m G}$, $\dim V^x \leq \dim V^y$ by upper semicontinuity of dimension.

Example 4.1. Suppose that $x \in \mathfrak{g}$ is noncentral semisimple and let $\mathfrak{b}$ be a Borel subalgebra containing x. Because x is not central, there is a root subgroup U_α in the corresponding Borel subgroup that does not commute with x. This implies that $x + \lambda y$ is in the same G-orbit as x for all $\lambda \in k$ and y in the corresponding root subalgebra, and similarly $\lambda x + y$ is in the same G-orbit as λx and in particular y is in the closure of $x^{\mathbb{G}_m G}$, so $\dim V^x = \dim V^{\lambda x + y} \leq \dim V^y$.

Lemma 4.2. *Suppose k is algebraically closed and let $G = \mathrm{GL}_n$ or SL_n. Let $x \in \mathfrak{g}$ be a semisimple element. Then there exists a nilpotent element $y \in \mathfrak{g}$ such that the following hold*:

(1) *The $\mathrm{Ad}(G)$-orbits of x and y have the same dimension.*

(2) *y is in the closure of $x^{\mathbb{G}_m G}$.*

(3) *If the matrix x has r distinct eigenvalues, then $y^{r-1} \neq 0$ and $y^r = 0$. In particular, if $p := \mathrm{char}\, k \neq 0$ and x is toral, then $y^{[p]} = 0$.*

(4) *The rank of y is the codimension of the largest eigenspace of x. In particular, if 0 is the eigenvalue of x with greatest multiplicity, then $\mathrm{rank}\, y = \mathrm{rank}\, x$.*

(5) *If V is a finite-dimensional rational G-module, then $\dim V^y \geq \dim V^x$ and $\dim y^G + \dim V^y \geq \dim x^G + \dim V^x$.*

Proof. Suppose first that $G = \mathrm{GL}_n$. We may assume that x is diagonal. Permuting the basis so that vectors with the same eigenvalue are adjacent, we may assume that x has $a_1, \ldots, a_r$ down the diagonal a_i appearing n_i times and $n_1 \geq n_2 \geq \cdots \geq n_r$. The centralizer of x in GL_n is $\prod_i \mathrm{GL}_{n_i}$ of dimension $\sum n_i^2$.

Let y be the block upper triangular matrix (with the blocks corresponding to the eigenspaces of x) such that the only nonzero blocks are the ones corresponding to the a_i, a_{i+1} block. In that block, take y to have 1's on the diagonal and 0's elsewhere; this block has rank n_{i+1} so $\mathrm{rank}\, y = \sum_{i=2}^r n_i$ as claimed in (4).

After conjugation by a permutation matrix, we deduce that the size of the Jordan blocks in the Jordan form of y are given by the partition of n conjugate to $(n_1, n_2, \ldots, n_r)$. The centralizer of such a matrix has dimension $\sum n_i^2$, cf. [Springer and Steinberg 1970, p. E-84, 1.7(iii); E-85, 1.8] or [Humphreys 1995, p. 14] and so (1) holds.

It follows that the largest Jordan block of y has size r whence the minimal polynomial of y has degree r (equal to the degree of the minimal polynomial of x).

Clearly, $x + ty \in x^G$ whence y is in the closure of $x^{\mathbb{G}_m G}$ and so $\dim V^x \leq \dim V^y$. This fact and (1) imply the last inequality in (5).

If x is toral, then x has all eigenvalues in $\mathbb{F}_p$ and so $r \leq p$, whence $y^{[p]} = 0$.

For $G = \mathrm{SL}_n$, each toral element is also toral in GL_n and one takes y as in the GL_n case. □

Generation.

Lemma 4.3. *Let $\rho\colon G \to \mathrm{GL}(V)$ be a representation of an algebraic group over a field k. Let X be an irreducible and G-invariant subset of $\mathfrak{g}$ such that X is open in $\overline{X}$. If, for some $Y \subseteq \overline{X}$, there exist $e > 0$ and $y_1, \ldots, y_e \in Y(k)$ that generate a subalgebra of $\mathfrak{g}$*

(1) *that has dimension at least d, for some d;*

(2) *that leaves no d-dimensional subspace of V invariant for some d;*

(3) *containing a G-invariant subalgebra M of $\mathfrak{g}$ such that M/N is an irreducible M-module and $\dim \mathfrak{g}/M < \dim M/N$, for some G-submodule N of M; or*

(4) *containing a strongly regular semisimple element (as defined in Example 5.1),*

then e generic elements of X do so as well.

We will use this lemma with $X = x^G$ and $Y = y^G$ for x and y. For a description of which nilpotent y lie in $\overline{x^G}$ for a given x, we refer to [Hesselink 1976, 3.10] for type A and, when $\operatorname{char} k \neq 2$, types B, C, and D. (A description can also be found in [Collingwood and McGovern 1993, §6.2].) For the other cases we use Lemma 11.2.

Alternatively, one can take x and y as in Example 4.1 or Lemma 4.2 and set $X = x^{\mathbb{G}_m G}$ and $Y = y^G$.

Proof of Lemma 4.3. For each of (1)–(4), we consider the subset U consisting of those $(y_1, \dots, y_e)$ in a product $\overline{X}^{\times e}$ of e copies of $\overline{X}$ that generate a subalgebra satisfying the given condition. Fix an $e > 0$ so that $U(k)$ is nonempty. It suffices to observe that U is open in $\overline{X}$, which is obvious for (1). Case (2) is argued as in [Breuillard et al. 2012, Lemma 3.6].

For (3), consider the set U' of $(y_1, \dots, y_e) \in \overline{X}^{\times e}$ such that $y_1, \dots y_e$ generate a subalgebra Q with Q acting irreducibly on M/N and $\dim Q \geq \dim M$; it is open as in (1) and (2). We claim that $U' = U$; the containment $\supseteq$ is clear. Conversely, if $(y_1, \dots, y_e)$ is in $U' \setminus U$, then $Q \cap M \subseteq N$ and $\dim Q \leq \dim \mathfrak{g}/M + \dim N < \dim M$, which is a contradiction.

For (4), the hypothesis is that some word w in variables is strongly regular semisimple for some collection of e elements of $Y(k)$. Since being strongly regular semisimple is an open condition, it follows that w is generically strongly regular semisimple. □

We also use the lemma in the form of the following corollary.

Corollary 4.4. *Let G be a simple algebraic group over a field k such that $\operatorname{char} k$ is not special for G and $(\text{type } G, \operatorname{char} k) \neq (A_1, 2)$. Let X be an irreducible and G-invariant subset of $\mathfrak{g}$ such that X is open in $\overline{X}$. If, for some $Y \subseteq \overline{X}$, there exist $e > 0$ and $y_1, \dots, y_e \in Y(k)$ that generate a subalgebra of $\mathfrak{g}$ containing $[\mathfrak{g}, \mathfrak{g}]$, then e generic elements of X do so as well.*

Proof. Set $M := \mathrm{d}\pi(\tilde{\mathfrak{g}}) = [\mathfrak{g}, \mathfrak{g}]$ (Lemma 3.1 (1)) and $N := \mathrm{d}\pi(\mathfrak{z}(\tilde{\mathfrak{g}})) = [\mathfrak{g}, \mathfrak{g}] \cap \mathfrak{z}(\mathfrak{g})$. Then M/N is, as a G-module, $L(\tilde{\alpha})$, an irreducible representation of M (Example 3.4). Moreover, $\dim \mathfrak{g}/M \leq \dim \mathfrak{z}(\tilde{\mathfrak{g}}) \leq 2 < \dim M/N$. Apply Lemma 4.3 (3). □

5. Quasiregular subalgebras

For this section, let T be a maximal torus in a reductive algebraic group G over an algebraically closed field k. Writing $\mathfrak{t} := \operatorname{Lie}(T)$ and $\mathfrak{g} := \operatorname{Lie}(G)$, the action of T on $\mathfrak{g}$ gives the Cartan decomposition $\mathfrak{g} = \mathfrak{t} \oplus \bigoplus_{\alpha \in \Phi} \mathfrak{g}_\alpha$, where Φ is the set of roots of G with respect to T and $\mathfrak{g}_\alpha$ is the 1-dimensional root

subalgebra for the root α. (Note that the action by $\mathfrak{t}$ induces a direct sum decomposition on $\mathfrak{g}$ that need not be as fine when $\operatorname{char} k = 2$, for in that case α and $-\alpha$ agree on $\mathfrak{t}$, and if furthermore $G = \mathrm{Sp}_{2n}$ for $n \geq 1$, then the centralizer of $\mathfrak{t}$ in $\mathfrak{g}$, the Cartan subalgebra, properly contains $\mathfrak{t}$.) We say that a subalgebra L of $\mathfrak{g}$ is *quasiregular with respect to* T if

$$L = (L \cap \mathfrak{t}) \oplus \begin{cases} \bigoplus_{\alpha \in \Phi} (L \cap \mathfrak{g}_\alpha) & \text{if } \operatorname{char} k \neq 2, \\ \bigoplus_{\alpha \in \Phi^+} (L \cap \mathfrak{g}_{\pm\alpha}) & \text{if } \operatorname{char} k = 2, \end{cases}$$

as a vector space, where $\mathfrak{g}_{\pm\alpha} := \mathfrak{g}_\alpha \oplus \mathfrak{g}_{-\alpha}$ and Φ^+ denotes the set of positive roots relative to some fixed ordering. We say simply that L is quasiregular if it is quasiregular with respect to some torus T.

For L quasiregular with respect to T, $\mathfrak{t}$ evidently normalizes L, i.e., $L + \mathfrak{t}$ is also a quasiregular subalgebra.

Example 5.1. Suppose there is a $t \in \mathfrak{t} \cap L$ such that

$$\pm\alpha(t) \neq \pm\beta(t) \quad \text{for all } \alpha \neq \beta \in \Phi^+ \cup \{0\}, \tag{5.2}$$

i.e., that has the same eigenspaces on $\mathfrak{g}$ as $\mathfrak{t}$. (We call such a t *strongly regular*.) Put $m(x)$ for the minimal polynomial of $\operatorname{ad}(t)$. For each $\alpha \in \Phi \cup \{0\}$, evaluating $m(x)/(x - \alpha(t))$ at $\operatorname{ad}(t)$ gives a linear map $\mathfrak{g} \to \mathfrak{g}$ with image $\mathfrak{g}_\alpha$ (if $\operatorname{char} k \neq 2$) or $\mathfrak{g}_{\pm\alpha}$ (if $\operatorname{char} k = 2$). Restricting t to L shows that $L \cap \mathfrak{g}_\alpha$ or $L \cap \mathfrak{g}_{\pm\alpha}$ is contained in L, i.e., *L is quasiregular.*

Example 5.3. Suppose $G = \mathrm{SL}_n$ or GL_n for $n \geq 4$. If L contains a copy of $\mathfrak{sl}_{n-1}$ (say, the matrices with zeros along the rightmost column and bottom row), then L is quasiregular. Indeed, taking T to be the diagonal matrices in G and $t \in \mathfrak{t}$ to have distinct indeterminates in the first $n-2$ diagonal entries and a zero in the last diagonal entry, we find that t satisfies (5.2). This L is quasiregular, but need not be regular, in the sense that it need not contain a maximal toral subalgebra of $\mathfrak{g}$.

Remark 5.4. Suppose $\operatorname{char} k \neq 2$ and $\mathfrak{g} = \mathfrak{sl}_n$, $\mathfrak{so}_n$, or $\mathfrak{sp}_{2n}$. *If L is a quasiregular Lie subalgebra and acts irreducibly on the natural module, then $L = \mathfrak{g}$.*

To see this, first suppose that L contains a maximal toral subalgebra $\mathfrak{t}$. Since $\operatorname{char} k \neq 2$, L is determined by $\mathfrak{t}$ and a closed subset of the root system of $\mathfrak{g}$, whose classification over k is the same as the Borel–de Siebenthal classification over $\mathbb{C}$. Now L cannot be contained in a maximal parabolic subalgebra, for such subalgebras act reducibly (even stabilizing a totally singular subspace for $\mathfrak{g} = \mathfrak{so}_n$ or $\mathfrak{sp}_{2n}$); see for example [Garibaldi and Carr 2006, §3]. Also, L cannot be contained in a semisimple subalgebra of maximal rank, since such subalgebras stabilize a nondegenerate subspace (compare for example [Dynkin 1952, Table 9]) and the claim follows. (This shows also that if $\operatorname{char} k \neq 2$ and $L \subseteq \mathfrak{gl}_n$ is a subalgebra that acts irreducibly and contains a maximal toral subalgebra, then $L = \mathfrak{gl}_n$.)

In the general case, let T be the torus with respect to which L is quasiregular. As $L + \mathfrak{t} = \mathfrak{g}$ by the preceding paragraph, every root subalgebra occurs in L, and we conclude that $L = \mathfrak{g}$ (compare the proof of Lemma 3.1).

See [Breuillard et al. 2012, Lemma 3.6] for a similar statement on the level of groups.

The subsystem subalgebra. Suppose L is a quasiregular subalgebra of $\mathfrak{g}$ with respect to T. Define L_0 to be the subalgebra of L generated by the $L \cap \mathfrak{g}_\alpha$ for $\alpha \in \Phi$.

Lemma 5.5. *If*

(1) $\operatorname{char} k \neq 2$ *or*

(2) $\operatorname{char} k = 2$, Φ *is irreducible, and all roots have the same length,*

then L_0 is an ideal in $L + \mathfrak{t}$.

Proof. If $\operatorname{char} k \neq 2$, then $L_0 \cap \mathfrak{g}_\alpha = L \cap \mathfrak{g}_\alpha$ for all α and the claim is trivial, so assume (2) holds. As L_0 is evidently stable under $\operatorname{ad} \mathfrak{t}$, it suffices to check, for $x_\beta \in \mathfrak{g}_\beta$, $x_{-\beta} \in \mathfrak{g}_{-\beta}$, $c \in k$ such that $x_\beta + cx_{-\beta} \in L$, and $x_\alpha \in L \cap \mathfrak{g}_\alpha \subseteq L_0$, that L_0 contains

$$[x_\beta + cx_{-\beta}, x_\alpha] = [x_\beta, x_\alpha] + c[x_{-\beta}, x_\alpha].$$

However, by hypothesis $\alpha + \beta$ and $\alpha - \beta$ cannot both be roots, so at least one of the two terms in the displayed sum is zero and the expression belongs to $L \cap \mathfrak{g}_{\alpha+\beta}$ or $L \cap \mathfrak{g}_{\alpha-\beta}$, hence to L_0. □

Example 5.6. Let L be the space of symmetric n-by-n matrices in $\mathfrak{gl}_n$. It is a Lie subalgebra when $\operatorname{char} k = 2$, and, in that case, it is quasiregular with respect to the maximal torus T of diagonal matrices in GL_n and $L_0 = 0$.

Lemma 5.7. *Suppose L is a quasiregular subalgebra of $\mathfrak{gl}(V)$ with respect to a maximal torus T. Then L_0 is irreducible on V if and only if $L_0 + \mathfrak{t}$ is irreducible on V if and only if $L_0 = \mathfrak{sl}(V)$, if and only if $L_0 + \mathfrak{t} = \mathfrak{gl}(V)$.*

Proof. The algebra L_0 is $(L_0 \cap \mathfrak{t}) \oplus \bigoplus_{\alpha \in S} \mathfrak{g}_\alpha$ where S is a closed subsystem of a root system of type A. Therefore $S = \Phi$ (in which case L_0 acts irreducibly and $L_0 = \mathfrak{sl}(V)$) or S is contained in a proper subsystem (which normalizes a proper T-invariant subspace of V). □

Application to type A.

Theorem 5.8. *Suppose L is a subalgebra of $\mathfrak{gl}_n$ for some $n \geq 2$ that is quasiregular and acts irreducibly on the natural representation of $\mathfrak{gl}_n$. Then*

(1) *L contains $\mathfrak{sl}_n$, or*

(2) $\operatorname{char} k = 2$ *and L is GL_n-conjugate to a subalgebra of symmetric n-by-n matrices containing the alternating matrices.*

Proof. Let T be the maximal torus with respect to which L is quasiregular. After conjugation by an element of $\mathrm{GL}_n(k)$, we may assume that T is the diagonal matrices. If $L_0 = L$ or even $L_0 + \mathfrak{t} = \mathfrak{gl}_n$, Lemma 5.7 gives that L contains $\mathfrak{sl}_n$.

Case: $L_0 \neq 0$. Suppose $L_0 \neq 0$. We claim that (1) holds. Replacing L with $L + \mathfrak{t}$, we may assume that $\mathfrak{t} \subseteq L$. We claim that $L_1 := L_0 + \mathfrak{t}$ acts irreducibly on the natural representation $V := k^n$ of $\mathfrak{gl}_n$. If $\dim V = 2$, the result is clear. So we assume that $\dim V \geq 3$.

Suppose $W \subseteq V$ is a subspace on which L_1 acts nontrivially and irreducibly. Conjugating by a monomial matrix, we may assume that W is the subspace consisting of vectors whose nonzero entries are in the first $w := \dim W$ coordinates. If $w = 1$, we can apply the graph automorphism that inverts T and permutes the root spaces and get a possibly different subalgebra L' which leaves invariant a hyperplane. Of course, it suffices to prove the result for L' and so we may take $w \geq 2$. Now $L_1 \cap \mathfrak{gl}(W)$ is a quasiregular subalgebra of $\mathfrak{gl}(W)$ acting irreducibly on W and it is generated by $\mathfrak{t} \cap \mathfrak{gl}(W)$ and those $\mathfrak{g}_\alpha$ contained in L, so by Lemma 5.7 it equals $\mathfrak{gl}(W)$.

If $W \neq V$, then there is a $\beta \in \Phi$ such that $\mathfrak{g}_{\pm\beta} \cap L_0 = 0$ yet $(\mathfrak{g}_{\pm\beta} \cap L)W \not\subseteq W$. That is, there exists $i > w$ and $j \leq w$ such that $E_{ij} - cE_{ji} \in L$ for some $c \in k^\times$, where E_{ij} denotes the matrix whose unique nonzero entry is a 1 in the (i, j)-entry. As $\dim W \geq 2$, there is $\ell \leq w$, $\ell \neq j$ and $E_{\ell j} \in \mathfrak{sl}(W) \subseteq L_0$. So $[E_{\ell j}, E_{ij} - cE_{ji}] = -E_{i\ell}$ is in L, and hence in L_0, yet $E_{i\ell}W \not\subseteq W$, which is a contradiction. Thus $W = V$, i.e., L_1 acts irreducibly on V and $L_0 = \mathfrak{sl}_n$.

Case: $L_0 = 0$. Suppose $L_0 = 0$. If $\operatorname{char} k \neq 2$, then $L + \mathfrak{t}$ cannot be irreducible (Remark 5.4), so assume $\operatorname{char} k = 2$. We prove (2).

Define $\widehat{L}$ to be the subspace of $\mathfrak{gl}(V)$ generated by $\mathfrak{t}$ and those $\mathfrak{g}_{\pm\alpha}$ with nonzero intersection with L. It is closed under the bracket. Indeed, fixing nonzero elements $x_\alpha \in \mathfrak{g}_\alpha$ for all $\alpha \in \Phi$, those $\mathfrak{g}_{\pm\alpha}$ that meet L are spanned by an element $x_\alpha + c_\alpha x_{-\alpha}$ for some $c_\alpha \in k^\times$. If $\mathfrak{g}_{\pm\beta}$ also meets L, then

$$[x_\alpha + c_\alpha x_{-\alpha}, x_\beta + c_\beta x_{-\beta}] \in \mathfrak{g}_{\pm(\alpha+\beta)} + \mathfrak{g}_{\pm(\alpha-\beta)},$$

whence the element on the left belongs to $\widehat{L}$ because at most one of $\alpha + \beta$, $\alpha - \beta$ is a root. As L acts irreducibly on V, so does $\widehat{L}$, and Lemma 5.7 gives that $\widehat{L} = \mathfrak{gl}_n$ and in particular $\mathfrak{g}_{\pm\alpha}$ meets L for every root α.

For each simple root α_i, set $h_i \colon \mathbb{G}_{\mathrm{m}} \to \mathrm{GL}_n$ to be a cocharacter such that $\alpha_j \circ h_i \colon \mathbb{G}_{\mathrm{m}} \to \mathbb{G}_{\mathrm{m}}$ is $t \mapsto 1$ if $i \neq j$ and $t \mapsto t^{r_i}$ for some $r_i \neq 0$ if $i = j$. As

$$\operatorname{Ad}(h_i(t))(x_{\alpha_i} + c_{\alpha_i} x_{-\alpha_i}) = t^{r_i} x_{\alpha_i} + \frac{c_{\alpha_i}}{t^{r_i}} x_{-\alpha_i},$$

there is a $t_i \in k^\times$ for each i so that $\operatorname{Ad}(h_i(t_i))(\mathfrak{g}_{\pm\alpha_i} \cap L)$ is generated by $E_{i,i+1} + E_{i+1,i}$. Conjugating L by $\prod h_i(t_i)$ arranges this for all simple roots α_i at once, and it follows that the resulting conjugate of L consists of symmetric matrices and intersects $\mathfrak{g}_{\pm\alpha}$ nontrivially for all $\alpha \in \Phi$, whence L contains the space of alternating matrices. □

6. Type *A* and char $k \neq 2$

Recall that $\mathfrak{sl}_n$ for $n \geq 2$ is either simple ($\operatorname{char} k$ does not divide n) or has a unique nontrivial ideal, the center (consisting of the scalar matrices, in case $\operatorname{char} k$ does divide n).

The next two items have no restrictions on the characteristic of k. We do not need the first result in characteristic 2.

Example 6.1. Suppose that x is regular nilpotent in $\mathfrak{sl}_n$ for some $n \geq 2$; we claim that $e(x) = 2$, i.e., 2 generic $\mathrm{SL}_n(k)$-conjugates of x generate $\mathfrak{sl}_n$. Up to conjugacy, x has 1's on the superdiagonal and 0's in all other entries. Choose a conjugate y of x whose only nonzero entries are $x_2, \dots, x_n$ on the subdiagonal. Then $w := [x, y]$ is diagonal with entries $z_1, \dots, z_n$ where $(z_1, \dots, z_n) = (-x_2, x_2 - x_3, \dots, x_{n-1} - x_n, x_n)$. For a nonempty open subvariety of $(x_2, \dots, x_n)$ the $z_i - z_j$ are distinct. Thus, the algebra generated by w and x contains all the positive simple root algebras and similarly the algebra generated by w and y contains all the negative simple root algebras, whence $\langle x, y \rangle = \mathfrak{sl}_n$. Since the condition on generating $\mathfrak{sl}_n$ is open (Lemma 4.3 (1)), this implies that 2 generic conjugates of x generate $\mathfrak{sl}_n$.

For $x \in \mathfrak{sl}_n$, put $\alpha(x)$ for the dimension of the largest eigenspace.

Lemma 6.2. *For noncentral $x \in \mathfrak{sl}_n$ with $n \geq 2$, if $e > (n-1)/(n-\alpha(x))$, then the subalgebra of $\mathfrak{sl}_n$ generated by e generic conjugates of x fixes no 1-dimensional subspace nor codimension-1 subspace of the natural module.*

The hypothesis that x is noncentral ensures that the denominator $n - \alpha(x)$ is not zero.

Proof. Suppose that the subalgebra generated by e generic conjugates of x fixes a line. Then, by Lemma 4.3 (2), every subalgebra generated by e conjugates fixes a line. Putting $X := x^{\mathrm{SL}_n}$, there is a map $G \times (\times^e X) \to \times^e X$ via $(g, x_1, \dots, x_e) \mapsto (\mathrm{Ad}(g)x_1, \dots, \mathrm{Ad}(g)x_e)$, and by hypothesis $\times^e X$ belongs to the image of $G \times (\times^e (X \cap \mathfrak{p}))$ where $\mathfrak{p}$ is the stabilizer of the first basis vector in the natural module, the Lie algebra of a parabolic subgroup P of $\mathfrak{sl}_n$. Thus

$$e \cdot \dim X \leq \dim \mathbb{P}^1 + e \cdot \dim(X \cap \mathfrak{p}),$$

and consequently

$$e(\dim X - \dim(X \cap \mathfrak{p})) \leq \dim(G/P) = n - 1. \tag{6.3}$$

Now consider the variety $Y \subset X \times \mathbb{P}^{n-1}$ with k-points

$$Y(k) = \{(y, \omega) \in X(k) \times \mathbb{P}(k^n) \mid y\omega = \omega\}.$$

The projection of Y on the first factor maps Y onto X with fibers of dimension $\alpha(x) - 1$. The projection of Y on the second factor maps Y onto $\mathbb{P}^{n-1}$ with fibers of dimension $\dim(X \cap \mathfrak{p})$. Consequently,

$$\dim X + \alpha(x) - 1 = \dim Y = (n-1) + \dim(X \cap \mathfrak{p}).$$

Combining this with (6.3) gives $e \leq (n-1)/(n-\alpha(x))$.

Now suppose each subalgebra $\mathfrak{g}$ generated by e generic conjugates of x fixes a codimension-1 subspace V of the natural module. Using the dot product we may identify the natural module k^n with its contragradient $(k^n)^*$, and it follows that the subalgebra $\{y^\top \mid y \in \mathfrak{g}\}$ fixes the line in $(k^n)^*$ of elements vanishing on V. Consequently $e \leq (n-1)/(n-\alpha(x^\top))$. As $\alpha(x^\top) = \alpha(x)$, the claim is proved. □

Proposition 6.4. *Assume* $\operatorname{char} k \neq 2$. *For each nonzero nilpotent* $x \in \mathfrak{sl}_n$ *with* $n \geq 3$, *e generic conjugates of* x *generate* $\mathfrak{sl}_n$, *where*:

(1) $e = 3$ *if* x *has Jordan canonical form with partition* $(2, 2, \ldots, 2)$ *or* $(2, 2, \ldots, 2, 1)$.

(2) $e = 2$ *if* $\alpha(x) \leq \lceil n/2 \rceil$ *but we are not in case (1).*

(3) $e = \lceil n/(n - \alpha(x)) \rceil$ *if* $\alpha(x) > \lceil n/2 \rceil$.

Proof. The conjugacy class of x is determined by its Jordan form, which corresponds to a partition $(p_1, \ldots, p_\alpha)$ of n, i.e., a list of numbers $p_1 \geq p_2 \geq \cdots \geq p_\alpha > 0$ such that $p_1 + \cdots + p_\alpha = n$. If x has partition (n), then $e(x) = 2$ by Example 6.1.

If x has partition $(2, 1, \ldots, 1)$, i.e., the Jordan form of x has a unique nonzero entry, then x generates a root subalgebra, and we may assume it corresponds to a simple root. The other root subalgebras for simple roots and for the lowest root suffice to generate $\mathfrak{sl}_n$, so in this case $n = \lceil n/(n - (n-1)) \rceil$ conjugates suffice to generate.

Thus we may assume that $n \geq 4$.

Suppose first that x has partition $(2, 2, \ldots, 2)$ and view x as the image of a regular nilpotent in $\mathfrak{sl}_2$ under the diagonal embedding in $\mathfrak{sl}_2^{\times n/2} \subset \mathfrak{sl}_n$. As in Example 6.1, two $\mathrm{SL}_2^{\times n/2}$-conjugates suffice to generate $\mathfrak{sl}_2^{\times n/2}$. As the adjoint representation of $\mathfrak{sl}_n$ restricts to a multiplicity-free representation of $\mathfrak{sl}_2^{\times n/2}$, there are only a finite number of Lie algebras lying between $\mathfrak{sl}_2^{\times n/2}$ and $\mathfrak{sl}_n$. Now x^{SL_n} generates $\mathfrak{sl}_n$ as a Lie algebra, so it is not contained in any of these proper subalgebras and the irreducible variety x^{SL_n} is not contained in the union of the proper subalgebras. This proves the claim that 3 conjugates suffice to generate $\mathfrak{sl}_n$.

If x has partition $(2, 2, \ldots, 2, 1)$, then we view it as the image of $x' \in \mathfrak{sl}_{n-1}$ where x' has partition $(2, 2, \ldots, 2)$, for which three SL_{n-1}-conjugates generate $\mathfrak{sl}_{n-1}$. That is, three generic SL_n-conjugates of x generate a subalgebra $\mathfrak{h}$ that is quasiregular (Example 5.3). Moreover, as $n = 2\alpha - 1$, $\mathfrak{h}$ does not fix a 1-dimensional or codimension-1 subspace of the natural module (Lemma 6.2), and therefore $\mathfrak{h}$ acts irreducibly and $\mathfrak{h}$ is the whole algebra $\mathfrak{sl}_n$ (Remark 5.4).

Now suppose $\alpha(x) \leq n/2$ and we are not in case (1). Then $p_1 \geq 3$ and by passing to a nilpotent element in the closure of x^{SL_n} as in Section 4, we can reduce to the cases

(a) n is even and x has partition $(3, 2, \ldots, 2, 1)$; or

(b) n is odd and x has partition $(3, 2, \ldots, 2)$.

In case (a), we see by induction that we can generate $\mathfrak{sl}_{n-1}$ with two SL_n-conjugates and we argue as in the preceding case.

In case (b), deform to $y \in \overline{x^{\mathrm{SL}_n}}$ with partition $(3, 2, \ldots, 2, 1, 1)$. It is the image of $y' \in \mathfrak{sl}_{n-1}$ with partition $(3, 2, \ldots, 2, 1)$. By induction on n, two SL_{n-1}-conjugates of y' generate a copy of $\mathfrak{sl}_{n-1}$. Arguing as in the preceding cases concludes the proof of (2).

Finally, suppose $\alpha(x) > \lceil n/2 \rceil$, so in particular $p_\alpha = 1$. Put $x' \in \mathfrak{sl}_{n-1}$ for a nilpotent with partition $(p_1, \ldots, p_{\alpha-1})$. By induction, we find that $\lceil n/(n-\alpha) \rceil$ SL_{n-1}-conjugates suffice to generate a copy of $\mathfrak{sl}_{n-1}$, and we complete the proof as before. □

Corollary 6.5. *Suppose* $\operatorname{char} k \neq 2$. *For noncentral* $x \in \mathfrak{gl}_n$ *with* $n \geq 2$ *such that* $x^{[p]} \in \{0, x\}$, *there exist* $e > 0$ *and elements* $x_1, \ldots, x_e \in x^{\mathrm{SL}_n}$ *that generate a subalgebra containing* $\mathfrak{sl}_n$ *such that* $e \cdot \dim x^{\mathrm{SL}_n} \leq \frac{9}{4}n^2$.

Proof. Suppose first that $x^{[p]} = 0$. If $n = 2$, then $e(x) = 2$ by Example 6.1 and $\dim x^{\mathrm{SL}_n} = 2$, so assume that $n \geq 3$. We consider the three cases in Proposition 6.4. In case (1), we have $\dim x^{\mathrm{SL}_n} \leq n^2/2$ and $e(x) = 3$, so the claim is clear. In case (2), $e = 2$ and $\dim x^{\mathrm{SL}_n} < n^2$. In case (3), among those nilpotent y with rank $n - \alpha(x)$, the one with the largest SL_n-orbit has partition $(n - \alpha(x) + 1, 1, \ldots, 1)$, whose orbit has dimension $n^2 - n - \alpha(x)^2 + \alpha(x)$. Consequently,

$$e(x) \cdot \dim x^{\mathrm{SL}_n} < (n + \alpha(x) - 1)(2n - \alpha(x)).$$

This is a quadratic polynomial in $\alpha(x)$ opening downwards with maximum at $(n + 1)/2$. As $\alpha \geq \lceil n/2 \rceil + 1 \geq n/2 + 1$, the right side is no larger than $\frac{9}{4}n^2 - 3n/2$ verifying the claim for x nilpotent.

For $x \in \mathfrak{sl}_n$ noncentral toral, let y be the nilpotent element provided by Lemma 4.2. Then $\dim x^G = \dim y^G$ and the same number of conjugates suffice to generate a subalgebra containing $\mathfrak{sl}_n$, as in Lemma 4.3 (3) with $M = \mathfrak{sl}_n$ and $N = \mathfrak{z}(\mathfrak{sl}_n)$. □

7. Type A and char $k = 2$

Proposition 7.1. *Suppose* $\operatorname{char} k = 2$ *and let* $x \in \mathfrak{sl}_n$ *with* $n \geq 2$ *be a nilpotent element of square* 0 *and rank* r. *Then* $\mathfrak{sl}_n$ *can be generated by* $e := \max\{3, \lceil n/r \rceil\}$ *conjugates of* x.

Proof. Note the result is clear if x is a root element by taking root elements in each of the simple positive root subalgebras and in the root subalgebra corresponding to the negative of the highest root. This gives the result for $n = 2, 3$ and shows that for $n = 4$, it suffices to consider $r = 2$. Choose two conjugates of x and y generating $\mathfrak{sl}_2 \times \mathfrak{sl}_2$. It is straightforward to see for a generic conjugate z of x, the elements x, y and z generate $\mathfrak{sl}_4$. So assume $n > 4$.

If n is odd, it follows by induction on n that e conjugates of x can generate an $\mathfrak{sl}_{n-1}$. On the other hand, the condition on the rank implies by Lemma 6.2 that e generic conjugates of x do not fix a 1-space or a hyperplane. Thus, generically e conjugates of x generate a subalgebra that acts irreducibly (as in the proof of Lemma 6.2) and is quasiregular by Example 5.3. Also, we see that generically the dimension of the Lie algebra generated by e conjugates has dimension at least $(e - 1)^2 - 1$. Since $n > 4$, this is larger than the dimension of the space of symmetric matrices, whence by Theorem 5.8, we see that e generic conjugates generate $\mathfrak{sl}_n$.

Now assume that n is even. By passing to closures we may assume that $r < n/2$ (since $n > 4$, $e = 3$ for both elements of rank $n/2$ and rank $n/2 - 1$). Now argue just as for the case that n is odd. □

Remark. The result also holds for idempotents (i.e., toral elements) of rank $e \leq n/2$ by a closure argument.

Corollary 7.2. *Suppose that* $\operatorname{char} k = 2$. *For noncentral* $x \in \mathfrak{gl}_n$ *with* $n \geq 2$ *such that* $x^{[2]} \in \{0, x\}$, *there exist* $e > 0$ *and elements* $x_1, \ldots, x_e \in x^{\mathrm{SL}_n}$ *that generate a subalgebra containing* $\mathfrak{sl}_n$ *such that* $e \cdot \dim x^{\mathrm{SL}_n} \leq 2n^2 - 2$.

Proof. Let $x \in \mathfrak{sl}_n \setminus \mathfrak{z}(\mathfrak{sl}_n)$ satisfy $x^{[2]} = 0$ and put r for the rank of x. Then $\dim x^{\mathrm{SL}_n} = n^2 - (r^2 + (n-r)^2) = 2r(n-r)$. If 3 conjugates of x generate $\mathfrak{sl}_n$, then $3 \cdot \dim x^{\mathrm{SL}_n} = 6r(n-r)$. This has a maximum at $r = n/2$, where it is $\frac{3}{2}n^2 \le 2n^2 - 2$. Otherwise $(n+r)/r$ conjugates suffice to generate, and we have $e \dim x^{\mathrm{SL}_n} \le 2(n^2 - r^2) \le 2n^2 - 2$.

Now suppose that $x \in \mathfrak{gl}_n$ is noncentral toral. Take $y \in \overline{x^{\mathbb{G}_{\mathrm{m}}\,\mathrm{GL}_n}}$ such that $y^{[2]} = 0$ as in Lemma 4.2, so $\dim y^{\mathrm{SL}_n} = \dim x^{\mathrm{SL}_n}$. Applying Lemma 4.3 (3) with $M = \mathfrak{sl}_n$ and $N = \mathfrak{z}(\mathfrak{sl}_n)$ gives that $e \cdot \dim x^{\mathrm{SL}_n} \le 2n^2 - 2$ also in this case. □

8. Type *C* and char $k \neq 2$

Proposition 8.1. *Assume* $\operatorname{char} k \neq 2$. *For every nonzero nilpotent* $x \in \mathfrak{sp}_{2n}$ *for* $n \ge 1$ *of rank* r, e *generic conjugates of* x *generate* $\mathfrak{sp}_{2n}$, *where*:

(1) $e = 3$ *if* x *has Jordan canonical form with partition* $(2, 2, \ldots, 2)$.

(2) $e = 2$ *if* $r \ge n$ *but we are not in case (1).*

(3) $e = 2\lceil n/r \rceil$ *if* $r < n$.

Proof. The conjugacy class of x is determined by its Jordan form, which corresponds to a partition $(p_1, \ldots, p_\alpha)$ of $2n$ with $p_1 \ge p_2 \ge \cdots \ge p_\alpha$ such that odd numbers appear with even multiplicity. Note that $\mathfrak{sp}_2 = \mathfrak{sl}_2$, so the $n = 1$ case holds by Example 6.1.

By specialization (replacing x with an element of $\overline{x^{\mathrm{Sp}_{2n}}}$ as in Section 4), we may replace in the partition of x

$$(2s+2, 1, 1) \rightsquigarrow (s+1, s+1, 2) \text{ or } (2s+1, 2s+1, 1, 1) \rightsquigarrow (2s, 2s, 2, 2) \text{ for } s \ge 2 \tag{8.2}$$

without changing the rank r of x nor whether the partition is $(2, \ldots, 2)$. In this way, we may assume that $p_\alpha \ge 2$ or $p_1 \le 4$.

Case (1). Suppose that x has partition $(2, 2, \ldots, 2)$. Two conjugates of x suffice to generate a copy of $\mathfrak{sl}_2^{\times n} \subset \mathfrak{sp}_{2n}$, and this contains a regular semisimple element of $\mathfrak{sp}_{2n}$. Furthermore, the natural representation of $\mathfrak{sp}_{2n}$ is multiplicity-free for $\mathfrak{sl}_2^{\times n}$, so one further conjugate suffices to produce a subalgebra that is irreducible on the natural module. Appealing to Remark 5.4, the claim follows in this case.

Case $\mathfrak{sp}_4$. For the case $n = 2$, it remains to consider x with partition (4), i.e., a regular nilpotent. A pair of generic conjugates generates an irreducible subalgebra. By passing to $(2, 2)$, we see it also generically contains an element as in (5.2), whence the result.

Case $\mathfrak{sp}_6$. Suppose $x \in \mathfrak{sp}_6$; it suffices to assume that x has rank at least 3 and $p_1 \ge 3$. We want to show that two conjugates of x can generate. By passing to closures, it suffices to assume that x is nilpotent with partition $(4, 1, 1)$. As in (8.2), the closure of the class of x contains the class corresponding to the partition $(2, 2, 2)$. Since two conjugates of the latter can generate an $\mathfrak{sl}_2 \times \mathfrak{sl}_2 \times \mathfrak{sl}_2$, we see via Lemma 4.3 (4) that generically two conjugates of x generate a Lie algebra containing a strongly regular semisimple element and so a quasiregular algebra.

By the $\mathfrak{sp}_4$ case, we see that generically the largest composition factor of the algebra generated by two conjugates of x is at least 4-dimensional and, by the paragraph above, the smallest is at least 2-dimensional. Thus, for generic $y \in x^G$, the subalgebra $\langle x, y\rangle$ generated by x and y is either irreducible or the module is a direct sum of nondegenerate spaces of dimension 4 and 2. However, this would imply that x and y would be trivial on the two dimensional space, which is a contradiction. Thus, a generic pair of conjugates of x and y generates an irreducible quasiregular subalgebra. Since we are in characteristic different from 2, this implies that generically $\langle x, y\rangle = \mathfrak{sp}_6$ as required.

Case $2n \geq 8$ *and* x *has partition* $(3, 3, 2, \ldots, 2, 1, 1)$. Suppose that x has partition $(3, 3, 2, \ldots, 2, 1, 1)$ so $r = n$. By the type A case (Proposition 6.4), 2 conjugates of x suffice to generate an $\mathfrak{sl}_n$ subalgebra and so generically our algebra contains a strongly regular semisimple element and also generically the smallest invariant subspace has dimension at least n. By induction, 2 conjugates of x can generate an $\mathfrak{sp}_{2n-2}$ subalgebra, so generically there is an irreducible submodule of dimension at least $2n-2$. Thus, generically the algebra is irreducible and contains strongly regular elements whence by Remark 5.4 is $\mathfrak{sp}_{2n}$.

Case $r \geq n$. We now consider the case where $r \geq n$ (and $2n \geq 8$).

If $p_\alpha \geq 2$, then, as $\alpha = 2n - r \leq n$ and we are not in case (1), we may replace $2s \rightsquigarrow (s, s)$ for $s \geq 3$, $(s, s) \rightsquigarrow (s-1, s-1, 1, 1)$ for $s \geq 4$, or $(4, 2) \rightsquigarrow (3, 3)$ as long as we retain the property that rank $x \geq n$. In this way, we may assume that $p_\alpha \leq 1$ or $p_1 \leq 3$.

So suppose $p_\alpha = 1$, in which case we may assume that $p_1 \leq 4$. We may replace $(4, 4, 1, 1) \rightsquigarrow (3, 3, 2, 2)$, $(4, 2) \rightsquigarrow (3, 3)$, or $(4, 3, 3, 1, 1) \rightsquigarrow (3, 3, 2, 2, 2)$ without changing the rank of x. Repeating these reductions and those in the previous paragraph, we are reduced to considering partitions $(4, 1, \ldots, 1)$ of rank 3 (excluded because $r \geq n \geq 4$) or $p_1 = 3$.

If there are at least four 3's, we substitute $(3^4, 1^2) \rightsquigarrow (3^2, 2^3, 1^2)$ if $p_\alpha = 1$ or $(3^4) \rightsquigarrow (3^2, 2^3)$ if $p_\alpha > 1$. Thus we may assume that x has partition $(3^2, 2^{r-4}, 1^t)$. As $2r \geq 2n = 2r - 2 + t$, we find that x has partition $(3^2, 2^{r-4}, 1^2)$ with $r = n$ (in which case the proposition has already been proved) or partition $(3^2, 2^{r-4})$ with $r = n + 1$, which specializes to the previous case.

Case $r < n$. Now suppose that x has rank $r < n$, so in particular $p_\alpha = 1$ and we may assume that $p_1 \leq 4$. Specializing as in (8.2) also with $s = 1$, we may assume that x has partition $(2^r, 1^{2n-2r})$. If $r = 1$, then $2n$ conjugates suffice to generate $\mathfrak{sp}_{2n}$ by, for example, [Cohen et al. 2001]. So assume $r \geq 2$.

Clearly, $n/r \leq n/2 < n - r$, so there are at least $2v+2$ 1-by-1 Jordan blocks in x for $e := 2\lceil n/r \rceil = 2v+2$. We then subdivide x into two blocks on the diagonal, with partitions $(2, 1^{2v})$ and $(2^{r-1}, 1^{2n-2r-2v})$. By the $r = 1$ case, e generic conjugates of the first generate an $\mathfrak{sp}_e$ subalgebra and by induction $\max\{3, 2\lceil (n-v-1)/(r-1)\rceil\}$ conjugates of the second generate an $\mathfrak{sp}_{2n-e}$ subalgebra. As $2n \leq re$, we have $(n-v-1)/(r-1) \leq n/r$, and the max in the preceding sentence is at most e. Note that $\mathfrak{sp}_e \times \mathfrak{sp}_{2n-e}$ contains a regular semisimple element of $\mathfrak{sp}_{2n}$ and the natural module has composition factors of size e, $2n - e$.

Alternatively, we may subdivide x into blocks with partitions $(2^r, 1^{2n-2r-2})$ and (1^2). By induction, e generic conjugates of this element give an $\mathfrak{sp}_{2n-2}$ subalgebra, with composition factors of size 1, 1, $2n - 2$. As this list does not meet the list of composition factors from the previous paragraph, the generic

subalgebra generated by e conjugates acts irreducibly on the natural module, and we are done via an application of Remark 5.4. □

Corollary 8.3. *Assume* $\operatorname{char} k \neq 2$. *For nonzero nilpotent or noncentral semisimple* $x \in \mathfrak{sp}_{2n}$ *with* $n \geq 1$, *there exist* $e > 0$ *and elements* $x_1 \ldots, x_e \in x^{\mathrm{Sp}_{2n}}$ *that generate* $\mathfrak{sp}_{2n}$ *such that* $e \cdot \dim x^{\mathrm{Sp}_{2n}} \leq 6n^2$.

Proof. Note that we are done if 3 conjugates of x suffice to generate $\mathfrak{sp}_{2n}$, as $\dim x^G \leq 2n^2$. Moreover, the case $n = 1$ holds by Corollary 6.5.

Recall that $\alpha(x)$ is the dimension of the largest eigenspace of x (and so for x nilpotent, the rank of x is $2n - \alpha(x)$).

Nilpotent case. Suppose that x is nonzero nilpotent and put $e(x)$ for the minimal number of conjugates of x needed to generate $\mathfrak{sp}_{2n}$. We may assume that $e(x) > 3$ and so $r < n$. In particular, $\alpha(x) > n$.

We have $e(x) \leq 2\lceil n/(2n-\alpha(x))\rceil$ by Proposition 8.1. To bound $\dim x^G$, we replace x with y such that $\alpha(y) = \alpha(x)$ and y specializes to x, i.e., x belongs to the closure of $y^{\mathbb{G}_m G}$. Then $\mathfrak{sp}_{2n}$ is also generated by $2\lceil n/(2n-\alpha(x))\rceil$ conjugates of y and $\dim x^G \leq \dim y^G$. The element x is given by a partition $(p_1, \ldots, p_\alpha)$ as in the proof of Proposition 8.1.

We claim that y can be taken to have partition $(2s, 2, 1^{\alpha(x)-2})$ or $(2s, 1^{\alpha(x)-1})$. Indeed, let $I := \{i \mid i > 1 \text{ and } p_i > 2\}$. Then the element y with partition $(p_1', p_2', \ldots, p_\alpha')$, where

$$p_i' = \begin{cases} 2 & \text{if } i \in I, \\ p_i & \text{if } i > 1 \text{ and } i \notin I, \\ p_1 + \sum_{i \in I}(p_i - 2) & \text{if } i = 1, \end{cases}$$

specializes to x, compare [Hesselink 1976, 3.10] or [Collingwood and McGovern 1993, 6.2.5]. Replacing x with y we find an element with partition $(2s, 2^r, 1^{\alpha(x)-r-1})$ for some $s \geq 1$ and some r. If $r \geq 2$ and $s > 1$, then we may replace x with an element with partition $(2s+2, 2^{r-2}, 1^{\alpha(x)-r+1})$ and repeating this procedure gives the claim.

The formula for $\dim C_{\mathrm{Sp}_{2n}(k)}(y)$ in [Liebeck and Seitz 2012, p. 39] gives that it is at least $n + (\alpha(x)^2 - 1)/2$. Applying $\lceil n/(2n-\alpha(x))\rceil < (3n-\alpha(x))/(2n-\alpha(x))$, we find that

$$e(x) \cdot \dim x^G < 6n^2 + \alpha(x)(n - \alpha(x)) + 1/(2n - \alpha(x)).$$

As $n - \alpha(x)$ is negative, we have verified the required inequality for x nilpotent.

Semisimple case. We may assume x is diagonal. Put α_0 for the number of nonzero entries in x; we will construct a nilpotent y in the closure of $x^{\mathbb{G}_m \mathrm{Sp}_{2n}}$. Recall that the diagonal of x consists of pairs $(t, -t)$ with $t \in k$.

Suppose first that $\alpha_0 \geq n$. We pick y to be block diagonal as follows. For a 4-by-4 block with entries $(0, 0, t, -t)$ for some $t \in k^\times$, we make a 4-by-4 block in y in the same location, where the 2-by-2 block in the upper right corner is generic for $\mathfrak{sp}_4$. As $\alpha_0 \geq n$, by permuting the entries in x we may assume that all pairs $(0, 0)$ on the diagonal of x are immediately followed by a $(t, -t)$ with $t \neq 0$. Thus, it remains to specify the diagonal blocks in y at the locations corresponding to the remaining 2-by-2 blocks $(t, -t)$ for

$t \neq 0$ in x, for which we take y to have a 1 in upper right corner. We have constructed a nilpotent y with rank $y \geq n$, so $e(x) \leq e(y) \leq 3$ by Proposition 8.1, and $e(x) \cdot \dim x^{\mathrm{Sp}_{2n}} \leq 6n^2$.

Now suppose $\alpha_0 < n$. Let x_0 be a $2\alpha_0$-by-$2\alpha_0$ submatrix consisting of all the nonzero diagonal entries in x together with α_0 zero entries. Take y_0 to be the nilpotent element constructed from x_0 as in the preceding paragraph, and extend it by zeros to obtain a nilpotent y with $\alpha(y) = 2n - \alpha_0 > n$. Then y is in the closure of $x^{\mathbb{G}_m \mathrm{Sp}_{2n}}$ and $e(y) \leq 2\lceil n/\alpha_0 \rceil < 2(n+\alpha_0)/\alpha_0$. On the other hand, the centralizer of x has dimension at least $\dim \mathrm{Sp}_{2n-\alpha_0} + \alpha_0/2 = 2n^2 - 2n\alpha_0 + \alpha_0^2/2 + n$. Thus $\dim x^{\mathrm{Sp}_{2n}} \leq 2n\alpha_0 - \alpha_0^2/2$. In summary, $e(x) \cdot \dim x^{\mathrm{Sp}_{2n}} < (n+\alpha_0)(4n-\alpha_0) = 4n^2 + 3\alpha_0 n - \alpha_0^2$. As a function of α_0, it is a parabola opening down with max at $\alpha_0 = 1.5n$, so its maximum for $\alpha_0 < n$ is where $\alpha_0 = n-1$, i.e., the max is at most $6n^2 - n - 1$. □

9. Types *B* and *D* with char $k \neq 2$

Proposition 9.1. *Assume* $\mathrm{char}\, k \neq 2$. *For every nonzero nilpotent* $x \in \mathfrak{so}_n$ *for* $n \geq 5$, $\max\{4, \lceil n/(n-\alpha(x))\rceil\}$ *conjugates of* x *generate* $\mathfrak{so}_n$.

Proof. The O_n-conjugacy class of x is determined by its Jordan form, which is given by a partition $(p_1, \ldots, p_\alpha)$ of n where even values occur with even multiplicity. We go by induction on n. As $\mathfrak{so}_5 \cong \mathfrak{sp}_4$, the $n = 5$ case is covered by Proposition 8.1, which gives 4 as the largest number of conjugates needed to generate. For $n = 6$, $\mathfrak{so}_6 \cong \mathfrak{sl}_4$, and this case is handled by Proposition 6.4. So assume $n \geq 7$.

Suppose first that the number δ of 1's in the partition for x is at most 1. Then we can find an element y in the closure of x^{SO_n} with partition

(i) $(2^{n/2})$ if $n \equiv 0 \bmod 4$;

(ii) $(2^{(n-1)/2}, 1)$ if $n \equiv 1 \bmod 4$;

(iii) $(3^2, 2^{(n-6)/2})$ if $n \equiv 2 \bmod 4$; or

(iv) $(3, 2^{(n-3)/2})$ if $n \equiv 3 \bmod 4$.

To see this, we specialize $(2s, 2s) \rightsquigarrow (s^4)$ for $s \geq 2$; $s \rightsquigarrow (s-4, 2, 2)$ for odd $s \geq 7$; or $(s, 1) \rightsquigarrow ((s+1)/2, (s+1)/2))$ for odd $s \geq 3$ and $\delta = 1$. Together with trivial reductions such as $(5^2) \rightsquigarrow (3^2, 2^2)$ brings us to a partition of the form $(3^b, 2^c, 1^\delta)$ for some $b \leq 3$ and some c from which the claim quickly follows. For such a y, 2 conjugates suffice to generate a copy of $\mathfrak{sl}_2^{\times n/2}$, $\mathfrak{sl}_2^{\times (n-1)/2}$, $\mathfrak{so}_3 \times \mathfrak{so}_3 \times \mathfrak{sl}_2^{\times (n-6)/2}$, or $\mathfrak{so}_3 \times \mathfrak{sl}_2^{(n-3)/2}$ respectively. As in the proof of Proposition 8.1, it follows that 3 conjugates are enough to generate $\mathfrak{so}_n$.

Now suppose there are more 1's in the partition for x. We specialize using

$$(2s+1, 1) \rightsquigarrow (s+1, s+1) \text{ for } s \geq 1 \quad \text{and} \quad (s, s, 1, 1) \rightsquigarrow (s-1, s-1, 2, 2) \text{ for } s \geq 4.$$

If, after a step in this specialization process, we find that only 0 or 1 1-by-1 blocks remain, we are done by the preceding paragraph. Therefore, we may assume that x has partition $(2^{2t}, 1^u)$ for $u \geq 2$.

Write out $t = 2t_0 + \delta$ for $\delta = 0$ or 1, and set $v = 2t_0\lceil u/(2t)\rceil$. We can view x as block diagonal where the first block has partition $(2^{2t_0}, 1^v)$ and the second has partition $(2^{2t_0+2\delta}, 1^{u-v})$. For the first block,

$$e := 2 + \left\lceil \frac{v}{2t_0} \right\rceil = 2 + \left\lceil \frac{u}{2t} \right\rceil$$

conjugates suffice to generate an $\mathfrak{so}_{2t_0e}$ subalgebra by induction on n. For the second block, we note that

$$\frac{u-v}{2t_0+2\delta} \le \frac{u}{2t},$$

so, by induction, e conjugates suffice to generate an $\mathfrak{so}_{n-2t_0e}$ subalgebra. Because $\mathfrak{so}_{2t_0e} \times \mathfrak{so}_{n-2t_0e}$ contains a regular semisimple element and the natural module has composition factors of size $2t_0e$ and $n - 2t_0e$, we conclude as in the proof of Proposition 8.1 that e conjugates of x suffice to generate $\mathfrak{so}_n$. □

Corollary 9.2. *Assume $p := \operatorname{char} k \neq 2$. For noncentral $x \in \mathfrak{so}_n$ with $n \geq 5$ such that $x^{[p]} \in \{0, x\}$, there exist $e > 0$ and elements $x_1, \ldots, x_e \in x^{\mathrm{SO}_n}$ that generate $\mathfrak{so}_n$ such that $e \cdot \dim x^{\mathrm{SO}_n} \leq 2(n-1)^2$.*

Proof. As $\operatorname{char} k \neq 2$, we identify $\mathfrak{spin}_n$ with $\mathfrak{so}_n$ via the differential of the covering map $\mathrm{Spin}_n \to \mathrm{SO}_n$. We argue as in the proof of Corollary 8.3, replacing $\mathfrak{sp}_{2n}$ with $\mathfrak{so}_n$ and references to Proposition 8.1 with references to 9.1. We may assume that $e(x) > 4$, for otherwise $e(x) \cdot x^{\mathrm{SO}_n} \leq 4 \cdot \left(\binom{n}{2} - \lfloor n/2 \rfloor\right) \leq 2(n-1)^2$.

Nilpotent case. Suppose that x is nonzero nilpotent. We have $e(x) \leq \lceil n/(n-\alpha)\rceil$ and in particular we may assume that $\alpha > \frac{2}{3}n$. Recall that the O_n-orbit of x is determined by a partition $(p_1, \ldots, p_\alpha)$ of n, where even numbers appear with even multiplicity. As in the proof of Corollary 8.3, we may replace x with y with partition $(p_1 + \sum_{i=2}^{\alpha}(p_i - 1), 1^{\alpha-1})$. This element has $\alpha(y) = \alpha(x)$ and orbit of size $\binom{n}{2} - \binom{\alpha}{2}$. As $e(x) < (2n-\alpha)/(n-\alpha)$, it follows that $e(x) \cdot x^{\mathrm{SO}_n} < \frac{1}{2}(2n-\alpha)(n+\alpha-1)$. The upper bound is maximized for $\frac{2}{3}n \leq \alpha \leq n$ at the lower bound, where it is $\frac{2}{9}n(5n-3) < 2(n-1)^2$.

Semisimple case. Suppose that x is noncentral diagonal in $\mathfrak{so}_n$.

Suppose first that n is even. If $\alpha_0 \geq n/2$, then pick y as in Corollary 8.3, so $\alpha(y) = n/2$, $e(y) \leq 4$, and we are done. If $\alpha_0 < n/2$, we perform the same construction as in the last paragraph of the proof of 8.3 to obtain y with $\alpha(y) = n - \alpha_0$, so $e(y) \leq \max\{4, \lceil n/\alpha_0\rceil\}$; suppose $\lceil n/\alpha_0\rceil > 4$, i.e., $n/\alpha_0 > 4$, i.e., $\alpha_0 < n/4$. The orbit of x has dimension at least $\dim \mathrm{SO}_n - \dim \mathrm{SO}_{n-\alpha_0} - \alpha_0/2$, whence $e(x) \cdot \dim x^{\mathrm{SO}_n} < (n+\alpha_0)(n-\alpha_0/2-1)$, where the right side is maximized at $\alpha_0 = n/4$ and again we verify that the upper bound is at most $2(n-1)^2$.

When n is odd, we view x as lying in the image of $\mathfrak{so}_{n-1} \hookrightarrow \mathfrak{so}_n$ and take y in this same image as constructed by the method in the previous paragraph. Computations identical to the ones just performed again verify $e(x) \cdot \dim x^{\mathrm{SO}_n} < 2(n-1)^2$. □

10. Type *D* with char *k* = 2

Concrete descriptions. For sake of precision, we first give concrete descriptions of the groups and Lie algebras associated with a nondegenerate quadratic from q on a vector space V of even dimension $2n$ over a field k (of any characteristic). The orthogonal group $\mathrm{O}(q)$ is the subgroup-scheme of $\mathrm{GL}(V)$

consisting of elements that preserve q, i.e., such that $q(gv) = q(v)$ for all $v \in V \otimes R$ for every commutative k-algebra R; the special orthogonal group $\mathrm{SO}(q)$ is the kernel of the Dickson invariant $\mathrm{O}(q) \to \mathbb{Z}/2$; and the groups of similarities $\mathrm{GO}(q)$ and proper similarities $\mathrm{SGO}(q)$ are the subgroup-schemes of $\mathrm{GL}(V)$ generated by the scalar transformations and $\mathrm{O}(q)$ or $\mathrm{SO}(q)$ respectively; see for example [Knus et al. 1998, §12 and p. 348; Knus 1991, Chapter IV]. For $n \geq 3$, the group $\mathrm{SO}(q)$ is semisimple of type D_n, but neither simply connected nor adjoint.

The statement that q is nondegenerate means that the bilinear form b on V defined by $b(v, v') := q(v + v') - q(v) - q(v')$ is nondegenerate. Viewing the Lie algebra of a group G over k as the kernel of the homomorphism $G(k[\varepsilon]) \to G(k)$ induced by the map $\varepsilon \mapsto 0$ from the dual numbers $k[\varepsilon]$ to k, one finds that $\mathfrak{o}(q)$ is the set of $x \in \mathfrak{gl}(V)$ such that $b(xv, v) = 0$ for all $v \in V$. Since $\mathrm{O}(q)/\mathrm{SO}(q) \cong \mathbb{Z}/2$, $\mathfrak{so}(q) = \mathfrak{o}(q)$. As b is nondegenerate, the equation $b(Tv, v') = b(v, \sigma(T)v')$ defines an involution σ on $\mathrm{End}(V)$. The set of alternating elements $\{T - \sigma(T) \mid T \in \mathrm{End}(V)\}$ is contained in $\mathfrak{so}(q)$ and also has dimension $2n^2 - n$ [Knus et al. 1998, 2.6], therefore the two subspaces are the same. The Lie algebra $\mathfrak{go}(q)$ of $\mathrm{GO}(q)$ and $\mathrm{SGO}(q)$ is the set of elements $x \in \mathfrak{gl}(V)$ such that there exists a $\mu_x \in k$ so that $b(xv, v) = \mu_x q(v)$ for all v. It has dimension one larger than $\mathfrak{so}(q)$.

We assume for the remainder of the section that $\operatorname{char} k = 2$.

Example 10.1. When $V = k^{2n}$ and q is defined by $q(v) = \sum_{i=1}^{n} v_i v_{i+n}$, we write $\mathfrak{so}_{2n}$ instead of $\mathfrak{so}(q)$, etc. The linear transformation x obtained by projecting on the first n coordinates satisfies $b(xv, v) = q(v)$ for all $v \in V$, so it and $\mathfrak{so}_{2n}$ span $\mathfrak{go}_{2n}$.

Suppose $x \in \mathfrak{go}_{2n}$ is a projection, i.e., $x^2 = x$, so x gives a decomposition $k^{2n} = \ker x \oplus \operatorname{im} x$ as vector spaces. If x belongs to $\mathfrak{so}_{2n}$, then this is an orthogonal decomposition and b is nondegenerate on $\ker x$ and $\operatorname{im} x$. Up to conjugacy, x stabilizes the subspaces spanned by vectors with nonzero entries only in the first n coordinates or the last n coordinates, which exhibits x as the image of some toral $\hat{x}$ under an inclusion $\mathfrak{gl}_n \hookrightarrow \mathfrak{so}_{2n}$ such that $2 \operatorname{rank} \hat{x} = \operatorname{rank} x$. Suppose $x \notin \mathfrak{z}(\mathfrak{so}_{2n})$, so $\hat{x} \notin \mathfrak{z}(\mathfrak{gl}_n)$. Let $\hat{y} \in \mathfrak{gl}_n$ denote the nilpotent obtained for $\hat{x}$ as in Lemma 4.2, and put $y \in \mathfrak{so}_{2n}$ for its image. Then y is in the closure of $x^{\mathbb{G}_m \mathrm{SO}_{2n}}$ and $\operatorname{rank} y \leq \operatorname{rank} x$ with equality if $\operatorname{rank} x \leq n$.

If $x \in \mathfrak{go}_{2n} \setminus \mathfrak{so}_{2n}$ has $x^2 = x$, then $\operatorname{im} x$ and $\ker x$ are maximal totally isotropic subspaces. To see this, note that if $q(v) \neq 0$, then $b(xv, v) = \mu_x q(v) \neq 0$, which is impossible if $xv \in \{0, v\}$.

We consider how many conjugates of an $x \in \mathfrak{so}_{2n}$ with $x^{[2]} \in \{0, x\}$ suffice to generate a subalgebra of $\mathfrak{so}_{2n}$ containing the derived subalgebra $[\mathfrak{so}_{2n}, \mathfrak{so}_{2n}]$. We apply Lemma 4.3 (3) with $G = \mathrm{GO}_{2n}$, $M = [\mathfrak{so}_{2n}, \mathfrak{so}_{2n}]$, and $N = \mathfrak{z}(M)$, so $\dim N = 0$ or 1, $\dim M/N \geq 2n^2 - n - 2$ and $\dim \mathfrak{g}/M = 2$.

Example 10.2. One can verify by computing with an example that for $x \in \mathfrak{so}_{2n}$ with $x^{[2]} = 0$, e conjugates suffice to generate $[\mathfrak{so}_{2n}, \mathfrak{so}_{2n}]$ in the cases (a) x is a root element and $n = e = 4$ or 5 or (b) $n = 7$ or 8, $e = 4$, and x has rank 4. (In the last case, note that x can be taken to have Jordan form with partition $(2^4, 1^{2n-4})$.) Magma code is provided with the arxiv version of this paper.

In the following, we say that an $\mathfrak{so}_{2m}$ subalgebra of $\mathfrak{so}_{2n}$ is *naturally embedded* if it arises from expressing k^{2n} as an orthogonal sum of a nondegenerate $2n - 2m$ and $2m$-dimensional spaces.

Lemma 10.3. *Let $\mathfrak{g} = \mathfrak{so}_{2n}$ with $n \geq 4$. If x is a root element, and $m \geq 4$, then m generic conjugates of x generate the derived subalgebra of a naturally embedded $\mathfrak{so}_{2m}$.*

Proof. The case $n = 4$ is from Example 10.2.

Now assume that $n > 4$ and $4 \leq m < n$. By induction on n, we know that m conjugates can generate the derived subalgebra of a copy of $\mathfrak{so}_{2m}$. Clearly any m conjugates have a fixed space of dimension at least $2n - 2m$ and generically this space will be nondegenerate, whence this $\mathfrak{so}_{2m}$ is naturally embedded.

Now assume that $m = n$; by Example 10.2 we may assume that $n \geq 6$. So now take $n - 2$ generic root elements, $x_1, \ldots, x_{n-2}$; they generate the derived subalgebra of a natural $\mathfrak{so}_{2n-4}$ by induction. Let us take a basis of k^{2n} as in Example 10.1. We identify our $\mathfrak{so}_{2n-4}$ as the one acting trivially on the subspace spanned by $v_1, v_{n+1}, v_2, v_{n+2}$.

Then consider two copies of the derived subalgebra of $\mathfrak{so}_{2n-2}$ acting on the spaces spanned by v_i and v_{n+i} for $1 \leq i < n$ and for $1 < i \leq n$. These both contain our $\mathfrak{so}_{2n-4}$ and by induction we can choose x, y respectively so that $x, x_1, \ldots, x_{n-2}$ generate the first copy of the derived subalgebra of $\mathfrak{so}_{2n-2}$ and $x_1, \ldots, x_{n-2}, y$ generate the second copy. These two copies generate the derived subalgebra of $\mathfrak{so}_{2n}$, as can be seen by considering the root elements in each one. □

Proposition 10.4. *Let $G = \mathrm{SO}_{2n}$ with $n \geq 3$ over an algebraically closed field k of characteristic 2. For noncentral $x \in \mathfrak{g}$ such that $x^{[2]} \in \{0, x\}$, $\max\{4, \lceil n/r \rceil\}$ conjugates of x generate a Lie subalgebra containing $[\mathfrak{g}, \mathfrak{g}]$ where $2r$ is the codimension of the largest eigenspace of x.*

For x as in the proposition: if (1) $x^{[2]} = x$ and $\operatorname{rank} x \leq n$ or (2) x is nilpotent, then $2r$ is the rank of x.

Proof. Suppose $x^{[2]} = 0$ and $n \geq 4$. The closure of x^G contains a nilpotent element y of the same rank with y contained in a Levi subalgebra $\mathfrak{gl}_n$ (by [Liebeck and Seitz 2012, Table 4.1], this reduces to the case of $\mathfrak{so}_4$ where the result is clear). Thus we may assume that x is nilpotent and is contained in $\mathfrak{sl}_n$. The case where x is a root element was considered in Lemma 10.3 (with $m = n$), so we may assume $r \geq 2$.

If $n/r \leq 3$, then by the result for $\mathfrak{sl}_n$ (Proposition 7.1), we can generate an $\mathfrak{sl}_n$ with 3 conjugates. Since $\mathfrak{g}/\mathfrak{sl}_n$ is multiplicity free as an $\mathfrak{sl}_n$-module, this implies the result.

Suppose that $n \leq 8$. The result follows by the previous paragraph unless $r = 2$ and $n = 7$ or 8. These cases were settled in Example 10.2.

Now suppose that $n \geq 9$ and put e for the maximum appearing in the statement. By the result for $\mathfrak{sl}_n$, e conjugates can generate an $\mathfrak{sl}_n$ and something containing the derived subalgebra of $\mathfrak{so}_{2n-2}$. Therefore generically, e conjugates generate an irreducible subalgebra of $\mathfrak{g}$ and in particular, the center is central in $\mathfrak{g}$.

Suppose that n is odd. On the irreducible module X with highest weight the highest root, there exist e conjugates with composition factors of dimensions $n^2 - 1$, $n(n-1)/2$, $n(n-1)/2$ and also one where there is a composition factor of dimension at least $(n-1)(2n-3) - 1$. Thus, generically there is a composition factor of dimension at most $2n^2 - 5n + 2$ and the smallest composition factor is at least $n(n-1)/2$. Since the sum of these two numbers (for $n \geq 9$) is greater than $\dim X = 2n^2 - n - 2$, we see that generically e conjugates acts irreducibly on X, whence they generate $\mathfrak{g}$ (by dimension).

Suppose that n is even. The same argument shows that e conjugates can generate a subalgebra having composition factors on $[\mathfrak{g},\mathfrak{g}]$ of dimensions 1, n^2-2, $n(n-1)/2$, $n(n-1)/2$ and another e conjugates having composition factors of dimensions 1, $2n^2-5n+2$, $2n-2$, $2n-2$. This implies that generically e conjugates act irreducibly on $[\mathfrak{g},\mathfrak{g}]/\mathfrak{z}(\mathfrak{g})$ and this implies they generate $[\mathfrak{g},\mathfrak{g}]$.

Suppose that $x^{[2]}=0$ and $n=3$. Then x is the image of a square-zero element under the differential of $\mathrm{SL}_4 \to \mathrm{SL}_4/\mu_2 \cong \mathrm{SO}_6$, and 4 conjugates of x suffice to generate $[\mathfrak{g},\mathfrak{g}]$ by Lemmas 7.1 and 3.1 (1).

Suppose now that $x^{[2]}=x$. If $\operatorname{rank} x \le n$, then let y be the nilpotent element provided by Example 10.1, so $\operatorname{rank} y = \operatorname{rank} x$ and the claim follows from the nilpotent case.

If $\operatorname{rank} x > n$, then set $x' = I_{2n} - x \in \mathfrak{so}_{2n}$, which is toral of rank $2r \le n$. Applying the previous case shows that $\max\{4, \lceil n/r\rceil\}$ conjugates of x' generate a Lie subalgebra containing $[\mathfrak{g},\mathfrak{g}]$. Therefore, since I_{2n} is central in $\mathfrak{so}_{2n}$, the same number of conjugates of x generate a Lie subalgebra containing $[\mathfrak{g},\mathfrak{g}]$ by Lemma 3.5. □

Example 10.5. Suppose $x \in \mathfrak{so}_{2n}$ satisfies $x^{[2]}=0$, so the Jordan form of x has $2r$ 2-by-2 blocks and $2n-4r$ 1-by-1 blocks for some $r \le n$. There are two possibilities for the conjugacy class of x; see [Hesselink 1979, 4.4; Liebeck and Seitz 2012, p. 70]. We focus on the larger class, the one where the restriction of the natural module to x includes a 4-dimensional indecomposable denoted by $W_2(2)$ in [Liebeck and Seitz 2012]. The centralizer of such an x in SO_{2n} has dimension

$$\sum_{i=1}^{2r} 2(i-1) + \sum_{i=2r+1}^{2n-2r} (i-1) = \binom{2n-2r}{2} + \binom{2r}{2},$$

and therefore $\dim x^{\mathrm{SO}_{2n}} = 4r(n-r)$. (The other class has dimension $2r(2n-2r-1)$.)

Corollary 10.6. *Suppose* $\operatorname{char} k = 2$. *For every noncentral* $x \in \mathfrak{go}_{2n}$ *with* $n \ge 4$ *such that* $x^{[2]} \in \{0, x\}$, *there exist* $e > 0$ *and elements* $x_1, \dots, x_e \in x^{\mathrm{SGO}_{2n}}$ *that generate a subalgebra containing* $[\mathfrak{so}_{2n}, \mathfrak{so}_{2n}]$ *such that* $e \cdot \dim x^{\mathrm{GO}_{2n}} \le 4n^2$.

Proof. Suppose x has $x^{[2]}=0$ and rank $2r$ as in Example 10.5. The condition we need is that $4n^2 \ge e\,4r(n-r)$. If the maximum in Proposition 10.4 is 4, i.e., if $r \ge n/4$, then as a function of r, $16r(n-r)$ has a maximum of $4n^2$ at $r=n/2$. Otherwise, the maximum is $e=\lceil n/r\rceil < (n+r)/r$, so $e \dim x^{\mathrm{GO}_{2n}} < 4(n^2-r^2)$. The right side has a maximum of $4n^2-4$ at $r=1$.

If $x^{[2]}=x$ and $x \in \mathfrak{so}_{2n}$, the centralizer of x in GO_{2n} has codimension 1 in $\mathrm{GO}_{2r'} \times \mathrm{GO}_{2(n-r')}$ when x has rank $2r'$. We may take $e=\max\{4, \lceil n/r\rceil\}$ where $2r$ is the dimension of the smallest eigenspace of x. If $r' \ge n/4$, then $4 \dim x^{\mathrm{GO}_{2n}} \le 4n^2$ as for nilpotent elements. So assume $r < n/4$. Then, as $r'=r$ or $n-r$, $e \dim x^{\mathrm{GO}_{2n}}$ is at most $(n+r')4r(n-r)/r' = 4(n^2-s^2)$ for $s=r$ or r', and again we conclude as for nilpotent elements.

If $x^{[2]}=x$ and $x \notin \mathfrak{so}_{2n}$, then x is determined by choosing an ordered pair of "parallel" maximal isotropic subspaces and so the dimension of $x^{\mathrm{GO}_{2n}}$ agrees with the dimension of the flag variety of D_n of parabolics with Levi subgroups of type A_{n-2}, which has dimension $(n^2+n-2)/2$. Up to conjugacy, we

may assume x is the element from Example 10.1. Let y_0 be an n-by-n nilpotent matrix of with $\lfloor n/2 \rfloor$ 2-by-2 rank 1 Jordan blocks down the diagonal. Then $y = \left(\begin{smallmatrix} 0 & y_0 \\ y_0 & 0 \end{smallmatrix}\right)$ is in $\mathfrak{so}_{2n}$, is nilpotent, and 4 conjugates of y suffice to generate a subalgebra containing $\mathfrak{so}_{2n}$ (Proposition 10.4), so 4 conjugates of x suffice as well. As $2n^2 + 2n - 4 < 4n^2$, the claim is proved in this case. □

11. Exceptional types

The aim of this section is to provide the necessary material to prove Theorem A for exceptional groups, but we begin with some general-purpose observations. Recall that a *root element* of a Lie algebra $\mathfrak{g}$ of G is a generator for a one-dimensional root subalgebra $\mathfrak{g}_\alpha$ of $\mathfrak{g}$.

Lemma 11.1. *Let G be a simple algebraic group such that $(G, \operatorname{char} k) \neq (\mathrm{Sp}_{2n}, 2)$ for $n \geq 2$. For each nonzero nilpotent $x \in \mathfrak{g}$, there is a root element in the closure of x^G.*

We ignore what happens in the excluded case.

Proof. Write $x = \sum_{\alpha \in S} X_\alpha$ where S is a nonempty set of positive roots (relative to some torus T) and X_α is a generator for $\mathfrak{g}_\alpha$. If $|S| = 1$ (e.g., if G has type A_1), then we are done. Otherwise, the hypothesis on $(G, \operatorname{char} k)$ guarantees that no root vanishes on T, so we can pick a subtorus T' of T that centralizes some X_α but not some $X_{\alpha'}$ for some $\alpha \neq \alpha' \in S$. Now in the closure of $x^{T'}$ we find a nonzero nilpotent supported on $S \setminus \{\alpha'\}$, and by induction we are done. □

We say that a root element in $\mathfrak{g}_\alpha$ is long (resp. short) if α is long (resp. short).

Lemma 11.2. *Let G be a simple linear algebraic group over a field k such that $\operatorname{char} k$ is not special for G. For every nonzero nilpotent $x \in \mathfrak{g}$, there is a **long** root element in the closure of x^G.*

Proof. By Lemma 11.1, we may assume that G has two root lengths and that x is a root element for a short root α.

Suppose first that G has rank 2, so G has type G_2 and $\operatorname{char} k \neq 3$ or G has type C_2 and $\operatorname{char} k \neq 2$. Let α be the short simple root, γ be the highest root (a long root), and take $\beta := \gamma - \alpha$. Let $x_\alpha, x_\beta \colon \mathbb{G}_a \to G$ be the corresponding root subgroups. These pick a generator $X_\alpha := \mathrm{d}x_\alpha(1)$ of $\mathfrak{g}_\alpha$ such that

$$\operatorname{ad}(x_\beta(t))X_\alpha = X_\alpha + N_{\beta,\alpha} X_\gamma,$$

where X_γ generates $\mathfrak{g}_\gamma$, cf. [Steinberg 2016, Chapter 3]. As $\operatorname{char} k$ is not special for G, $N_{\beta,\alpha}$ is not zero in k, and arguing as in the proof of Lemma 11.1 we conclude that $k^\times X_\gamma$ meets the closure of $(X_\alpha)^G$, proving the claim in case G has rank 2.

If G has rank at least 3, pick a long root β that is not orthogonal to α and let G' be the subgroup of G corresponding to the rank 2 sub-root-system generated by α, β. The ratio of the square-lengths of α, β is 2 so G' has type C_2 and $\operatorname{char} k \neq 2$. Then the closure of $x^{G'}$ contains a long root element in G', hence in G. □

Remark 11.3. Suppose that G is a simple linear algebraic group over k such that $\operatorname{char} k$ is special for G. The short root subalgebras generate a G-invariant subalgebra $\mathfrak{n}$ of $\mathfrak{g}$. Omitting the case where

$(G, \operatorname{char} k) = (\mathrm{Sp}_{2n}, 2)$ for $n \geq 2$, the same argument as in the proof of Lemma 11.1 shows that for a nonzero nilpotent $x \in \mathfrak{g} \setminus \mathfrak{n}$ (resp., $\in \mathfrak{n}$), there is a long (resp. short) root element in the closure of x^G.

Now we focus on exceptional groups. Table 2 appears on page 1606.

Proposition 11.4. *Suppose G is simple of exceptional type over a field k such that* $\operatorname{char} k$ *is not special for G. For e as in Table 2, $b(G)$ as in Table 1, and $x \in \mathfrak{g}$ noncentral, we have*:

(1) *there are $x_1, \ldots, x_e \in x^G$ generating a Lie subalgebra of $\mathfrak{g}$ containing $[\mathfrak{g}, \mathfrak{g}]$, and*

(2) $e \cdot \dim x^G \leq b(G)$.

Proof. The crux is to prove (1). By taking closures as in Corollary 4.4, we may assume that the orbit x^G of x consists of root elements. Moreover, as k is not special, by Lemma 11.2 we may assume that x^G consists of long root elements. In view of Lemma 3.1 (1), we may assume $\mathfrak{g}$ is simply connected.

If $p \neq 2$, we can apply the result of [Cohen et al. 2001] to obtain (1). We now prove the result for $p = 2$; in most cases, the argument also gives another proof for all p.

If $G = G_2$, we consider the A_2 subalgebra $\mathfrak{h}$ generated by the long roots so $\mathfrak{g}/\mathfrak{h}$ has the weights of $k^3 \oplus (k^3)^*$ as a representation of $\mathfrak{h}$, so it is multiplicity free. As $\mathfrak{h}$ can be generated with 3 root elements (Proposition 7.1), the claim follows.

If $G = E_n$, one uses that 4 root elements generate the D_4 inside E_n (Example 10.2) and argue as for G_2, or one computes directly that five random root elements generate $\mathfrak{g}$. This completes the proof of (1).

Claim (2) follows because

$$b(G) = e \cdot (\dim G - \operatorname{rank} G) \geq e \cdot \dim x^G. \qquad \square$$

12. Proof of Theorem A

Lemma 12.1. *Let G be a simple algebraic group over a field k such that $p := \operatorname{char} k$ is not special. Then for $b(G)$ as in Table 1 and all noncentral $x \in \mathfrak{g}$ such that $x^{[p]} \in \{0, x\}$, there exists $e > 0$ and elements $x_1, \ldots, x_e \in x^G$ generating a subalgebra $\mathfrak{s}$ of $\mathfrak{g}$ containing $[\mathfrak{g}, \mathfrak{g}]$ such that $e \cdot \dim x^G \leq b(G)$.*

Proof. We apply Proposition 11.4 if G has exceptional type.

Put $\pi : \widetilde{G} \to G$ for the simply connected cover of G. If $\mathrm{d}\pi : \tilde{\mathfrak{g}} \to \mathfrak{g}$ is an isomorphism, then we apply Corollary 6.5 or 7.2 for type A, 9.2 for types B or D if $p \neq 2$, and 8.3 for type C.

Therefore, we may assume that $G = \mathrm{SL}_n / \mu_m$ and $p \mid m$, or G has type D_n and $p = 2$. In these cases, Corollaries 6.5, 7.2, and 10.6 concern not G but a group $G'' := (G' \times \mathbb{G}_{\mathrm{m}})/Z(G')$ for some G' isogenous to G. In particular, putting $q : G' \to \overline{G}$ for the natural surjection onto the adjoint group, the induced map $\mathrm{d}q : \mathrm{Lie}(G'') \to \mathrm{Lie}(\overline{G})$ is also a surjection.

Consider now the case $G = \overline{G}$. Pick $y \in \mathrm{Lie}(G'')$ such that $\mathrm{d}q(y) = x$. The results cited in the second paragraph of the proof provide elements $y_1, \ldots, y_e \in y^{G''}$ such that $\mathfrak{s}'' := \langle y_1, \ldots, y_e \rangle$ contains $[\mathfrak{g}'', \mathfrak{g}'']$, and $e \cdot \dim y^{G''} \leq b(G)$. Taking $x_i := \mathrm{d}q(y_i)$, we obtain the desired result.

In the general case, write now q for the natural map $G \to \overline{G}$. For $z := \mathrm{d}q(x)$, let $z_1, \ldots, z_e \in z^{\overline{G}}$ be the elements provided by the adjoint case of the lemma. Pick $g_i \in G(k)$ such that $z_i = \mathrm{Ad}(g_i)z$ and set

$x_i := \mathrm{Ad}(g_i)x$. Then $x_1, \dots, x_e$ generate a subalgebra $\mathfrak{s}$ such that $\mathrm{d}q(s) \supseteq [\bar{\mathfrak{g}}, \bar{\mathfrak{g}}]$. Lemma 3.5 completes the proof. □

We prove the following result, which has the same hypotheses as Theorem A and a stronger conclusion.

Theorem 12.2. *Let G be a simple linear algebraic group over a field k such that $p := \mathrm{char}\, k$ is not special for G. If $\rho\colon G \to \mathrm{GL}(V)$ is a representation of G such that V has a G-subquotient X with $X^{[\mathfrak{g},\mathfrak{g}]} = 0$ and $\dim X > b(G)$ for $b(G)$ as in Table 1, then $\dim x^G + \dim V^x < \dim V$ for all noncentral $x \in \mathfrak{g}$ with $x^{[p]} \in \{0, x\}$.*

Proof. If $V = X$, combining Lemma 12.1 with Section 1 shows that the desired inequality holds. This implies the inequality for general V as in Example 2.1. □

Proof of Theorem A. Combine Theorem 12.2 with Lemmas 1.1 and 1.7. □

13. Small examples; proof of Corollary B

Before proving Corollary B, we provide an example that we treat in greater generality than is required for proving the corollary. We put $\mathrm{S}^2 V$ for the second symmetric power of the vector space V.

Lemma 13.1. *Suppose $\mathrm{char}\, k \neq 2$. Let $G = \mathrm{SO}(V)$ with $\dim V = n$. Let W be the irreducible composition factor of $\mathrm{S}^2 V$ of dimension $n(n+1)/2-1$ if $\mathrm{char}\, k$ does not divide n, or $n(n+1)/2-2$ if $\mathrm{char}\, k$ divides n. Then $\mathfrak{g}$ acts generically freely on V.*

Proof. By fixing a basis for V, we may identify $\mathrm{S}^2 V$ with n-by-n symmetric matrices and $\mathfrak{g}$ with skew-symmetric matrices. Then we see W inside $\mathrm{S}^2 V$ (with $\mathfrak{g}$ acting via the Lie bracket in $\mathfrak{gl}_n$).

If $\mathrm{char}\, k$ does not divide n, W is just the trace zero matrices in $\mathrm{S}^2 V$. If $\mathrm{char}\, k$ divides n, then W is the set of trace zero matrices modulo scalars.

If we take an element of trace zero that is diagonal and generic, then its centralizer in $\mathfrak{gl}_n$ is just diagonal matrices (and even so for commuting modulo scalars). Thus, its centralizer in $\mathfrak{g}$ is 0, whence the generic stabilizer in $\mathfrak{g}$ is 0. □

Proof of Corollary B. As $\mathrm{char}\, k$ is not special and we may assume that $\mathrm{d}\rho \neq 0$, we have $\ker \mathrm{d}\rho \subseteq \mathfrak{z}(\mathfrak{g})$. If $\dim V < \dim G - \dim \mathfrak{z}(\mathfrak{g})$, we have $\dim \mathrm{d}\rho(\mathfrak{g})v \leq \dim V < \dim \mathrm{d}\rho(\mathfrak{g})$, whence $\mathfrak{g}$ does not act virtually freely.

At the other extreme, if $\dim V > b(G)$ as in Table 1, then V is virtually free by Theorem A because $V^{[\mathfrak{g},\mathfrak{g}]} = 0$ by Lemma 3.1 (2).

For groups of classical type, the possible V with $\dim V \leq \dim G$ are listed in [Lübeck 2001a, Table 2]. The cases with $\dim G - \dim \mathfrak{z}(\mathfrak{g}) \leq \dim V \leq \dim G$ are settled in Lemma 13.1 and Example 3.4, so assume $\dim V > \dim G$.

Consider first G of type A_ℓ. By [Lübeck 2001a, Theorem 5.1], $\dim V > \ell^3/8$. If $\ell \geq 20$, then $\ell^3/8 > b(G)$ and we are done. For $16 \leq \ell \leq 19$, the tables in [Lübeck 2001a][1] show that there is no restricted dominant μ so that $\dim G < \dim L(\mu) \leq b(G)$, completing the argument for type A_ℓ.

[1] For A_{18} and A_{19}, we refer to the extended table available on Lübeck's web page [2001b].

type of G	e	type of G	e
A_ℓ ($\ell \geq 1$) or B_ℓ ($\ell \geq 3$)	$\ell+1$	G_2	4
C_ℓ ($\ell \geq 2$)	2ℓ	F_4, E_6, E_7, E_8	5
D_ℓ ($\ell \geq 4$)	ℓ		

Table 2. Number of conjugates e needed to generate, as in Proposition 14.1.

For G of type B_ℓ, C_ℓ, or D_ℓ, the argument is the same but easier, with $\ell^3/8$ replaced by ℓ^3.

Suppose now that G has exceptional type. The case $V = L(\tilde{\alpha})$ has been treated in Example 3.4. Otherwise, Tables A.49–A.53 in Lübeck provide the following list of possibilities for V with $b(G) \geq \dim V \geq \dim G - \dim \mathfrak{z}(\mathfrak{g})$, up to graph automorphism and assuming $\operatorname{char} k$ is not special, where we denote the highest weights as in [Lübeck 2001a]: G_2 with highest weight 02 and dimension 26 or 27 (where ρ factors through SO_7 and so is virtually free by Lemma 13.1); G_2 with highest weight 11 and dimension 38 and $\operatorname{char} k = 7$; F_4 with highest weight 0010 and dimension 196 and $\operatorname{char} k = 3$; E_6, with highest weight 000002 or 000010 and dimension 324 or 351. These representations have $\dim V > \dim G$ and are virtually free by [Guerreiro 1997, Theorem 4.3.1]. Note that for any particular V and $\operatorname{char} k$, one can verify that the representation is virtually free using a computer, as described in [Garibaldi and Guralnick 2020a]. □

14. How many conjugates are needed to generate Lie(G)?

The results in the previous section suffice to prove the following, which generalizes the main result (Theorem 8.2) of [Cohen et al. 2001].

Proposition 14.1. *Let G be a simple linear algebraic group over an algebraically closed field k such that* $\operatorname{char} k$ *is not special for G, and let e be as in Table 2. If $x \in \mathfrak{g}$ is noncentral, then there exist e G-conjugates of x that generate a subalgebra containing* $[\mathfrak{g}, \mathfrak{g}]$.

Recall that when G is simply connected (and $\operatorname{char} k$ is not special), $\mathfrak{g} = [\mathfrak{g}, \mathfrak{g}]$ as in Lemma 3.1 (1).

The new results here are types A, D, E, and G_2 when $\operatorname{char} k = 2$. The related result in [Cohen et al. 2001] is stated for long root elements only, but the proof below shows that the long root elements are the main case.

Proof. If G is of exceptional type, we apply Proposition 11.4, so assume that G has type A, B, C, or D.

In the case $(G, \operatorname{char} k) = (\mathrm{SL}_2, 2)$, the subalgebra $[\mathfrak{g}, \mathfrak{g}]$ is the 1-dimensional center of $\mathfrak{sl}_2$, so we prove instead the stronger result that 2 G-conjugates of a noncentral x generate $\mathfrak{sl}_2$.

We first assume that x is a long root element and G is simply connected. For type A_1, a long root element is regular nilpotent and 2 conjugates suffice by Example 6.1. For type A_ℓ (i.e., $G = \mathrm{SL}_{\ell+1}$) with $\ell \geq 2$, $\ell+1$ conjugates suffice by Proposition 6.4 (3) if $\operatorname{char} k \neq 2$ and Proposition 7.1 if $\operatorname{char} k = 2$. For type C_ℓ ($\mathrm{Sp}_{2\ell}$) with $\ell \geq 2$, 2ℓ conjugates suffice by Proposition 8.1 (3). For types B and D, long root elements have rank 2 so Proposition 9.1 gives the claim. If $\operatorname{char} k = 2$ and G has type D_ℓ, then the claim follows for $\mathfrak{so}_{2\ell}$ by Lemma 10.3. The claim follows for groups isogenous to G by Lemma 3.1.

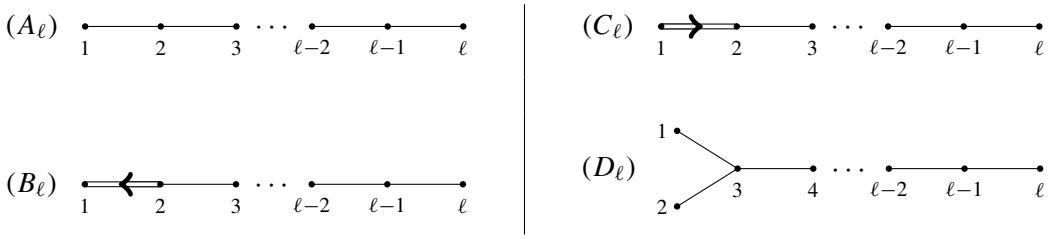


Table 3. Dynkin diagrams of simple root systems of classical type, with simple roots numbered as in [Lübeck 2001a].

If x is nonzero nilpotent, then by Lemma 11.2 and deforming as in Section 4 we are reduced to the previous case.

Generally, x has a Jordan decomposition $x = x_s + x_n$ where x_s is semisimple and x_n is nilpotent and we may assume $x_s \neq 0$. If x_s is noncentral, then we replace x with x_s (whose orbit is closed in the closure of x^G) and then replace x_s with a root element as in Example 4.1.

Therefore, we may assume that $x_s, x_n \neq 0$ and x_s is central. Deforming, it suffices to treat the case where x_n is a root element. The line $tx_s + x_n$ for $t \in k$ has an open subset consisting of elements such that e conjugates suffice to generate a subalgebra containing $[\mathfrak{g}, \mathfrak{g}]$ (resp., $\mathfrak{g}$ in case $(G, \operatorname{char} k) = (\mathrm{SL}_2, 2)$), and this set is nonempty because it contains x_n, so it contains $t_0 x_s + x_n$ for some $t_0 \in k^\times$. The element x_n and $t_0^{-1} x_n$ are in the same G-orbit, so the same is true of x and $x_s + t_0^{-1} x_n$; this proves the claim. □

In the proof, the final paragraph could have been replaced by an argument that maps x into the Lie algebra of the adjoint group of G and applies the result for nilpotent elements there together with Lemma 3.5.

15. The generic stabilizer in G as a group scheme

Let G be an algebraic group over k and $\rho\colon G \to \mathrm{GL}(V)$ a representation. We say that G acts *generically freely* on V if there is a dense open subset U of V such that stabilizer G_u is the trivial group scheme 1 for every $u \in U$. Of course, $\ker \rho \subseteq G_u$ for all u, so it is natural to replace G with $\rho(G)$ and assume that G acts faithfully on V, i.e., $\ker \rho$ is the trivial group scheme.

In this section, we announce results on determining the generic stabilizer as a group scheme when V is faithful and irreducible. The proofs are combinations of the main results in this paper, the sequels [Garibaldi and Guralnick 2020a; 2020b] (which build on this paper), and [Guralnick and Lawther 2019].

Theorem C. *Let $\rho\colon G \to \mathrm{GL}(V)$ be a faithful irreducible representation of a simple algebraic group over an algebraically closed field k.*

(1) *G_v is finite étale for generic $v \in V$ if and only if $\dim V > \dim G$ and (G, V) does not appear in Table 4.*

(2) *G acts generically freely on V if and only if $\dim V > \dim G$ and (G, V) appears in neither Table 4 nor Table 5.*

G	char k	rep'n	dim V	dim $\mathfrak{g}_v$	G	char k	high weight	dim V	dim $\mathfrak{g}_v$
SL_8/μ_4	2	$\wedge^4$	70	3	Sp_8	3	0100	40	2
SL_9/μ_3	3	$\wedge^3$	84	2	Sp_4	5	11	12	1
Spin_{16}/μ_2	2	half-spin	128	4	SL_4	p odd	$01p^e$, $e \geq 1$	24	1
					SL_4/μ_2	2	012^e, $e \geq 2$	24	1

Table 4. Irreducible and faithful representations V of simple G with $\dim V > \dim G$ that are not generically free for $\mathfrak{g}$, up to graph automorphism. For each, the stabilizer $\mathfrak{g}_v$ of a generic $v \in V$ is a toral subalgebra. The weights on the right side are numbered as in Table 3.

Proof. The stabilizer G_v of a generic $v \in V$ is finite étale if and only if the stabilizer $\mathfrak{g}_v$ of a generic $v \in V$ is zero, i.e., if and only if $\mathfrak{g}$ acts generically freely on V. By Theorem A in [Garibaldi and Guralnick 2020a], this occurs if and only if $\dim V > \dim G$ and $(G, \mathrm{char}\, k, V)$ does not appear in Table 4, proving (1).

For (2), we must enumerate in Table 5 those representations V such that $\dim V > \dim G$, V does not appear in Table 4, and the group of points $G_v(k)$ is not trivial. Those V with the latter property are enumerated in [Guralnick and Lawther 2019], completing the proof. □

The results above settle completely the question of determining which faithful irreducible representations of simple G are generically free. It is natural to ask which of these hypotheses are necessary. For example, if char k is special for G, there are irreducible but nonfaithful representations that factor through the very special isogeny; whether or not these are virtually free for $\mathfrak{g}$ is settled in [Garibaldi and Guralnick 2020b]. Another way that G may fail to act faithfully is if V is the Frobenius twist of a representation V_0; in that case $\mathfrak{g}$ acts trivially on V, so G acts virtually freely if and only if the group $G(k)$ of k-points acts virtually freely on V_0.

Combining the results of this paper with [Garibaldi and Guralnick 2020a; 2020b] proves the following extension of Corollary B; see [Garibaldi and Guralnick 2020b] for details.

G	char k	V	dim V	G	char k	V	dim V
A_1	$\neq 2, 3$	S^3	4	A_2	$\neq 2, 3$	S^3	10
A_1	$\neq 2, 3$	S^4	5	A_3	$\neq 2$	$L(2\omega_2)$	$20 - \varepsilon$
A_8	$\neq 3$	$\wedge^3$	84	A_7	$\neq 2$	$\wedge^4$	70
A_3	3	$L(\omega_1 + \omega_2)$	16	A_ℓ	$p \neq 0$	$L(\omega_1 + p^i\omega_\ell)$, $L(\omega_1 + p^i\omega_1)$	$(\ell+1)^2$
B_ℓ $(\ell \geq 2)$	$\neq 2$	$L(2\omega_\ell)$	$2\ell^2 + 3\ell - \varepsilon$	C_4	$\neq 2$	"spin"	$42 - \varepsilon$
D_ℓ $(\ell \geq 4)$	$\neq 2$	$L(2\omega_\ell)$	$2\ell^2 + \ell - 1 - \varepsilon$	D_8	$\neq 2$	half-spin	128

Table 5. Irreducible faithful representations V of simple G with $\dim V > \dim G$ such that G_v is finite étale and $\neq 1$ for generic $v \in V$, up to graph automorphism, adapted from [Guralnick and Lawther 2019]. The symbol ε denotes 0 or 1 depending on the value of char k.

Theorem D. *Let $\rho\colon G \to \mathrm{GL}(V)$ be an irreducible representation of G whose highest weight is restricted. If $\mathfrak{g}$ does **not** act virtually freely on V, then $\dim V < \dim G$ or $\mathfrak{g}_v$ is a toral subalgebra for generic $v \in V$.*

One could ask: *What about analogues of the main results for G semisimple?*

One could also ask for a stronger bound in Theorem A. What is the smallest constant c such that the conclusion holds when we set $b(G) = c \dim G$? What about to guarantee G_v étale? Or $G_v = 1$? Table 4 shows that c must be greater than 1. Does $c = 2$ suffice?

Acknowledgements

We thank the referees for thoughtful remarks, which improved the paper. We also thank Brian Conrad for his helpful advice on group schemes. Part of this research was performed while Garibaldi was at the Institute for Pure and Applied Mathematics (IPAM) at UCLA, which is supported by the National Science Foundation. Guralnick was partially supported by NSF grant DMS-1600056 and DMS-1901595.

References

[Andreev and Popov 1971] E. M. Andreev and V. L. Popov, "Stationary subgroups of points of general position in the representation space of a semisimple Lie group", *Funkcional. Anal. i Priložen.* **5**:4 (1971), 1–8. In Russian; translated in *Funct. Anal. Appl.* **5**:4 (1971), 265–271. MR

[Andreev et al. 1967] E. M. Andreev, E. B. Vinberg, and A. G. Elashvili, "Orbits of greatest dimension in semisimple linear Lie groups", *Funkcional. Anal. i Priložen.* **1**:4 (1967), 3–7. In Russian; translated in *Funct. Anal. Appl.* **1**:4 (1967), 257–261. MR Zbl

[Auld 2001] A. Auld, *Exceptional modular representations of special linear Lie algebras*, Ph.D. thesis, University of Manchester, 2001, https://search.proquest.com/docview/2036850183.

[Breuillard et al. 2012] E. Breuillard, B. Green, R. Guralnick, and T. Tao, "Strongly dense free subgroups of semisimple algebraic groups", *Israel J. Math.* **192**:1 (2012), 347–379. MR Zbl

[Brosnan et al. 2010] P. Brosnan, Z. Reichstein, and A. Vistoli, "Essential dimension, spinor groups, and quadratic forms", *Ann. of Math.* (2) **171**:1 (2010), 533–544. MR Zbl

[Cohen et al. 2001] A. M. Cohen, A. Steinbach, R. Ushirobira, and D. Wales, "Lie algebras generated by extremal elements", *J. Algebra* **236**:1 (2001), 122–154. MR Zbl

[Collingwood and McGovern 1993] D. H. Collingwood and W. M. McGovern, *Nilpotent orbits in semisimple Lie algebras*, Van Nostrand Reinhold, New York, 1993. MR Zbl

[Demazure and Gabriel 1970] M. Demazure and P. Gabriel, *Groupes algébriques, I: Géométrie algébrique, généralités, groupes commutatifs*, Masson & Cie, Paris, 1970. MR Zbl

[Dynkin 1952] E. B. Dynkin, "Semisimple subalgebras of semisimple Lie algebras", *Mat. Sb.* (*N.S.*) **30(72)**:2 (1952), 349–462. In Russian; translated in *Amer. Math. Soc. Transl. (2)* **6** (1957), 111–244. MR Zbl

[Elashvili 1972] A. G. Elashvili, "Canonical form and stationary subalgebras of points in general position for simple linear Lie groups", *Funkcional. Anal. i Priložen.* **6**:1 (1972), 51–62. In Russian; translated in *Funct. Anal. Appl.* **6**:1 (1972), 44–53. MR Zbl

[Garibaldi and Carr 2006] S. Garibaldi and M. Carr, "Geometries, the principle of duality, and algebraic groups", *Expo. Math.* **24**:3 (2006), 195–234. MR Zbl

[Garibaldi and Guralnick 2017] S. Garibaldi and R. M. Guralnick, "Spinors and essential dimension", *Compos. Math.* **153**:3 (2017), 535–556. MR Zbl

[Garibaldi and Guralnick 2020a] S. Garibaldi and R. M. Guralnick, "Generically free representations, II: Irreducible representations", *Tranform. Groups* **25**:3 (2020), 793–817.

[Garibaldi and Guralnick 2020b] S. Garibaldi and R. M. Guralnick, "Generically free representations, III: Extremely bad characteristic", *Tranform. Groups* **25**:3 (2020), 819–841.

[Gordeev and Saxl 2002a] N. Gordeev and J. Saxl, "Products of conjugacy classes in Chevalley groups, I: Extended covering numbers", *Israel J. Math.* **130** (2002), 207–248. MR Zbl

[Gordeev and Saxl 2002b] N. Gordeev and J. Saxl, "Products of conjugacy classes in Chevalley groups, II: Covering and generation", *Israel J. Math.* **130** (2002), 249–258. MR Zbl

[Guerreiro 1997] M. Guerreiro, *Exceptional representations of simple algebraic groups in prime characteristic*, Ph.D. thesis, University of Manchester, 1997, https://search.proquest.com/docview/2036801443.

[Guralnick and Lawther 2019] R. M. Guralnick and R. Lawther, "Generic stabilizers in actions of simple algebraic groups, I: Modules and first Grassmanian varieties", preprint, 2019. arXiv

[Guralnick and Saxl 2003] R. M. Guralnick and J. Saxl, "Generation of finite almost simple groups by conjugates", *J. Algebra* **268**:2 (2003), 519–571. MR Zbl

[Hesselink 1976] W. Hesselink, "Singularities in the nilpotent scheme of a classical group", *Trans. Amer. Math. Soc.* **222** (1976), 1–32. MR Zbl

[Hesselink 1979] W. H. Hesselink, "Nilpotency in classical groups over a field of characteristic 2", *Math. Z.* **166**:2 (1979), 165–181. MR Zbl

[Hogeweij 1978] G. M. D. Hogeweij, *Ideals and automorphisms of almost-classical Lie algebras*, Ph.D. thesis, University of Utrecht, 1978.

[Humphreys 1995] J. E. Humphreys, *Conjugacy classes in semisimple algebraic groups*, Math. Surv. Monogr. **43**, Amer. Math. Soc., Providence, RI, 1995. MR Zbl

[Karpenko 2010] N. A. Karpenko, "Canonical dimension", pp. 146–161 in *Proc. Int. Congress Math., II* (Hyderabad, India, 2010), edited by R. Bhatia et al., Hindustan, New Delhi, 2010. MR Zbl

[Knus 1991] M.-A. Knus, *Quadratic and Hermitian forms over rings*, Grundlehren der Math. Wissenschaften **294**, Springer, 1991. MR Zbl

[Knus et al. 1998] M.-A. Knus, A. Merkurjev, M. Rost, and J.-P. Tignol, *The book of involutions*, Amer. Math. Soc. Colloq. Publ. **44**, Amer. Math. Soc., Providence, RI, 1998. MR Zbl

[Liebeck and Seitz 2012] M. W. Liebeck and G. M. Seitz, *Unipotent and nilpotent classes in simple algebraic groups and Lie algebras*, Math. Surv. Monogr. **180**, Amer. Math. Soc., Providence, RI, 2012. MR Zbl

[Lötscher 2013] R. Lötscher, "A fiber dimension theorem for essential and canonical dimension", *Compos. Math.* **149**:1 (2013), 148–174. MR Zbl

[Lötscher 2015] R. Lötscher, "Essential dimension of separable algebras embedding in a fixed central simple algebra", *Doc. Math.* Merkurjev special issue (2015), 443–459. MR Zbl

[Lötscher et al. 2013] R. Lötscher, M. MacDonald, A. Meyer, and Z. Reichstein, "Essential dimension of algebraic tori", *J. Reine Angew. Math.* **677** (2013), 1–13. MR Zbl

[Lübeck 2001a] F. Lübeck, "Small degree representations of finite Chevalley groups in defining characteristic", *LMS J. Comput. Math.* **4** (2001), 135–169. MR Zbl

[Lübeck 2001b] F. Lübeck, "Tables of Weight Multiplicities", 2001, http://www.math.rwth-aachen.de/~Frank.Luebeck/chev/WMSmall/index.html.

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

[Popov 1987] A. M. Popov, "Finite isotropy subgroups in general position of irreducible semisimple linear Lie groups", *Trudy Moskov. Mat. Obshch.* **50** (1987), 209–248. In Russian; translated in *Trans. Moscow Math. Soc.* **1988** (1988), 205–249. MR Zbl

[Popov 1988] V. L. Popov, "Closed orbits of Borel subgroups", *Mat. Sb.* (*N.S.*) **135(177)**:3 (1988), 385–402. In Russian; translated in *Math. USSR-Sb.* **63**:2 (1989), 375–392. MR Zbl

[Popov and Vinberg 1994] V. L. Popov and E. B. Vinberg, "Invariant theory", pp. 123–278 in *Algebraic geometry, IV: Linear algebraic groups, invariant theory*, edited by A. N. Parshin and I. R. Shafarevich, Encycl. Math. Sci. **55**, Springer, 1994. Zbl

[Premet 1997] A. Premet, "Support varieties of non-restricted modules over Lie algebras of reductive groups", *J. Lond. Math. Soc.* (2) **55**:2 (1997), 236–250. MR Zbl

[Reichstein 2010] Z. Reichstein, "Essential dimension", pp. 162–188 in *Proc. Int. Congress Math., II* (Hyderabad, India, 2010), edited by R. Bhatia et al., Hindustan, New Delhi, 2010. MR Zbl

[Seitz 1987] G. M. Seitz, *The maximal subgroups of classical algebraic groups*, Mem. Amer. Math. Soc. **365**, Amer. Math. Soc., Providence, RI, 1987. MR Zbl

[SGA 3 II 1970] M. Demazure and A. Grothendieck, *Schémas en groupes, Tome II: Groupes de type multiplicatif, et structure des schémas en groupes généraux, Exposés VIII–XVIII* (Séminaire de Géométrie Algébrique du Bois Marie 1962–1964), Lecture Notes in Math. **152**, Springer, 1970. MR Zbl

[SGA 3 III 2011] M. Demazure and A. Grothendieck, *Schémas en groupes, Tome III: Structure des schémas en groupes réductifs, Exposés XIX–XXVI* (Séminaire de Géométrie Algébrique du Bois Marie 1962–1964), revised ed., Doc. Math **8**, Soc. Math. France, Paris, 2011. MR Zbl

[Springer and Steinberg 1970] T. A. Springer and R. Steinberg, "Conjugacy classes", pp. 167–266 in *Seminar on Algebraic Groups and Related Finite Groups* (Princeton, 1968/1969), edited by A. Dold and B. Eckmann, Lecture Notes in Math. **131**, Springer, 1970. MR Zbl

[Steinberg 1963] R. Steinberg, "Representations of algebraic groups", *Nagoya Math. J.* **22** (1963), 33–56. MR Zbl

[Steinberg 2016] R. Steinberg, *Lectures on Chevalley groups*, revised ed., Univ. Lecture Series **66**, Amer. Math. Soc., Providence, RI, 2016. MR Zbl

[Strade and Farnsteiner 1988] H. Strade and R. Farnsteiner, *Modular Lie algebras and their representations*, Monogr. Textbooks Pure Appl. Math. **116**, Dekker, New York, 1988. MR Zbl

[Winter 1977] P. W. Winter, "On the modular representation theory of the two-dimensional special linear group over an algebraically closed field", *J. Lond. Math. Soc.* (2) **16**:2 (1977), 237–252. MR Zbl

Communicated by Pham Huu Tiep
Received 2019-05-06 Revised 2019-11-19 Accepted 2020-02-06

sgaribaldi@fastmail.com *Center for Communications Research, San Diego, CA, United States*

guralnic@usc.edu *Department of Mathematics, University of Southern California, Los Angeles, CA, United States*

msp

ALGEBRA AND NUMBER THEORY 14:6 (2020)
https://doi.org/10.2140/ant.2020.14.1613

Classification of some vertex operator algebras of rank 3

Cameron Franc and Geoffrey Mason

We discuss the classification of strongly regular vertex operator algebras (VOAs) with exactly three simple modules whose character vector satisfies a monic modular linear differential equation with irreducible monodromy. Our main theorem provides a classification of all such VOAs in the form of one infinite family of affine VOAs, one individual affine algebra and two Virasoro algebras, together with a family of eleven exceptional character vectors and associated data that we call the U-series. We prove that there are at least 15 VOAs in the U-series occurring as commutants in a Schellekens list holomorphic VOA. These include the affine algebra $E_{8,2}$ and Höhn's baby monster VOA $\mathrm{VB}^{\natural}_{(0)}$ but the other 13 seem to be new. The idea in the proof of our main theorem is to exploit properties of a family of vector-valued modular forms with rational functions as Fourier coefficients, which solves a family of modular linear differential equations in terms of generalized hypergeometric series.

1. Introduction and statement of the main theorem

It is a natural problem to classify (2-dimensional) rational conformal field theories, which we conflate with the classification of rational vertex operator algebras V (VOAs). In order to do this one needs some *invariants* of V. They should be computable and yet capable of reflecting enough of the structure of V so

Franc was supported by an NSERC Discovery Grant. Mason was supported by grant #427007 from the Simons Foundation.
MSC2010: primary 17B69; secondary 11F03, 17B65.
Keywords: vertex operator algebras, vector-valued modular forms, modular linear differential equations.

that they can distinguish between isomorphism classes of VOAs, or at least come close to this ideal. In fact we work with *strongly regular* VOAs V [Mason 2014]. Among other properties, these are simple VOAs of CFT-type which are also rational and C_2-cofinite. In particular, they have only finitely many (isomorphism classes of) simple modules.

Before continuing, let us develop some notation. If V has n simple modules $V := M_0, M_1, \ldots, M_{n-1}$ it is convenient to say that V *has rank* n. The q-character of M_i is defined in the usual way, namely

$$f_i(\tau) = \mathrm{Tr}_{M_i} q^{L(0)-c/24}.$$

Notation here is standard, and in particular V has central charge c, τ lies in the complex upper half-plane $\mathcal{H}$, and $q := e^{2\pi i\tau}$. The *character vector* of V is the n-vector

$$F(\tau) := (f_0, f_1, \ldots, f_{n-1})^T,$$

and we let $\mathfrak{ch}_V$ denote the *span* of the $f_i(\tau)$. By Zhu's theorem [1996], $\mathfrak{ch}_V$ is a right Γ-module, where $\Gamma := \mathrm{SL}_2(\mathbb{Z})$ and the action is induced by $\gamma : f_i(\tau) \mapsto f_i(\gamma\tau)$ $(\gamma \in \Gamma)$.

Another way to state these facts is in the language of *vector-valued modular forms* (VVMFs): there is representation $\rho : \Gamma \to \mathrm{GL}_n(\mathbb{C})$ such that

$$F(\gamma\tau) = \rho(\gamma)F(\tau),$$

which says that $F(\tau)$ is a VVMF of weight zero on Γ. For a survey of VVMFs, including their connections to Riemann–Hilbert type problems (which we consider below) but *not* their applications to VOAs, we refer the reader to [Franc and Mason 2016a].

A striking property of the character vector is its *modularity* [Huang 2008], which may be stated as follows: the kernel of ρ is a *congruence subgroup* of Γ. This entails that each q-character $f_i(\tau)$ is a modular function of weight zero on some congruence subgroup of Γ. One might therefore think that the character vector could serve as a good invariant for V of the type we are seeking. In fact experience shows that there is a more useful and more subtle invariant that we will explain here: it is a *modular linear differential equation* (MLDE); see [Franc and Mason 2016a]. For the case at hand this may be taken to be a linear differential equation of *weight* k with modular coefficients of the form

$$(P_0D^n + P_1D^{n-1} + \cdots + P_{n-1}D + P_n)u = 0. \tag{1}$$

Here, each $P_\ell \in \mathbb{C}[E_4, E_6]$ is a holomorphic modular form of weight $k + 2\ell - 2n$ and D is the *modular derivative* defined on modular forms of weight k by the formula $D_k = q(d/dq) - (k/12)E_2$. In this paper, since the character vector of a VOA is of weight zero, we require the case where $D = D_0$ and so

$$D^n = D_{2n-2} \circ \cdots \circ D_2 \circ D_0.$$

Then one knows that *there is an MLDE of some weight* k *whose solution space is* $\mathfrak{ch}_V$.

The MLDE (1) may be taken as the desired invariant of V. Not only does it implicitly include $\mathfrak{ch}_V$ as the space of solutions of (1), but in addition it carries a *monodromy representation* defined by analytic

continuation of the solutions around the singularities. Because of the special nature of the differential equation (1) this monodromy is essentially the representation ρ of Γ acting on $\mathfrak{ch}_V$.

The purpose of the present paper is to prove the analog of the Mathur–Mukhi–Sen theorem [Mathur et al. 1988; Mason et al. 2018] in rank 3. The extra dimension gives rise to a great deal of additional complication and difficulties. Some of these were discussed in [Mason 2020] where our main theorem appeared as Problem 4. In particular, while it has long been recognized that VOAs have a strong arithmetic vein, the current proof of Theorem 1 (the main theorem) includes an unprecedented amount of number theoretic complications.

We shall now state our main result precisely and outline its proof: we characterize strongly regular VOAs V of rank 3 whose associated MLDE (1) has weight 0 so that it takes the form

$$(D^3 + aE_4D + bE_6)u = 0, \quad (a, b \in \mathbb{C}).$$

An MLDE of weight zero such as this is said to be *monic*. Additionally, we assume that the monodromy ρ is an *irreducible* representation of Γ. With these conditions and definitions we establish the following main result:

Theorem 1 (rank 3 Mathur–Mukhi–Sen). *Let V be a strongly regular VOA with exactly three simple modules. Suppose that the q-characters of the simple V-modules furnish a fundamental system of solutions for an MLDE of order 3 that is*

(i) *monic, and*

(ii) *has irreducible monodromy.*

Then one of the following holds:

(a) *V is isomorphic to one of the following*:

$$B_{\ell,1} \quad (\ell \geq 2), \qquad A_{1,2}, \qquad \mathrm{Vir}(c_{3,4}),\ \mathrm{Vir}(c_{2,7}).$$

(b) *V lies in the U-series* (*see Remark 2*).

(*Here, and below, $\mathcal{G}_{\ell,k}$ denotes an affine algebra of type $\mathcal{G}$, rank ℓ, and level k*; $\mathrm{Vir}(c)$ *is a Virasoro VOA of central charge c.*)

Remark 2. The U-series[1] refers both to 11 sets of datum indexed by an integer k in the range $0 \leq k \leq 10$, and to a family of VOAs uniformly described by the data, each of which satisfies the hypotheses of Theorem 1. The data arises from the residual cases in our approach to the proof of Theorem 1.

Two VOAs in the U-series are well-known. These are the affine algebra $E_{8,2}$ and the baby monster VOA $\mathrm{VB}^\natural_{(0)}$ [Höhn 1996]. These two VOAs correspond to $k = 8$ and $k = 0$ respectively.

We will show that 13 additional VOAs in the U-series, corresponding to $k = 1, 2, \ldots, 6$, may also be constructed as commutants in a Schellekens list VOA. This is strongly suggested by, and depends on, the

[1] In an earlier preprint U stood for "unknown" or "undecided". Although the question of existence is now decided in many cases — subject to a standard conjecture — it is still a useful mnemonic.

work of Gaberdiel, Hampapura and Mukhi [Gaberdiel et al. 2016] and Xingjun Lin [2017]. For further details we refer the reader to Section 9.

Remark 3. Since the original submission of this paper we have been able to prove that Theorem 1 remains true without the irreducibility assumption (ii). Were we to include details, however, it would significantly add to the length of the present paper, so we skip them here.

The idea of classifying 2-dimensional conformal field theories is an old dream of physicists, dating from the late 1980s, and the influential paper of Moore and Seiberg [1989] is often cited in this regard. The idea of attacking the problem based on the method of MLDEs as we have explained it was propounded by Mathur, Mukhi and Sen [1988], where they discussed the classification of rank 2 VOAs at the level of physical rigor. Until recently mathematicians have hesitated to get on this bandwagon, perhaps because of the lack of a sufficiently solid theory of MLDEs and VVMFs, however that trend has now reversed itself. The rank 2 theory of Mathur, Mukhi and Sen was put on a solid mathematical foundation in [Mason et al. 2018], and Tener and O'Grady [2018] extended this in developing the theory of rank 2 *extremal* VOAs.

As for the rank 3 theory treated here, Theorem 1 (our main theorem) subsumes a number of results in both the mathematical and physical literature. Hampapura and Mukhi [2016] treated the baby monster VOA from the MLDE perspective. This example together with $E_{8,2}$ was considered by Gerald Höhn [1996]. Gaberdiel, Hampapura and Mukhi [2016] also constructed several VOAs related to, and conjecturally equal to, some VOAs in the U-series. In Appendix C of [Mathur et al. 1989] one finds a discussion of the infinite series of affine algebras intervening in Theorem 1. Arike, Nagatomo, Kaneko and Sakai [Arike et al. 2016] discussed the MLDEs satisfied by these and many other affine algebras. Arike, Nagatomo and Sakai [Arike et al. 2017] characterized some low-dimensional Virasoro algebras according to their MLDEs, and the results of both this and a preprint of Mason, Nagatomo and Sakai [Mason et al. 2018] characterizing some VOAs with $c = 8$ or 16 are special cases of Theorem 1.

Theorem 1 is proved by exploiting the fact, proved in [Franc and Mason 2016a], that a monic MLDE of degree three can be solved in terms of generalized hypergeometric series. This solution describes an algebraic family of modular forms that vary according to choices of local exponents at the cusp for the MLDE. The important point for our analysis is that this family of modular forms has Fourier coefficients that are *rational functions* of the local exponents. Since our goal is to classify specializations of the family that have Fourier coefficients that are nonnegative integers, we proceed as follows:

(1) It is known that the monodromy representation is congruence, and in Section 3 we give a direct proof of this fact (see Theorem 7). Indeed, together with the results of [Franc and Mason 2016b], our results establish the *unbounded denominator conjecture* for 3-dimensional irreducible representations of the modular group (whereas [Franc and Mason 2016b] treated the case of imprimitive representations). The 2-dimensional case was proved in [Franc and Mason 2014]. The main result of Section 3 details the 3-dimensional irreducible representations of Γ and makes precise some computations from [Beukers and Heckman 1989].

(2) Next in Section 5 we study the divisors of the first nontrivial Fourier coefficients of the character vector. The signs of the coefficients are constant on the connected components of the complement of the divisors, so that we may restrict our search to a reasonably small and manageable subset of all possible parameters. This is explained in Theorem 21 and displayed graphically in Figure 2 on page 1635.

(3) The remaining characters are tested for integrality in Section 6, where we use arithmetic properties of hypergeometric series discussed in [Franc et al. 2018]. The output is one infinite family of possible character vectors, in addition to a finite list of additional exceptional possibilities tabulated in Tables 4 and 5 on pages 1644 and 1645.

(4) Next in Section 7 we apply further tests arising from the theory of VOAs, namely symmetry of the S-matrix and the Verlinde formula [Huang 2008], to whittle the remaining examples down to the statement of Theorem 1.

(5) In Section 8 we complete the proof of Theorem 1 by discussing the infinite family of possible character vectors, and we explain how they are in fact realized by VOAs.

Finally, Section 9 discusses the U-series.

It is worth noting that a significant feature of our proof, indeed, of the general approach to VOA classification through VVMFs and MLDEs, is the difficulty in distinguishing VOAs that have *more* than three simple modules but which satisfy $\dim \mathfrak{ch}_V = 3$. A good part of our proof goes through under the weaker assumption that $\dim \mathfrak{ch}_V = 3$. But in order to readily apply the symmetry of the S-matrix we must assume that V has rank 3. A similar circumstance already revealed itself in [Mason et al. 2018] in the rank 2 case.

It is well-known that the VOAs listed in Theorem 1 are strongly regular, have exactly three simple modules, and satisfy the other conditions of Theorem 1. For the case of the affine algebras this is easily deduced from [Arike et al. 2016] and for the Virasoro algebras, see, e.g., [Lepowsky and Li 2004]. In Table 1 on page 1618 we have collected some relevant data for these VOAs.

2. Background on VOAs

2.1. ***The invariants c, $\tilde{c}$, ℓ.*** In this subsection we discuss the numerical invariants c, $\tilde{c}$ and ℓ associated with a strongly regular VOA $V = (V, Y, \mathbf{1}, \omega)$ that we will use in the following sections. For additional background and discussion we refer the reader to [Mason 2014]. We note that one of our results, Theorem 4, is new and improves upon an inequality of Dong and Mason [2004]. In this subsection we do *not* make any assumptions about the number of irreducible modules that V may have, merely that they are finite in number.

The invariant c, the *central charge* of V, is of course well-known and a standard invariant that is part of the definition of V. We sometimes write $c = c_V$. Because V is strongly regular then it has only finitely many (isomorphism classes of) irreducible modules, which we label as $M_0, M_1, \ldots, M_{n-1}$. And because V is necessarily simple then one of the M_i is isomorphic to V, and we will always choose notation so that $V = M_0$. Each M_i has a *conformal weight* h_i defined to be the least nonvanishing eigenvalue of the $L(0)$-operator. Thus M_i has (conformal) grading $M_i = \bigoplus_{n\geq 0} M_{n+h_i}$, and the *q-character* of M_i is defined by

$$\text{q-char } M_i := \text{Tr}_{M_i}\, q^{L(0)-c/24} = q^{h_i - c/24} \sum_{n\geq 0} \dim(M_i)_{n+h_i} q^n. \tag{2}$$

VOA	c	h_1	h_2
$A_{1,2}$	$\frac{3}{2}$	$\frac{3}{16}$	$\frac{1}{2}$
$B_{\ell,1},\ \ell \geq 2$	$\frac{1}{2}(2\ell+1)$	$\frac{1}{16}(2\ell+1)$	$\frac{1}{2}$
$E_{8,2}$	$\frac{31}{2}$	$\frac{3}{2}$	$\frac{15}{16}$
$\mathrm{Vir}(c_{2,7})$	$-\frac{68}{7}$	$-\frac{2}{7}$	$-\frac{3}{7}$
$\mathrm{Vir}(c_{3,4})$	$\frac{1}{2}$	$\frac{1}{16}$	$\frac{1}{2}$

Table 1. Some VOAs with three simple modules.

Throughout we use the notation

$$m := \dim V_1.$$

In particular, and as part of the definition of a strongly regular VOA, we have

$$\text{q-char } V := \mathrm{Tr}_V\, q^{L(0)-c/24} = q^{-c/24} \sum_{n\geq 0} \dim V_n q^n = q^{-c/24}(1 + mq + \cdots)$$

We note that c and each h_i lies in $\mathbb{Q}$, the field of rational numbers [Dong et al. 2000].

The *effective central charge* $\tilde{c} = \tilde{c}_V$ is defined as

$$\tilde{c} := c - 24h_{\min},$$

where $h_{\min}$ is the *least* of the rational numbers h_i. Note that $h_0 = 0$ by our convention, in particular we always have $c \leq \tilde{c}$, and of course $\tilde{c} \in \mathbb{Q}$. The effective central charge will play an important rôle in our efforts to characterize certain VOAs. Its relevance is partially explained by noticing that among the set of q-characters (2), the *least* of the leading q-powers is precisely $q^{-\tilde{c}/24}$.

The invariant ℓ is defined to be the *Lie rank* of V_1. It is well-known that the homogeneous space V_1 of a strongly regular VOA carries the structure of a Lie algebra with respect to the bracket $[ab] := a(0)b$. Indeed, V_1 is a *reductive Lie algebra* [Dong and Mason 2004]. Then ℓ is the dimension of a Cartan subalgebra of V_1. The following equality involving ℓ and $\tilde{c}$ is known [loc. cit.]

$$\tilde{c} \geq \ell, \quad \text{and} \quad \tilde{c} = 0 \quad \text{only if} \quad V = \mathbb{C}.$$

In particular, if $V \neq \mathbb{C}$ then at least one of the q-characters (2) has a *pole* at $q = 0$.

In [Dong and Mason 2004] it was shown that the simultaneous equalities $c = \tilde{c} = \ell$ characterize lattice VOAs V_Λ (some positive-definite even lattice Λ), and the authors expected that the equality $\tilde{c} = \ell$ should suffice to characterize this class of VOAs. Here, we shall prove this and more.

Theorem 4. *Suppose that V is a strongly regular VOA satisfying $\tilde{c} < \ell + 1$. Then $c = \tilde{c}$. In particular, if $\tilde{c} = \ell$ then V is isomorphic to a lattice theory V_Λ for some even lattice Λ.*

Proof. We shall do this by modifying the proof of Theorem 7 of [Mason 2014]. Theorem 1 of [Mason 2014] says that V contains a subVOA $T \subseteq V$ with the following properties:

(a) T is a *conformal subalgebra* of V, i.e., V and T have the *same* Virasoro element, and in particular $c_V = c_T$.

(b) T is a tensor product $T \cong W \otimes C$ of a pair of subVOAs W isomorphic to a lattice theory V_Λ of rank ℓ, and C isomorphic to a discrete series Virasoro VOA $\mathrm{Vir}(c_{p,q})$.

Actually, in this set-up we have $0 \le c_{p,q} < 1$, so that C is in the *unitary* discrete series. We have the following series of inequalities that proves what is needed:

$$c_V \le \tilde{c}_V \le \tilde{c}_T = \tilde{c}_{V_\Lambda} \tilde{c}_{\mathrm{Vir}(c_{p,q})} = c_{V_\Lambda} c_{\mathrm{Vir}(c_{p,q})} = c_T = c_V.$$

Here, the first inequality was pointed out before; the second inequality holds because T is a conformal subalgebra of V; the first equality holds because effective central charge is multiplicative over tenor products; the second equality holds because central charge and effective central charge *coincide* for both lattice theories and unitary discrete series of Virasoro VOAs; and finally the third equality holds because central charge is also multiplicative over tensor products. □

As a corollary of this proof, we have:

Corollary 5. *Suppose* $2c \in \mathbb{Z}$. *Then one of the following holds*:

(a) $\tilde{c} - \ell \ge 1$;

(b) $\tilde{c} - \ell = \frac{1}{2}$ *and* $\tilde{c} = c$;

(c) $\tilde{c} = \ell = c$.

Proof. If (a) is false then $\tilde{c} - \ell < 1$ and Theorem 4 tells us that $\tilde{c} = c$. Moreover, as the proof shows, V contains a conformal subVOA isomorphic to $V_\Lambda \otimes \mathrm{Vir}(c_{p,q})$ where Λ is an even lattice of rank ℓ. The Virasoro tensor factor lies in the unitary discrete series because its central charge is less than 1. It follows that there is an integer $z \ge 2$ such that

$$c = \ell + 1 - \frac{6}{z(z+1)}.$$

Because $2c \in \mathbb{Z}$, this can only happen if $z = 2$ or 3. These two possibilities correspond to (c) and (b), respectively. This completes the proof. □

2.2. ***The space*** $\mathfrak{ch}_V$ ***of q-characters.*** We retain the notation of the previous subsection and in particular V denotes a strongly regular VOA. For the rest of this subsection we assume that $\dim \mathfrak{ch}_V = 3$ and that $\mathfrak{ch}_V$ is the solution space of a *monic* MLDE that has an *irreducible* monodromy representation $\rho : \mathrm{SL}_2(\mathbb{Z}) \to \mathrm{GL}(\mathfrak{ch}_V)$; see [Franc and Mason 2016a; 2016b]. In particular, the MLDE in question must look like

$$(D_0^3 + aE_4 D_0 + bE_6) f = 0, \quad a, b \in \mathbb{Q}. \tag{3}$$

Here E_4 and E_6 are the holomorphic Eisenstein series of level one and weights 4 and 6, respectively, normalized so that the constant terms are 1. In [Franc and Mason 2016a; 2016b] this MLDE arose as the differential equation satisfied by forms of minimal weight for ρ. It is worth noting that the form of

minimal weight for a given representation (and choice of exponents for $\rho(T)$) is rarely 0, so that the modular forms arising as character vectors of VOAs are almost never of minimal weight. Nevertheless, the computations of [Franc and Mason 2016a; 2016b] may be used to study the solutions of (3), and we discuss this next.

Because ρ is irreducible it is easy to see, and it is a special case of a result of [Tuba and Wenzl 2001], that the T-matrix $\rho(T)$ has *distinct eigenvalues*. A general result [Dong et al. 2000] says that $\rho(T)$ has *finite order* (although in the present context this can be seen more directly), and in any case there are *distinct* $r_0, r_1, r_2 \in \mathbb{Q} \cap [0, 1)$ and a basis of $\mathfrak{ch}_V$ such that if we assume that ρ is written with respect to this choice of basis then

$$\rho(T) = \begin{pmatrix} e^{2\pi i r_0} & 0 & 0 \\ 0 & e^{2\pi i r_1} & 0 \\ 0 & 0 & e^{2\pi i r_2} \end{pmatrix}. \tag{4}$$

Because $\mathfrak{ch}_V$ *spans* the solution space of the MLDE (3) then it is easy to see that the three eigenfunctions for $\rho(T)$ may be taken to be the q-characters of three irreducible V-modules, and that moreover we may take the first of these V-modules to be $V = M_0$. Let M_1, M_2 be the other two irreducible V-modules. The *character* vector of V is thus the vector-valued modular form

$$F(\tau) := \begin{pmatrix} f_0(\tau) \\ f_1(\tau) \\ f_2(\tau) \end{pmatrix},$$

where

$$f_0(\tau) := \mathrm{Tr}_V\, q^{L(0)-c/24} = q^{-c/24} + O(q^{1-c/24}),$$
$$f_i(\tau) := \mathrm{Tr}_{M_i}\, q^{L(0)-c/24} = \dim(M_i)_{h_i} q^{h_i - c/24} + O(q^{1+h_i-c/24}), \quad i = 1, 2,$$

and furthermore

$$r_0 \equiv -\frac{c}{24} \pmod{\mathbb{Z}},$$
$$r_i \equiv h_i - \frac{c}{24} \pmod{\mathbb{Z}}, \quad i = 1, 2.$$

There is an important identity that accrues from the special shape of the MLDE (3), namely:

Lemma 6. *The following hold*:

(a) $c = 8\big(h_1 + h_2 - \frac{1}{2}\big)$.

(b) $\det \rho(T) = -1$.

Proof. (a) The *indicial equation* (at ∞) for (3) is readily found to be

$$x^3 - \tfrac{1}{2}x^2 + \big(a + \tfrac{1}{18}\big)x + b = 0,$$

and in particular the corresponding indicial roots sum to $\frac{1}{2}$. However these roots are the leading exponents of q for the functions $f_i(\tau)$ $(i = 0, 1, 2)$, namely $-\frac{c}{24}$, $h_1 - \frac{c}{24}$ and $h_2 - \frac{c}{24}$. Part (a) follows immediately.

As for (b), using (a) we have $\det \rho(T) = e^{2\pi i(r_0+r_1+r_2)} = e^{2\pi i(h_1+h_2-c/8)} = -1$. □

2.3. ***Things hypergeometric.*** It is fundamental for this paper that with a suitable change of variables the MLDE (3) becomes a generalized hypergeometric differential equation that is solved by generalized hypergeometric functions ${}_3F_2$. This circumstance is explained in [Franc and Mason 2016a; 2016b], where, in particular, motivation for using the level 1 hauptmodul $K : \mathcal{H} \cup \{\infty\} \to \mathbb{P}^1(\mathbb{C})$ defined by

$$K = \frac{E_4^3}{E_4^3 - E_6^2} = \frac{1728}{j} = 1728q + \cdots$$

is provided. The well-known [Beukers and Heckman 1989], which describes the monodromy of all generalized hypergeometric differential equations of all orders, may also be referenced here. We shall only need the case of order 3. In terms of the differential operator $\theta_K := K(d/dK)$, the MLDE (3) becomes (see [Franc and Mason 2016a, Example 15])

$$\left(\theta_K^3 - \frac{2K+1}{2(1-K)}\theta_K^2 + \frac{18a+1-4K}{18(1-K)}\theta_K + \frac{b}{1-K}\right) f = 0.$$

Following [Beukers and Heckman 1989, Section 2], upon multiplying the previous differential operator by $1-K$ we obtain the following alternate formulation:

$$\left\{(\theta_K+\beta_1-1)(\theta_K+\beta_2-1)(\theta_K+\beta_3-1) - K(\theta_K+\alpha_1)(\theta_K+\alpha_2)(\theta_K+\alpha_3)\right\} f = 0$$

for scalars $\alpha_1, \ldots, \beta_3$ satisfying

$$\begin{gathered}\alpha_1+\alpha_2+\alpha_3 = 1, \quad \alpha_1\alpha_2+\alpha_1\alpha_3+\alpha_2\alpha_3 = \tfrac{2}{9}, \quad \alpha_1\alpha_2\alpha_3 = 0, \quad \beta_1+\beta_2+\beta_3-3 = -\tfrac{1}{2}, \\ (\beta_1-1)(\beta_2-1)+(\beta_1-1)(\beta_3-1)+(\beta_2-1)(\beta_3-1) = \tfrac{1}{18}+a, \quad (\beta_1-1)(\beta_2-1)(\beta_3-1) = b.\end{gathered} \tag{5}$$

The local indices at the three singularities $K = 0, 1, \infty$ are

$$\begin{array}{ccccc} 1-\beta_1, & 1-\beta_2, & 1-\beta_3 & & \text{at } K=0, \\ \alpha_1 = 0, & \alpha_2 = \frac{1}{3}, & \alpha_3 = \frac{2}{3} & & \text{at } K=\infty, \\ 0, & 1, & \frac{1}{2} & & \text{at } K=1. \end{array} \tag{6}$$

Inasmuch as $K(\infty) = 0$, $K(e^{2\pi i/3}) = \infty$ and $K(i) = 1$, these sets of indices correspond to the local monodromies $\rho(T)$, $\rho(R)$, $\rho(S)$ respectively (where $R = -ST$ — see Section 3.1 below for the notation). For example, we see that

$$\det\rho(T) = -1, \quad \det\rho(R) = 1, \quad \det\rho(S) = -1.$$

The generalized hypergeometric function ${}_3F_2$ is defined by

$${}_3F_2(a_1, a_2, a_3; b_1, b_2; z) := 1 + \sum_{n\geq 1} \frac{(a_1)_n(a_2)_n(a_3)_n}{(b_1)_n(b_2)_n}\frac{z^n}{n!},$$

where $(t)_n := t(t+1)\cdots(t+n-1)$ is the rising factorial. Here, a_1, a_2, a_3, b_1, b_2 are arbitrary scalars subject to the exclusion that b_1, b_2 are *not* nonpositive integers. With this convention, ${}_3F_2$ converges for $|z| < 1$, has singularities at $z = 0, 1, \infty$, and is defined by analytic continuation elsewhere.

With the assumption that *no two of the β_i differ by an integer*, a fundamental system of solutions near $K = 0$ of our hypergeometric differential equation is given as in equation (2.9) of [Beukers and Heckman 1989] by

$$\begin{aligned} &K^{1-\beta_1}{}_3F_2(1+\alpha_1-\beta_1, 1+\alpha_3-\beta_1, 1+\alpha_1-\beta_1; 1+\beta_2-\beta_1, 1+\beta_3-\beta_1; K) \\ &K^{1-\beta_2}{}_3F_2(1+\alpha_1-\beta_2, 1+\alpha_3-\beta_2, 1+\alpha_1-\beta_2; 1+\beta_1-\beta_2, 1+\beta_3-\beta_2; K) \\ &K^{1-\beta_3}{}_3F_2(1+\alpha_1-\beta_3, 1+\alpha_3-\beta_3, 1+\alpha_1-\beta_3; 1+\beta_1-\beta_3, 1+\beta_2-\beta_3; K) \end{aligned} \tag{7}$$

In this way one obtains explicit and useful formulas for the character vector $F(\tau)$ of Section 2.2. We shall exploit this hypergeometric formula, which describes a family of vector-valued modular forms varying over a space of indices for the differential equation (3), to classify possible character vectors of VOAs having exactly 3 irreducible modules and irreducible monic monodromy. The key points are that the Fourier coefficients of this family are rational functions in the local indices, and that the arithmetic behavior of these coefficients are very well-studied; see [Dwork 1990; Franc et al. 2018].

3. Classification of the monodromy

The purpose of this section is to enumerate the possible monodromies ρ of the MLDE attached to $\mathfrak{ch}_V$ (see Section 2.2). Essentially, this amounts to cataloging certain equivalence classes of 3-dimensional irreducible representations of $\mathrm{SL}_2(\mathbb{Z})$. We shall do this, and in particular we will calculate the possible sets of exponents r_i of the T-matrix (4). These rational numbers (and in particular their *denominators*) will play an important rôle in the arithmetic analysis in later sections.

Beukers and Heckman [1989] described the monodromy of all hypergeometric functions ${}_nF_{n-1}$, so in principle they already solved the problem that concerns us in this section because, as we have explained, our MLDE is hypergeometric. However there are several reasons why we prefer to develop our results independently. Firstly, the results of Beukers and Heckman are couched indirectly in terms of what they refer to as *scalar shifts*, making their general answer that applies to all ranks too imprecise for our specific purpose. Secondly, they work with representations of the free group of rank 2 whereas our monodromy groups factor through the modular group $\mathrm{SL}_2(\mathbb{Z})$. So the question of the *modularity* of ρ does not arise in [Beukers and Heckman 1989]. Finally, we anticipate that the details of our explicit enumeration will be useful in further work involving MLDEs of order 3.

Some of the main arithmetic results are summarized in the following:

Theorem 7. *Let V be a strongly regular VOA V and suppose that the third order MLDE (3) associated with $\mathfrak{ch}_V$ is monic with **irreducible** monodromy representation ρ. Then ρ is a **congruence representation**, and one of the following holds*:

(1) *ρ is **imprimitive** and both h_1 and h_2 are rational with denominators dividing 16. Moreover, either*
   (a) *one of h_1 or h_2 lies in $\frac{1}{2}\mathbb{Z}$, or*
   (b) *the denominators of h_1 and h_2 are equal to each other.*

(2) ρ *is* ***primitive*** *and the denominators of* h_1 *and* h_2 *are both equal to each other and to one of* 5 *or* 7.

We describe how to classify the representations of Theorem 7, and give more detailed information about them, in the following sections.

3.1. ***Some generalities.*** We begin with some general facts about Γ and the representation ρ that we shall need.

Let $\Gamma := \mathrm{SL}_2(\mathbb{Z})$ and let U be the left $\mathbb{C}[\Gamma]$-module furnished by the representation ρ of Γ associated to our MLDE (3). In effect, $U = \mathfrak{ch}_V$, though this particular realization of U will be unhelpful in this subsection. We use the following notation for elements in Γ:

$$R := \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}, \quad S := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Lemma 8. *The following hold*:

(a) *If* $\gamma \in \Gamma$ *then* $\det \rho(\gamma) = \pm 1$.

(b) $\rho(S^2) = I$.

Proof. Because ρ is irreducible then $\rho(R)$ has the 3 cube roots of unity as eigenvalues, and in particular $\det \rho(R) = 1$. However $\Gamma = \langle R, T \rangle$, and we have seen in Lemma 6(b) that $\det \rho(T) = -1$. Now part (a) of the present lemma follows.

To prove part (b) assume that it is false. Then $\rho(S^2) = -I$, and it follows from (a) that there is a subgroup $G \trianglelefteq \Gamma$ of index 2 such that $\Gamma = G \times \langle S^2 \rangle$. But this is impossible, because G must contain the congruence subgroup $\Gamma(2)$, whereas $S^2 \in \Gamma(2)$. This completes the proof. □

Part (b) informs us that ρ is an *even* representation, i.e., it factors through the quotient $\mathrm{PSL}_2(\mathbb{Z}) := \Gamma/\langle \pm I \rangle$. Furthermore, we have:

Corollary 9. *The subgroup of* $\rho(\Gamma)$ *that acts on* U *with determinant* 1 *has index* 2.

Proof. This follows from Lemmas 8(a) and 6(b). □

The next result is well-known. We give a proof for completeness.

Lemma 10. *The following hold*:

(a) *Suppose that* $N \trianglelefteq \Gamma$ *and that* $\Gamma/N \cong L_2(7)$. *Then* $N = \Gamma(7)\langle S^2 \rangle$.

(b) $A_6 \cong L_2(9)$ *is not a quotient of* Γ.

Proof. The proofs of each of these assertions are essentially the same. We deal with (a) and skip the proof of (b). We may, and shall, calculate in the group $\Gamma/\{\pm I\}$.

Part (a) is essentially explained by the *automorphism group* $\mathrm{PGL}_2(7)$ of $L_2(7)$, which has order 336.

Count ordered pairs of elements of orders 2 and 3 that generate the abstract group $L_2(7)$: if this set is denoted by X, we claim that X is a $\mathrm{PGL}_2(7)$-*torsor*, i.e., $\mathrm{PGL}_2(7)$ acts transitively (by conjugation) on X and $|X| = |\mathrm{PGL}_2(7)|$. The action is evident, so it suffices to check the cardinality of X.

For example, the total number of pairs of elements of order 2 and 3 respectively equal $21 \cdot 56$, whereas the number of S_3-pairs is $6 \cdot 28$, the number of A_4-pairs is $2 \cdot 7 \cdot 24$, and the number of S_4-pairs is $2 \cdot 7 \cdot 24$. Therefore we find that the number of $L_2(7)$-pairs is $21 \cdot 56 - 12(14 + 28 + 28) = 336$.

Finally, let $\nu : \Gamma/\{\pm I\} \to L_2(7)$ be reduction mod 7, and let $\varphi : \Gamma/\{\pm I\} \to L_2(7)$ be any surjection.

$$\begin{array}{ccc} \Gamma/\{\pm I\} & \xrightarrow{\nu} & L_2(7) \\ & \searrow^{\varphi} & \downarrow{\scriptstyle \alpha} \\ & & L_2(7) \end{array}$$

Because X is a $\mathrm{PGL}_2(7)$-torsor, there is $\alpha \in \mathrm{PGL}_2(7)$ that makes the diagram commute. Therefore, $\varphi = \alpha \circ \nu$ has kernel $\Gamma(7)\langle S^2\rangle/\langle S^2\rangle$. This completes the proof of part (a). □

3.2. ***The imprimitive case.*** Suppose that $N \trianglelefteq \Gamma$ is a normal subgroup. Suppose further that the *restriction* $U|_N$ of U to N is *not* irreducible. Then there is a direct sum decomposition into 1-dimensional N-submodules

$$U|_N \cong U_0 \oplus U_1 \oplus U_2$$

and there are just two possibilities for the *Wedderburn structure*, namely

(i) (one Wedderburn component) the U_j are pairwise *isomorphic* as N-modules;

(ii) (three Wedderburn components) the U_j are pairwise *nonisomorphic* as N-modules, and they are transitively permuted among themselves by the action of Γ.

Care is warranted because the U_j may *not* be the three T-eigenspaces. If case (ii) pertains, the representation ρ is called *imprimitive*. Otherwise, it is *primitive*.

Lemma 11. *Suppose that N has one Wedderburn component. Then $\rho(N) \subseteq Z(\rho(\Gamma))$ and $\rho(N)$ is isomorphic to a subgroup of $\mathbb{Z}/6\mathbb{Z}$.*

Proof. By hypothesis, each element $\gamma \in N$ is such that $\rho(\gamma)$ acts on each W_j as multiplication by the *same* scalar. In other words, $\rho(\gamma)$ is a scalar matrix. As such it lies in the center $Z(\rho(\Gamma))$. This proves the first assertion of the lemma. Suppose that λ is the eigenvalue for such a $\rho(\gamma)$. Then we must have $\lambda^6 = 1$ by Corollary 9, and the second assertion of the lemma follows. □

We now assume that ρ is imprimitive, and choose a maximal element K in the poset of normal subgroups $K_1 \trianglelefteq \Gamma$ with the property that $U|_{K_1}$ is not irreducible. Let the Wedderburn decomposition be

$$U|_K = W_0 \oplus W_1 \oplus W_2.$$

Note that elements of K are represented by *diagonal matrices*, whence $\rho(K)$ is *abelian*.

By assumption, Γ permutes the subspaces W_j among themselves and acts transitively on this set. The *kernel* of this action is a normal subgroup leaving each W_j invariant, and by the maximality of K, it is none other than K itself. Hence Γ/K is isomorphic to one of $\mathbb{Z}/3\mathbb{Z}$ or S_3, being a transitive subgroup of S_3 in its action on 3 letters.

It follows from the previous paragraph that one of the powers T^s $(s = 1, 2, 3)$ lies in K. It is well-known (e.g., [Knopp et al. 1965]) that the *normal closure* of T^s in Γ is the principal congruence subgroup $\Gamma(s)$. Hence $\Gamma(s) \subseteq K$. Now note that because $K \neq \Gamma$ then $s \neq 1$.

Next we show that the assumption $\Gamma/K \cong \mathbb{Z}/3\mathbb{Z}$ leads to a contradiction, so assume it is true. Then K is the unique normal subgroup of index 3, and as such it has just three classes of subgroups of order 4 which generate K. It follows that $K/K'\langle S^2\rangle \cong (\mathbb{Z}/2\mathbb{Z})^2$. But $\rho(K)$ is abelian, hence $\rho(K) \cong (\mathbb{Z}/2\mathbb{Z})^2$, $K = \Gamma(3)\langle S^2\rangle$, and $\Gamma/K \cong A_4$. But then $\rho(T)$ has order 3, contradicting Lemma 6(b).

This reduces us to the case when $\Gamma/K \cong S_3$. Suppose also that $s = 3$. Then R and T jointly generate a subgroup of index 2 in Γ, a contradiction because they are generators of Γ. It follows that $s = 2$. In this case we must have $K = \Gamma(2)$ because $\Gamma/\Gamma(2) \cong S_3$. Now $\Gamma(2)/\langle S^2\rangle$ is a free group of rank 2. Therefore because $\rho(K)$ is abelian it is a homocyclic quotient of $\mathbb{Z}^2$ (remember that $\rho(S^2) = I$). Now because $\rho(T)$ has distinct eigenvalues, then it *cannot* have order 2. Therefore $\rho(T^2)$ is a nonidentity torsion element of $\rho(K)$. This implies that $\rho(K) \cong (\mathbb{Z}/t\mathbb{Z})^2$ for some integer t, and in particular $\rho(\Gamma)$ is *finite* (of order $6t^2$).

At this point we have maneuvered ourselves into a position where we can apply the results of [Franc and Mason 2016b] concerning finite-image, imprimitive, irreducible representations of $\Gamma/\langle S^2\rangle$. Indeed, setting $H = \Gamma_0(2)$, ρ is an induced representation $\rho = \mathrm{Ind}_H^\Gamma \chi$ for some linear character

$$\chi : \Gamma_0(2) \to \mathbb{C}^\times.$$

of *finite order*. In the notation of [Franc and Mason 2016b], there is a positive integer n and a primitive n-th root of unity λ such that

$$\chi(U) = \lambda, \quad \chi(V) = 1, \quad \chi(S^2) = 1,$$

where the images of $U := \left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$ and $V := \left(\begin{smallmatrix} -1 & 1 \\ -2 & 1 \end{smallmatrix}\right)$ generate the abelianization of $H/\langle S^2\rangle$. In [Franc and Mason 2016b] χ takes the value $\epsilon = \pm 1$ on V, however the condition $\det \rho(T) = -1$ demands that $\epsilon = 1$. Furthermore, the irreducibility of ρ implies that $n \neq 1$ or 3.

Proposition 12. *The following hold*:

(a) ρ *is a* ***congruence representation***, *i.e.*, $\ker \rho$ *is a congruence subgroup, and all elements in* $\mathfrak{ch}_V$ *are modular functions of weight* 0 *and level* $2n$.

(b) $n \mid 24$ *and* $n \neq 1, 3$.

Proof. By construction, $\mathfrak{ch}_V$ is spanned by functions having q-expansions with *integral Fourier coefficients*. Now the proposition is essentially a restatement of Theorem 21 of [Franc and Mason 2016b]

The only assertion not explicitly stated in [Franc and Mason 2016b] is the statement that the level is $2n$. This amounts to showing that $\rho(T)$ has order $2n$, and this is follows from a knowledge of the eigenvalues of $\rho(T)$, which are as follows [Franc and Mason 2016b, Proposition 2]:

$$\{\lambda, \pm\sigma\}, \qquad \text{where } \sigma^2 = \bar{\lambda}. \qquad \square$$

From Proposition 12 together with the above equation, there is an *even divisor* n of 24 and an integer k coprime to n such that the eigenvalues of $\rho(T)$ are $\{e^{2\pi ik/n}, e^{-2\pi ik/2n}, e^{2\pi i(n-k)/2n}\}$. The three *exponents*

occurring here are equal $\pmod{\mathbb{Z}}$, and in some order, to the exponents $\{r_0, r_1, r_2\}$ occurring in (4). These in turn are equal $\pmod{\mathbb{Z}}$, and in the same order, to $\left\{-\frac{c}{24}, h_1-\frac{c}{24}, h_2-\frac{c}{24}\right\}$.

It follows that $\{h_1, h_2\}$ is congruent $\pmod{\mathbb{Z}}$ to one of $\left\{-\frac{3k}{2n}, \frac{n-3k}{2n}\right\}$, $\left\{\frac{3k}{2n}, \frac{1}{2}\right\}$ or $\left\{\frac{3k+n}{2n}, \frac{1}{2}\right\}$. Because n is an even divisor of 24, all of the rational numbers involved here have denominators equal to 2, 4, 8 or 16 and in fact we obtain the following more precise result:

Proposition 13. *One of the following holds*:

(a) *One of h_1 or h_2 is an element of $\frac{1}{2}+\mathbb{Z}$, and the other has denominator equal to* 4, 8 *or* 16;

(b) *The denominators of h_1 and h_2 are equal, and both are equal to* 4, 8 *or* 16.

Furthermore, we always have $2c \in \mathbb{Z}$, and in particular the conclusions of Corollary 5 apply.

Proof. The assertion regarding the central charge c follows from (a) and (b) together with Lemma 6(a). The lemma follows. □

3.3. ***The primitive case.*** The purpose of this section is to establish results that parallel those of Section 3.2 but now in the case that ρ is *primitive*. This means that if $N \trianglelefteq \Gamma$ then either $U|_N$ is irreducible, or else N is a central subgroup of order dividing 6 (see Lemma 11). We assume that this holds throughout this subsection.

In the imprimitive case we were able to rely on the results of [Franc and Mason 2016b] to restrict the possibilities for ρ to a manageable list. For the case that now presents itself, we will prove:

Proposition 14. *Suppose that ρ is* ***primitive****. Then*

$$\rho(\Gamma) \cong L_2(p) \times \mathbb{Z}/r\mathbb{Z}, \qquad (p = 5 \text{ or } 7,\ r = 2 \text{ or } 6). \tag{8}$$

In all cases ρ is a congruence representation of level pr.

Proof. Let $Z := Z(\rho(\Gamma))$ and note that Z is cyclic of order dividing 6. This holds because $U|_Z$ is necessarily reducible. In particular $\rho(\Gamma) \neq Z$, so we may choose a *minimal nontrivial normal subgroup* $M/Z \trianglelefteq \rho(\Gamma)/Z$.

Case 1: M is solvable. We will show that this case cannot occur. Otherwise, $M/Z \cong (\mathbb{Z}/\ell\mathbb{Z})^d$ for some prime ℓ and integer d. Now $U|_M$ is irreducible, and this forces $\ell = 3$, moreover the Sylow 3-subgroup of M, call it P, satisfies $P \trianglelefteq \rho(\Gamma)$. Indeed, $d = 2$ and P is an extra-special group $P \cong 3^{1+2}$. Because P acts irreducibly on U its centralizer consists of scalar matrices which therefore lie in Z. As a result, it follows that $\rho(\Gamma)/Z$ is isomorphic to a subgroup of the group of automorphisms of P that acts trivially on $Z(P)$. This latter group is $(\mathbb{Z}/3\mathbb{Z})^2 \rtimes \mathrm{SL}_2(3)$. Because $\rho(\Gamma)$ has a subgroup of index 2 (Corollary 9) the only possibilities are that $\rho(\Gamma)/PZ$ is isomorphic to subgroup of $\mathbb{Z}/12\mathbb{Z}$, where we use the fact that the abelianization of Γ is *cyclic* to eliminate some possibilities. Indeed, this abelianization is $\mathbb{Z}/12\mathbb{Z}$, generated by the image of T, and furthermore $T^6\Gamma' = S^2\Gamma'$. It follows that in fact $\rho(\Gamma)/PZ$ is isomorphic to subgroup of $\mathbb{Z}/6\mathbb{Z}$. But in all such cases, $M = PZ$ is *not* a minimal normal subgroup. This completes the proof in Case 1.

Case 2: M is nonsolvable. Here, the only quasisimple groups with a 3-dimensional faithful projective representation are $L_2(5)$, $L_2(7)$, $3.L_2(9)$, and the latter group is excluded thanks to Lemma 10(b). We deduce that $M \cong L_2(p) \times Z$ with $p = 5$ or 7. Furthermore $\mathrm{Aut}(L_2(p)) = \mathrm{PGL}_2(p)$ does *not* have a 3-dimensional faithful representation, so $\Gamma = M$. Let $Z \cong Z/r\mathbb{Z}$ with $r \mid 6$. Because $\rho(\Gamma)$ has a subgroup of index 2, then $2 \mid r$, so that $r = 2$ or 6.

Finally, use Lemma 10(a) and the fact that $\Gamma'\langle S^2\rangle$ is a congruence subgroup of level 6 to see that $\ker\rho$ is also a congruence subgroup, of level pr. This completes the proof of the proposition. □

With this result in hand we turn to a description of the possible sets of eigenvalues for $\rho(T)$. Because T generates the abelianization of Γ and the level of $\ker\rho$ is pr, there is a generator z of $Z \cong \mathbb{Z}/r\mathbb{Z}$ and an element $x \in L_2(p)$ of order p such that $\rho(T) = xz$. Noting that $L_2(p)$ has a pair of conjugate irreducible representations of dimension 3, it follows that ρ falls into one of just 12 equivalence classes and similarly there 12 possible sets of eigenvalues for $\rho(T)$. Thus if $p = 5$ then the eigenvalues for $\rho(T)$ are of the form $\{\mu, \mu\lambda, \mu\bar{\lambda}\}$ where λ and μ are primitive 5-th and r-th roots of unity, respectively. Similarly, if $p = 7$ the eigenvalues for $\rho(T)$ are of the form $\{\mu\lambda, \mu\lambda^2, \mu\lambda^4\}$, where λ and μ are primitive 7-th and r-th roots of unity, respectively. Hence the possible exponents (mod $\mathbb{Z}$) are as follows:

$$\begin{aligned}
&(p,r) = (5,2).\ \left\{\tfrac{1}{2}, \tfrac{3}{10}, \tfrac{7}{10}\right\}, \left\{\tfrac{1}{2}, \tfrac{1}{10}, \tfrac{9}{10}\right\}\\
&(p,r) = (5,6).\ \left\{\tfrac{1}{6}, \tfrac{11}{30}, \tfrac{29}{30}\right\}, \left\{\tfrac{1}{6}, \tfrac{17}{30}, \tfrac{23}{30}\right\}, \left\{\tfrac{5}{6}, \tfrac{1}{30}, \tfrac{19}{30}\right\}, \left\{\tfrac{5}{6}, \tfrac{7}{30}, \tfrac{13}{30}\right\}\\
&(p,r) = (7,2).\ \left\{\tfrac{1}{14}, \tfrac{9}{14}, \tfrac{11}{14}\right\}, \left\{\tfrac{3}{14}, \tfrac{5}{14}, \tfrac{13}{14}\right\}\\
&(p,r) = (7,6).\ \left\{\tfrac{13}{42}, \tfrac{19}{42}, \tfrac{31}{42}\right\}, \left\{\tfrac{25}{42}, \tfrac{37}{42}, \tfrac{1}{42}\right\}, \left\{\tfrac{41}{42}, \tfrac{5}{42}, \tfrac{17}{42}\right\}, \left\{\tfrac{11}{42}, \tfrac{23}{42}, \tfrac{29}{42}\right\}.
\end{aligned} \tag{9}$$

Finally we summarize these computations in the following:

Proposition 15. *If ρ is a primitive representation then one of the following holds*:

(1) *If $p = 5$, then the pairs of rational numbers $\{h_1, h_2\}$ (mod $\mathbb{Z}$) take **all** possible values $\left\{\frac{u}{5}, \frac{v}{5}\right\}$ with $1 \le u < v \le 4$.*

(2) *If $p = 7$, then the pairs of rational numbers $\{h_1, h_2\}$ (mod $\mathbb{Z}$) takes each of the 6 values $\left\{\frac{1}{7}, \frac{3}{7}\right\}$, $\left\{\frac{1}{7}, \frac{5}{7}\right\}$, $\left\{\frac{2}{7}, \frac{3}{7}\right\}$, $\left\{\frac{2}{7}, \frac{6}{7}\right\}$, $\left\{\frac{4}{7}, \frac{5}{7}\right\}$, $\left\{\frac{4}{7}, \frac{6}{7}\right\}$ exactly 3 times, and the other 9 values are **omitted**.*

Remark 16. In what follows, the critical points to observe in Propositions 13 and 15 are that the *denominators* of h_1 and h_2 are divisors of 16 in the imprimitive cases, and they are divisors of 5 or 7 in the primitive cases.

4. The elliptic surface

Thanks to the results in Sections 2 and 3, we are now prepared to tackle the arithmetic classification of possible character vectors $F(\tau)$ for strongly regular VOAs V with exactly 3-simple modules and irreducible monic monodromy. It will then remain to analyze which of the possible character vectors are in fact realized by a VOA.

The next step in our classification specializes (7) to yield the following formula for the character vector $F(\tau)$ corresponding to a VOA V with simple modules V, M_1 and M_2: we have $F(\tau)=(f_0, f_1, f_2)^T$, where

$$f_0 = j^{\frac{1}{6}(2x+2y+3)}{}_3F_2\Big(-\frac{2x+2y+3}{6}, -\frac{2x+2y+1}{6}, -\frac{2x+2y-1}{6}; -x, -y; \frac{1728}{j}\Big),$$
$$f_1 = A_1 j^{\frac{1}{6}(2y-4x-3)}{}_3F_2\Big(\frac{4x-2y+3}{6}, \frac{4x-2y+5}{6}, \frac{4x-2y+7}{6}; x+1, x-y; \frac{1728}{j}\Big),$$
$$f_2 = A_2 j^{\frac{1}{6}(2x-4y-3)}{}_3F_2\Big(\frac{4y-2x+3}{6}, \frac{4y-2x+5}{6}, \frac{4y-2x+7}{6}; y+1, y-x; \frac{1728}{j}\Big),$$

and $c = 8(x+y)+12$, $h_1 = x+1$, $h_2 = y+1$, $A_1 = \dim(M_1)_{h_1}$, $A_2 = \dim(M_2)_{h_2}$.

While Section 3 showed that we need only consider certain rational values of x and y whose denominators divide 16, 5 or 7, it is useful to observe that $F(\tau)$ is in fact an algebraic family of vector-valued modular forms varying with the parameters x and y, in the sense that the Fourier coefficients of this family are rational functions in x and y. If $F(\tau)$ corresponds to a VOA, then the coefficients must in fact be nonnegative integers. Since A_1 and A_2 are unknown positive integers, in this section we focus on f_0. More precisely, if we write $f_0(q) = q^{-c/24}(1+mq+O(q^2))$ as in Section 2, then the hypergeometric expression for f_0 above shows that m, x and y satisfy an algebraic equation that defines an elliptic surface:

$$0 = (4(x+y)+6)((4(x+y)+2)(4(x+y)-2)-62xy)+mxy. \tag{10}$$

As a fibration over the m-line, a theorem of Siegel (Theorem 7.3.9 of [Bombieri and Gubler 2006]) tells us that all of the good fibers of this surface have finitely many rational solutions subject to our restrictions on the monodromy from Section 3. It does not appear to be easy to classify all of the relevant rational solutions directly, and so ultimately our analysis will rely on properties of this elliptic surface, in addition to properties of vector-valued modular forms and generalized hypergeometric series. Nevertheless, we shall describe some facts on the geometry and arithmetic of this surface that were crucial in our initial studies on this classification problem, but which will otherwise not be used in the sequel.

Begin by homogenizing (10): we are interested in the curve $E/\mathbb{C}(m)$ defined by $F(x, y, z) = 0$, where

$$F(x, y, z) = (4(x+y)+6z)((4(x+y)+2z)(4(x+y)-2z)-62xy)+mxyz.$$

Notice that E meets the line at infinity defined by $z=0$ in three distinct points:

$$P_1 = (1:-1:0), \quad P_2 = (15+\sqrt{-31}:16:0), \quad P_3 = (15-\sqrt{-31}:16:0).$$

Taking $P_1 := \infty$ for the identity of the group, the inversion for the group law on E is given by swapping x and y. At the level of VOAs this corresponds to interchanging the nontrivial modules M_1 and M_2 for V. The group law of (10) itself has a more complicated expression in terms of m that we will not write down explicitly.

Consider the change of coordinates

$$(U:V:W) = (x:y:z)\begin{pmatrix} -24(65m^2-24552m-353648) & -6912m(m-248)(m-496) & 248 \\ -24(65m^2-24552m-353648) & 6912m(m-248)(m-496) & 248 \\ -3(m^3-732m^2+97712m-4243776) & 0 & 372-m \end{pmatrix}$$

This change of coordinates turns (10) into the Weierstrass form $H(U, V, W) = 0$, where

$$H(U, V, W) = -V^2W + U^3 - 27(m^3 - 844m^2 + 210992m + 1049536)(m + 124)UW^2 + 54(m^6 - 1080m^5 + 353904m^4 - 78209280m^3 + 16393117440m^2 + 465661052928m + 1484665229312)W^3.$$

The discriminant of this elliptic curve over $\mathbb{C}(m)$ is

$$\Delta = 2^{27} \cdot 3^{13} \cdot (m+4)m^2(m-248)^2(m-496)^2\left(m^2 + \tfrac{123}{3}m + \tfrac{8464}{3}\right)$$

and the j-invariant is

$$j = \frac{(m+124)^3(m^3 - 844m^2 + 210992m + 1049536)^3}{2^{15} \cdot 3 \cdot m^2(m-248)^2(m-496)^2(m+4)\left(m^2 + \frac{128}{3}m + \frac{8464}{3}\right)}.$$

Setting $y = 0$ in (10) yields three rational points

$$Q_1 = \left(\tfrac{1}{2} : 0 : 1\right), \quad Q_2 = \left(-\tfrac{1}{2} : 0 : 1\right), \quad Q_3 = \left(-\tfrac{3}{2} : 0 : 1\right),$$

such that $Q_1 + Q_2 + Q_3 = \infty$. One can show that these points have infinite order in the fiber E_m of $E/\mathbb{C}(m)$ for all rational values of m except when $m = -32, -4, 0, \frac{633}{3}, 248$ and 496. Thus, the rational fibers E_m typically have Mordell–Weil rank at least 2. This might sound surprising, as the average Mordell–Weil rank of a rational elliptic curve is expected to be $\frac{1}{2}$. But in fact, families such as (10) with large rank are not so uncommon — see for example [Elkies 2007] for an interesting discussion of such matters.

We began our study of (10) directly via the fibration over $\mathbb{C}(m)$. It turns out that fibering over y is more useful for classifying the VOAs under discussion here: indeed, all but finitely many of the infinite number of VOAs identified in Theorem 1 correspond to $y = -\frac{1}{2}$. Nevertheless, we shall record here a result that allows the effective enumeration of solutions (m, x, y) to (10) for fixed rational m and rational x and y with bounded denominator that was crucial in our initial studies of (10).

The idea is to first study the rational points of the quotient surface obtained by modding out (10) by the inverse for the elliptic curve group law. Since inversion is given by swapping x and y in (10), we are interested in the rational solutions to the equation

$$0 = 4(2u+3)(8u^2 - 2 - 31v) + mv.$$

Solving for x and y via $x + y = u$ and $xy = v$ yields solutions of (10) defined over a quadratic extension of $\mathbb{Q}$. It will be convenient to work with the corresponding projectivized equation

$$0 = 4(2u+3w)(8u^2 - 2w^2 - 31vw) + mvw^2. \tag{11}$$

Equation (11) defines a one-parameter family of singular cubic curves that, generically, are connected (and there are a finite number of fibers equal to a conic times a line). The rational points in the smooth locus of a connected rational singular cubic can be parametrized by linear projection from a rational singularity. The point $P = (0 : 1 : 0)$ is a rational singular point of every fiber, and this is the point that we will project from. The general line meeting P is given by the equation $au + bw = 0$

First suppose that $b = 0$. This means we wish to describe the solutions to (11) with $u = 0$. These are the point P, along with the points

$$\left(0 : \frac{24}{m-372} : 1\right)$$

with $m \neq 372$.

Henceforth we may assume that b and u are nonzero. After reparametrizing our line, we may assume $w = au$. Substituting this into (11) and using $u \neq 0$ yields

$$a(-ma + 372a + 248)v = -8(a-2)(a+2)(3a+2)u.$$

If $a = 0$ then this equation forces $u = 0$, and we have already classified such points. We are thus now free to assume $a \neq 0$. If $-ma + 372a + 248 = 0$ then we must have $a = 2, -2$ or $a = -\frac{2}{3}$. This implies that away from the fibers for $m = 0$, 248 and 496, we may assume $-ma + 372a + 248 \neq 0$. Therefore, away from these values of m we can solve for v above to obtain the family of points

$$\left(u : \frac{8(2u-1)(2u+1)(2u+3)}{(372 - m + 248u)} : 1\right).$$

Notice that if we set $u = 0$ we recover the preceding family of points.

It remains to consider whether the fibers have other rational singularities besides P (as those points can't be accessed via projection), and to consider the fibers above $m = 0$, 248 and 496.

First we treat the singularities. The v-partial derivative of (11) yields

$$w(-wm + 248u + 372w) = 0.$$

Thus, singular solutions in a fiber of (11) must satisfy either $w = 0$ or $u = \frac{1}{248}(m - 372)w$. When $w = 0$ we find, by consideration of the v-partial, that the only possible additional rational singularity is $\left(1 : \frac{12}{31} : 0\right)$. The u-partial does not vanish at this point, and hence this is not in fact a singularity of the fibers. The other case is when $w \neq 0$ and

$$Q = \left(\frac{m - 372}{248} : v : 1\right).$$

Substituting this into (11) yields $m = 0, 248$ or 496. When $m = 0$ we obtain the unique additional singularity $\left(-\frac{3}{2} : \frac{16}{31} : 1\right)$, when $m = 248$ we obtain the unique additional singularity $\left(-\frac{1}{2} : -\frac{8}{31} : 1\right)$, and when $m = 496$ we obtain the unique additional singularity $\left(\frac{1}{2} : \frac{16}{31} : 1\right)$. These are all the missing singularities, and all the missing points on the fibers corresponding to $m = 0, 248$ and 496. Thus, we have described all rational solutions to (11). We have nearly proven the following:

Proposition 17. *Suppose that (m, x, y) is a rational solution to (10). Then if $u = x + y$ and $v = xy$, the rational point (u, v, m) is equal to*

$$\left(u, \frac{8(2u-1)(2u+1)(2u+3)}{(372 - m + 248u)}, m\right)$$

and $u \neq \frac{1}{248}(m - 372)$.

Proof. We have seen that the only other possible rational solutions (m, x, y) correspond to (u, v, m) equal to one of the singular points $\left(-\frac{3}{2}, \frac{16}{31}, 0\right)$, $\left(-\frac{1}{2}, -\frac{8}{31}, 248\right)$ or $\left(\frac{1}{2}, \frac{16}{31}, 496\right)$. But none of these correspond to *rational* values of x and y. □

Theorem 18. *Let $N > 0$ be an integer and let m be a rational number. Then the number of solutions $a_m(N)$ to (10) with rational x, y of denominator dividing N satisfies*

$$a_m(N) \leq 2 + N \max\left(\frac{16|m - 372|}{31}, 6148\right).$$

Proof. Let (m, x, y) be a rational solution to (10), and let (u, v, m) be the corresponding solution to (11) with $u = x + y$, $v = xy$. Then (u, v, m) is equal to one of the points in Proposition 17. Since the polynomial $T^2 - uT + v$ has rational roots by hypothesis, it follows that the discriminant

$$u^2 - 4v = u^2 - \frac{32(2u-1)(2u+1)(2u+3)}{(372 - m + 248u)}$$

must be a rational square. In particular,

$$1 \geq \frac{32}{31} \frac{\left(1 - \frac{1}{(2u)^2}\right)\left(1 + \frac{3}{2u}\right)}{\left(1 + \frac{372-m}{248u}\right)}.$$

As $|u|$ grows, the right-hand side converges to $\frac{32}{31}$, so that in fact, there are only finitely many solutions in each fiber. We knew this already by a result of Siegel, but we can now use the parametrization to obtain precise bounds.

First assume that $|(372 - m)/248u| < 1/A$ for some big constant A that we will specify later. Then for $A > 31$ we find

$$1 > \frac{31(A+1)}{32A} \geq \left(1 - \frac{1}{4u^2}\right)\left(1 + \frac{3}{2u}\right),$$

and this will produce contradictions for large $|u|$. Choose numbers $e_1, e_2 \in (0, 1)$ with $e_1 + e_2 = 1$. We will find explicit bounds on u that ensure

$$(1 - (2u)^{-2}) > (31(A+1)/32A)^{e_1},$$
$$(1 + 3/(2u)) > (31(A+1)/32A)^{e_2}.$$

The first bound is equivalent with

$$1 - \left(\frac{31(A+1)}{32A}\right)^{e_1} > \frac{1}{(2u)^2}$$

which is equivalent with

$$|u| > \frac{1}{2}\left(1 - \left(\frac{31(A+1)}{32A}\right)^{e_1}\right)^{-1/2}.$$

The second bound is equivalent with

$$1-\left(\frac{31(A+1)}{32A}\right)^{e_2} > -\frac{3}{2u}$$

This is always true if $u > 0$ by choice of A and e_2, since the left side is positive, so that the second bound will hold whenever

$$|u| > \frac{3}{2}\left(1-\left(\frac{31(A+1)}{32A}\right)^{e_2}\right)^{-1}$$

Thus, if $|u|$ is bigger than the max of these, we have a contradiction. Therefore, we must have

$$|u| \leq \max\left(\frac{A|m-372|}{248}, \frac{1}{2}\left(1-\left(\frac{31(A+1)}{32A}\right)^{e_1}\right)^{-1/2}, \frac{3}{2}\left(1-\left(\frac{31(A+1)}{32A}\right)^{e_2}\right)^{-1}\right).$$

Now to optimize parameters. First off, our choice of A must ensure that $1 > 31(A+1)/(32A)$, and we'd like it to be as small as possible. A natural choice is $A = 32$, but any A satisfying $31 < A \leq 32$ would work. To be definite take $A = 32$, so that

$$|u| \leq \max\left(\frac{4|m-372|}{31}, \frac{1}{2}\left(1-\left(\frac{1023}{1024}\right)^{e_1}\right)^{-1/2}, \frac{3}{2}\left(1-\left(\frac{1023}{1024}\right)^{e_2}\right)^{-1}\right).$$

Next we would like to optimize the choice of e_1 and e_2 so that this maximum is minimized. Computations show that the minimum of the last two values above is achieved for e_1 somewhere between $1/5000$ and $1/10000$. For example, using $e_1 = 1/5000$ we obtain

$$|u| \leq \max\left(\frac{4|m-372|}{31}, 1537\right).$$

We are only interested in the values of u of the form $u = i/N$ in this range, and there are at most

$$2N \max\left(\frac{4|m-372|}{31}, 1537\right) + 1$$

of these. For each such choice, we have at most two rational solutions (m, x, y) and (m, y, x) to (10). This concludes the proof. □

Remark 19. In the proof above, many values of u correspond to points for which the discriminant

$$u^2 - \frac{32(2u-1)(2u+1)(2u+3)}{(372-m+248t)} \geq 0$$

is not a rational square. In such cases the corresponding pair of points (m, x, y) and (m, y, x) satisfying (10) have x and y values contained in a real quadratic extension of $\mathbb{Q}$. Thus, it seems possible that the linear bound on $a_m(N)$ above could be improved by making stronger use of the discriminant condition.

Remark 20. For fixed values of m, the preceding proof yields an explicit and efficient algorithm for enumerating all rational solutions to (10) satisfying the divisibility conditions of Theorem 18. The steps are as follows:

(1) Fix a rational value of m.

(2) List the finite number of values $u = i/N$ satisfying the inequality

$$|u| \leq \max\left(\frac{4|m-372|}{31}, 1537\right).$$

(3) For each value of u from the previous step, test whether the discriminant

$$D(u,m) = u^2 - \frac{32(2u-1)(2u+1)(2u+3)}{(372-m+248u)}$$

is a rational square.

(4) If $D(u,m)$ is a rational square, then set $x = (u+\sqrt{D})/2$ and $y = (u-\sqrt{D})/2$. This contributes solutions (m,x,y) and (m,y,x) to (10) (note that it's possible to have $x = y$).

We have run this algorithm for $m = 0$ through $m = 20,000$, and one finds that it is most common to have $a_m(16) = 8$ and $a_m(5) = a_m(7) = 0$ in that range. Note that $a_m(16) \geq 8$ for all m due to the existence of the points $\pm Q_1, \pm Q_2, \pm Q_3$ on the elliptic curve over $\mathbb{C}(m)$ defined by (10), as well as the points $\pm Q_4$, where

$$Q_4 = Q_1 - Q_2 = \left(-\frac{m}{16} - 1 : -\frac{m}{16} - \frac{1}{2} : 1\right).$$

Notice that the existence of this family of points shows that the bound on $|u|$ used in the proof of Theorem 18 is essentially optimal, since this family of points corresponds to $u = -\frac{1}{8}m - \frac{3}{2}$.

In general, for each m (10) has many rational solutions that do not correspond to VOAs. To aid us in eliminating many of these solutions we shall next analyze all three coordinates of the corresponding (in general hypothetical) characters corresponding to a solution of (10).

5. Positivity restrictions

Let (m,x,y) denote a solution to (10) that corresponds to a VOA as in Theorem 1, and let $F(\tau)$ be the corresponding character vector. In this section we exploit the fact that the Fourier coefficients of $F(\tau)$ must be nonnegative. Since these coefficients are reducible rational functions, we can gain some traction by studying their divisors, as the sign of the coefficient is constant in the connected components of the complement of the divisor.

Theorem 21. *If (m,x,y) denotes a solution to (10) realized by a VOA satisfying the restrictions of Theorem 1, and if $|x+1| > \frac{5}{2}$ or $|y+1| > \frac{5}{2}$, then exactly one of the following holds*:

(1) $|x-y| \leq 1$.

(2) $-2 \leq y \leq 0$.

(3) $-2 \leq x \leq 0$.

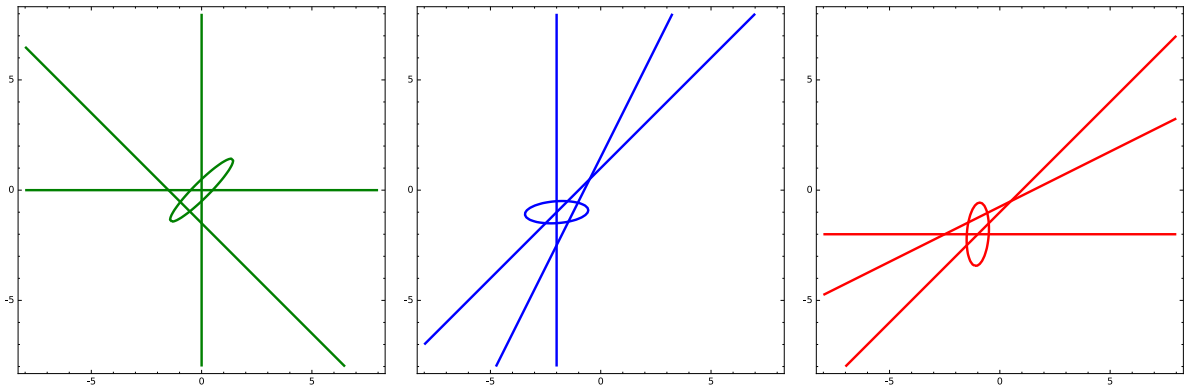


Figure 1. The divisors of m (left), F_1 (center) and F_2 (right).

Proof. Begin by writing

$$\begin{pmatrix} f_0 \\ f_1 \\ f_2 \end{pmatrix} = \operatorname{diag}\left(q^{-\frac{1}{6}(2x+2y+3)}, A_1 q^{-\frac{1}{6}(2y-4x-3)}, A_2 q^{-\frac{1}{6}(2x-4y-3)}\right) \begin{pmatrix} 1+mq+O(q^2) \\ 1+F_1 q+O(q^2) \\ 1+F_2 q+O(q^2) \end{pmatrix}$$

Equation (10) gives an explicit formula for m in terms of x and y. From the expressions for f_1 and f_2 in terms of generalized hypergeometric series, one finds similarly that

$$F_1(x, y) = \frac{4(2y-4x-3)(x^2-xy+8y^2+3x+14y+8)}{(x+2)(y-x-1)}$$

and $F_2(x, y) = F_1(y, x)$. Observe that the divisors of F_1 dissect the plane into a finite number of regions, and the sign of F_1 is constant in each region. Figure 1 shows the divisors of each of m, F_1 and F_2. Figure 2 plots all three divisors. Outside of the boxed region enclosed by the dashed lines, the only regions where m, F_1 and F_2 are simultaneously positive are the shaded regions in Figure 2, and these regions correspond to the statement of the theorem. □

Remark 22. Since $h_1 = x+1$ and $h_2 = y+1$, the following condition holds for a strongly regular VOA with exactly 3 simple modules and whose character vector satisfies a monic MLDE of degree 3 with irreducible monodromy. If $|h_1| > \frac{5}{2}$ or $|h_2| > \frac{5}{2}$, then one of the following holds:

(1) $|h_1 - h_2| \leq 1$,

(2) $|h_1| \leq 1$, or

(3) $|h_2| \leq 1$.

This is a relatively simple consequence of the fact that the Fourier coefficients of $F(\tau)$ are rational functions of h_1 and h_2.

The region bounded by the dashed lines in Figure 2 contains a finite number of points (x, y), where x and y are rational numbers satisfying the restrictions of Section 3 (recall that Section 3 showed that x and y are necessarily rational numbers with denominators that divide 5, 7 or 16). It is thus a simple matter

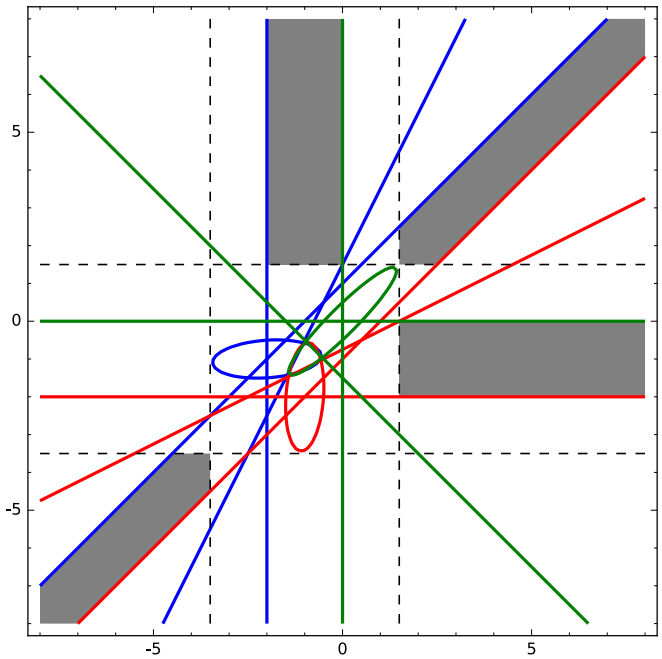


Figure 2. Regions corresponding to VOAs.

to enumerate them. Therefore, by symmetry we may now focus our attention on the shaded regions in Figure 2 on page 1635 below the diagonal $x = y$. The shaded regions contain a finite number of horizontal and diagonal slices of the elliptic surface defined by (10) of relevance to our classification. These slices turn out to be singular cubic curves whose rational points are parametrized and studied in Section 6 below. In the next section we exploit this geometry and the hypergeometric nature of F to find all values of x and y where f_0 has positive integer coefficients, and where f_1 and f_2 have positive coefficients.

Remark 23. Due to the unknown scalars $A_j = \dim(M_j)_{h_j}$ for $j = 1, 2$, we cannot yet make use of the fact that f_1 and f_2 have integer coefficients.

6. The remaining fibers

6.1. ***The horizontal fibers.*** In this section we regard (10) as a fibration over y. In order to homogenize the equation, let m be of degree 1 and let y be of degree 0. Then the homogenized version of (10) is

$$0 = (4(x + yz) + 6z)((4(x + yz) + 2z)(4(x + yz) - 2z) - 62xyz) + mxyz \tag{12}$$

and there is a (unique) singular point $(m : x : z) = (1 : 0 : 0)$ at infinity in every fiber. Therefore, the smooth locus of each fiber can be rationally parametrized by projection from $(1 : 0 : 0)$.

Before proceeding to this we shall classify all additional singular points in the affine patches with $z \neq 0$, as such points cannot be obtained by projection from $(1 : 0 : 0)$. First off, the vanishing of the

m-partial of (12) implies that either $x = 0$ or $z = 0$ at a singularity. The vanishing of the partials at points with $x = 0$ corresponds to the polynomial equations

$$0 = z(56y^2z - my + 180yz + 16z),$$
$$0 = (2y - 1)(2y + 1)(2y + 3)z^2.$$

It follows that if $y \neq \pm\frac{1}{2}$ and $-\frac{3}{2}$, then the only singular point in the fiber is the point $(1 : 0 : 0)$ at infinity. Thus, the entire fiber of (12) can be described by projection from infinity, as long as $y \neq \pm\frac{1}{2}$ and $-\frac{3}{2}$. In the exceptional fibers we find the following additional singular points, corresponding to a conic intersecting a line in two points:

$$\begin{aligned} y = \tfrac{1}{2}: &\quad (240 : 0 : 1), \\ y = -\tfrac{1}{2}: &\quad (120 : 0 : 1), \\ y = -\tfrac{3}{2}: &\quad \left(\tfrac{256}{3} : 0 : 1\right). \end{aligned}$$

Note that $y = \frac{1}{2}$ is outside of the shaded region, so there are in fact only two exceptional fibers that we must consider.

Thus, we now suppose that $-2 \leq y \leq 0$ with $y \neq -\frac{1}{2}, -\frac{3}{2}$, and we will treat these two exceptional fibers separately afterward. In order to rule out the existence of a VOA corresponding to all but (an explicitly computable) finite number of such solutions to (10), we will use the fact that the character of the hypothetical VOA

$$f_0 = j^{\frac{1}{6}(2x+2y+3)}\,{}_3F_2\left(-\frac{2x+2y+3}{6}, -\frac{2x+2y+1}{6}, -\frac{2x+2y-1}{6}; -x, -y; \frac{1728}{j}\right)$$

must have nonnegative integers as coefficients.

Let B_k denote the k-th coefficient of the underlying hypergeometric series (without the j-factors taken into account) defining f_0. If we can show that some hypergeometric coefficient B_k has a prime divisor in its denominator that does not divide the denominators of $x/6$ and $y/6$, then it will also appear in the denominator of the k-th coefficient of f_0. Notice that since we are only interested in solutions (x, y) to (10) with denominators equal to 5, 7 or a divisor of 16, by Section 3, this means that only primes $p \leq 96$ could possibly divide some denominator of a coefficient B_k but not divide any denominators in f_0. Thus, below we restrict to primes $p \geq 96$ and consider only the coefficients B_k, rather than the more complicated coefficients of f_0.

Recall from [Franc et al. 2018, Theorem 3.4] that if $c_p(x, y)$ denotes the number of p-adic carries required to compute the p-adic addition $x + y$, and if v_p denotes the p-adic valuation normalized so that $v_p(p) = 1$, then

$$v_p(B_k) = c_p\left(-\tfrac{1}{3}(x + y) - \tfrac{3}{2}, k\right) + c_p\left(-\tfrac{1}{3}(x + y) - \tfrac{7}{6}, k\right) + c_p\left(-\tfrac{1}{3}(x + y) - \tfrac{5}{6}, k\right) - c_p(-x - 1, k) - c_p(-y - 1, k). \quad (13)$$

$y=-\frac{1}{5}$	$-\frac{1}{3}y-\frac{3}{2}$	$-\frac{1}{3}y-\frac{7}{6}$	$-\frac{1}{3}y-\frac{5}{6}$	$-y-1$	$-\frac{1}{3}$
$p\equiv 1$	$\frac{1}{30}(13p-43)$	$\frac{1}{30}(3p-33)$	$\frac{1}{30}(23p-23)$	$\frac{1}{30}(24p-24)$	$\frac{1}{3}(p-1)$
$p\equiv 7$	$\frac{1}{30}(19p-43)$	$\frac{1}{30}(9p-33)$	$\frac{1}{30}(29p-23)$	$\frac{1}{30}(12p-24)$	$\frac{1}{3}(p-1)$
$p\equiv 11$	$\frac{1}{30}(23p-43)$	$\frac{1}{30}(3p-33)$	$\frac{1}{30}(13p-23)$	$\frac{1}{30}(24p-24)$	$\frac{1}{3}(2p-1)$
$p\equiv 13$	$\frac{1}{30}(p-43)$	$\frac{1}{30}(21p-33)$	$\frac{1}{30}(11p-23)$	$\frac{1}{30}(18p-24)$	$\frac{1}{3}(p-1)$
$p\equiv 17$	$\frac{1}{30}(29p-43)$	$\frac{1}{30}(9p-33)$	$\frac{1}{30}(19p-23)$	$\frac{1}{30}(12p-24)$	$\frac{1}{3}(2p-1)$
$p\equiv 19$	$\frac{1}{30}(7p-43)$	$\frac{1}{30}(27p-33)$	$\frac{1}{30}(17p-23)$	$\frac{1}{30}(6p-24)$	$\frac{1}{3}(p-1)$
$p\equiv 23$	$\frac{1}{30}(11p-43)$	$\frac{1}{30}(21p-33)$	$\frac{1}{30}(p-23)$	$\frac{1}{30}(18p-24)$	$\frac{1}{3}(2p-1)$
$p\equiv 29$	$\frac{1}{30}(17p-43)$	$\frac{1}{30}(27p-33)$	$\frac{1}{30}(7p-23)$	$\frac{1}{30}(6p-24)$	$\frac{1}{3}(2p-1)$

Table 2. List of zeroth p-adic digits of hypergeometric parameters.

The key point here is that if there exists a prime $p\geq 96$ such that the zeroth p-adic digit of $-y-1$ is largest among the 5 arguments above, say $-y-1\equiv y_0 \pmod p$, then in (13), $c_p(-y-1,\, p-y_0)\geq 1$, while each other term $c_p(*,\, p-y_0)$ will be zero. Therefore, for such primes we have $v_p(A_{p-y_0})\leq -1$ and hence f_0 is *not* integral. The arithmetic difficulty that arises in our argument for the exceptional cases when $y=-\frac{1}{2},-\frac{3}{2}$ is that $-y-1$ has zeroth p-adic digit asymptotic to $p/2$ for *all* odd primes. Hence we shall treat those cases separately.

Suppose first that the denominators of x and y are both equal to 5. We shall give all the details in this case and omit the details for the cases of the other possible denominators, as the arguments are identical save for adjusted constants. The exceptions are the fibers $y=-\frac{1}{2}$ and $y=-\frac{3}{2}$, which we shall also treat in detail. Note that since we are interested in irreducible monodromy representations, we may assume that $5x$ and $5y$ are both integral and relatively prime to 5, and also $5x\not\equiv 5y \pmod 5$. The key result in this case is the following:

Proposition 24. *Let (m,x,y) be a solution to (10) with $|y+1|<1$, such that $5x$ and $5y$ are integers coprime to 5, and such that $5x\not\equiv 5y \pmod 5$. Then if $x>18188$, the series f_0 does not have integral Fourier coefficients.*

Proof. There is a unique nonzero congruence class $p_0 \pmod{30}$ such that for all primes $p\equiv p_0 \pmod{30}$ big enough (e.g., $p>96$ suffices), the zeroth p-adic digit of $-y-1$ is of the form $(4p+A)/5$ where $4p+A\equiv 0 \pmod 5$, and the zeroth p-adic digit of $-\frac{1}{3}$ is $(p-1)/3$ (this second condition just forces $p_0\equiv 1 \pmod 3$). Note that A depends on y, but there are finitely many choices for y, so it's bounded absolutely. For example, Table 2 lists the zeroth p-adic digits of some relevant quantities when $y=-\frac{1}{5}$.

A similar table exists for each choice of y, and the important feature is that there is always a unique column where $-y-1$ has zeroth digit asymptotic to $4p/5$, and $-\frac{1}{3}$ has zeroth p-adic digit $(p-1)/3$.

When $y = -\frac{1}{5}$ this is the column $p \equiv 1 \pmod{30}$, but in general it is some class mod 30 such that $p \equiv 1 \pmod{3}$.

So far we have ignored the occurrences of x in the formula (13) for $v_p(B_k)$. We incorporate this information next. Taking account of x has the effect of shifting the digits in first three rows of the table above by a uniform amount (the zeroth p-adic digit of $-x/3$) modulo p. The key is to find primes p such that this shift does not make one of the entries in the first three rows larger than the zeroth p-adic digit of $-y-1$. Therefore, given x, it will suffice to prove that there exists a prime $p \equiv p_0 \pmod{30}$ satisfying $p > 96$ and

$$0 < \left[\frac{(p-1)x}{3}\right]_p < \frac{p}{30}, \tag{14}$$

(where $[\alpha]_p$ denotes the least nonnegative residue of an integer α mod p). This is due to the fact that $[(p-1)x/3]_p$ is the zeroth p-adic digit of $-x/3$, which is the amount that we are shifting p-adic digits by.

Observe that if we write $x = x_0/5$ then

$$\left[\frac{(p-1)x}{3}\right]_p = \begin{cases} \left\{\frac{x_0}{15}\right\}p - \frac{x_0}{15} & p \equiv 1 \pmod{30},\ p > \frac{x_0}{[x_0]_{15}}, \\ \left\{\frac{x_0 - 3[x_0]_5}{15}\right\}p - \frac{x_0}{15} & p \equiv 7 \pmod{30},\ p > \frac{x_0}{[x_0 - 3[x_0]_5]_{15}}, \\ \left\{\frac{x_0 - 9[x_0]_5}{15}\right\}p - \frac{x_0}{15} & p \equiv 13 \pmod{30},\ p > \frac{x_0}{[x_0 - 9[x_0]_5]_{15}}, \\ \left\{\frac{x_0 - 12[x_0]_5}{15}\right\}p - \frac{x_0}{15} & p \equiv 19 \pmod{30},\ p > \frac{x_0}{[x_0 - 12[x_0]_5]_{15}}. \end{cases}$$

In each case there is an integer A (in fact $A = 1, 3, 9$ or 12) such that we win if there exists a prime $p \equiv p_0 \pmod{30}$ with

$$\frac{x_0}{[x_0 - A[x_0]_5]_{15}} < p \quad \text{and} \quad \left\{\frac{x_0 - A[x_0]_5}{15}\right\}p - \frac{x_0}{15} < \frac{p}{30}.$$

These two inequalities are equivalent with

$$\frac{x_0}{[x_0 - A[x_0]_5]_{15}} < p < \frac{x_0}{[x_0 - A[x_0]_5]_{15} - \frac{1}{2}}.$$

If we set $X = x_0/([x_0 - A[x_0]_5]_{15})$ then this is equivalent with

$$X < p < \left(\frac{[x_0 - A[x_0]_5]_{15}}{[x_0 - A[x_0]_5]_{15} - \frac{1}{2}}\right)X$$

In all cases, the complicated scalar factor in the rightmost inequality above is minimized as $\frac{28}{27}$. Therefore, if we can show that for $X > N$ for an explicit N, there is always a prime $p \equiv p_0 \pmod{30}$ that satisfies $X < p < \frac{28}{27}X$, then we will be done by the discussion following (13).

It is a standard argument from analytic number theory that such generalizations of Bertrand's postulate (incorporating more general scalar factors, and restricting to congruence classes of primes) can be proven if one has a sufficiently good understanding of zeros of Dirichlet L-functions. For an explicit discussion

involving effective results, see Appendix A. In particular, Theorem 39 of Appendix A implies that f_0 will not be integral as long as $X > 6496$. Therefore, f_0 is not integral if $x_0 > 14 \cdot 6496$. Since $x = x_0/5$, the proposition follows. □

Proposition 24 allows the classification of all solutions to (10) with $|y+1| < 1$ and y of the form $y = y_0/5$ such that the corresponding function f_0 has positive integral Fourier coefficients, and such that the first two Fourier coefficients of f_1 and f_2 are nonnegative. We computed the first thousand Fourier coefficients of f_0, f_1 and f_2 for all solutions to (10) as in Proposition 24, but with $x \leq 18188$, and tabulated which have the property that

(1) the first thousand coefficients of f_0 are nonnegative integers;

(2) the first thousand coefficients of f_1 and f_2 are nonnegative.

Using only the first thousand coefficients already cut the number of possibilities for f_0 down dramatically. The results of this computation are in Table 4.

A similar argument works for all other y-fibers with $|y+1| < 1$ of interest to us, save for those with $y = -\frac{1}{2}$ and $y = -\frac{3}{2}$. As mentioned above, the issue in these two cases is that the p-adic expansion of $-y-1$ has a zeroth coefficient asymptotic to $p/2$, so it is harder to use the technique described above to find primes such that its zeroth digit is the largest among the five hypergeometric parameters appearing in (13). Thus, we treat these two cases next.

Upon specialization to these two values of y, (10) factors as

$$\begin{aligned} y = -\tfrac{1}{2}: &\quad x(128x^2 + 248x - m + 120) = 0, \\ y = -\tfrac{3}{2}: &\quad x(128x^2 + 360x - 3m + 256) = 0. \end{aligned}$$

Therefore, among the horizontal fibers, it remains to consider solutions (m, x, y) to (10) of the form

$$\left(m, 0, -\tfrac{1}{2}\right), \quad \left(m, 0, -\tfrac{3}{2}\right), \quad \left(\tfrac{1}{512}(n^2-64), \tfrac{1}{256}(-248 \pm n), -\tfrac{1}{2}\right), \quad \left(\tfrac{1}{1536}(n^2+1472), \tfrac{1}{256}(-360 \pm n), -\tfrac{3}{2}\right)$$

The first two sections of (10) with $x = 0$ correspond to reducible monodromy representations, since x is an integer, and so we can ignore them for the present classification of VOAs with associated monodromy representation that is irreducible. Thus, since it remains to consider solutions to (10) in the horizontal region with $x > \frac{3}{2}$, the other points having already been tabulated, it remains in this region to consider the two families of solutions:

$$\left(\tfrac{1}{512}(n^2 - 64), \tfrac{1}{256}(n - 248), -\tfrac{1}{2}\right) \quad n > 632, \qquad \left(\tfrac{1}{1536}(n^2 + 1472), \tfrac{1}{256}(n - 360), -\tfrac{3}{2}\right) \quad n > 744.$$

Any points above corresponding to a finite monodromy representation as classified in Section 3 will necessarily correspond to imprimitive representations. In order to be irreducible, the x values cannot be in $\frac{1}{2}\mathbb{Z}$, and thus by Section 3 they must have denominator equal to 4, 8 or 16 when expressed in lowest terms. Hence in the first case we are only interested in values of n such that $\frac{1}{256}(n - 248) = \frac{1}{16}\alpha$ for an integer $\alpha \not\equiv 0 \pmod 8$, while in the second we are only interested in values of n such that $\frac{1}{256}(n - 360) = \frac{1}{16}\beta$

$y=-\frac{3}{2}$	$p\equiv 1 \pmod 3$	$p\equiv 2 \pmod 3$
$-1-\frac{1}{48}\beta$	$\frac{1}{2}(p-1)+\frac{1}{2}(p-1)p$	$\frac{1}{2}(p-1)+\frac{1}{2}(p-1)p$
$-\frac{2}{3}-\frac{1}{48}\beta$	$\frac{1}{6}(p-1)+\frac{1}{6}(p-1)p$	$\frac{1}{6}(5p-1)+\frac{1}{6}(p-5)p$
$-\frac{1}{3}-\frac{1}{48}\beta$	$\frac{1}{6}(5p+1)+\frac{1}{6}(5p-5)p$	$\frac{1}{6}(p+1)+\frac{1}{6}(5p-1)p$
$-1-\frac{3}{48}\beta$	$\frac{1}{2}(p+1)+\frac{1}{2}(p-1)p$	$\frac{1}{2}(p+1)+\frac{1}{2}(p-1)p$
$\frac{1}{2}$	$\frac{1}{2}(p+1)+\frac{1}{2}(p-1)p$	$\frac{1}{2}(p+1)+\frac{1}{2}(p-1)p$

Table 3. List of zeroth p-adic digits of hypergeometric parameters when $y=-\frac{3}{2}$.

for $\beta \not\equiv 0 \pmod 8$. Thus, taking this integrality condition into consideration, we need only consider solutions of the form

$$\left(\tfrac{1}{2}(\alpha+15)(\alpha+16),\ \tfrac{1}{16}\alpha,\ -\tfrac{1}{2}\right), \qquad \left(\tfrac{1}{6}(\beta^2+45\beta+512),\ \tfrac{1}{16}\beta,\ -\tfrac{3}{2}\right),$$

where $\alpha, \beta > 24$ are integers such that $\alpha, \beta \not\equiv 0 \pmod 8$. Notice that $m=\frac{1}{2}(\alpha+15)(\alpha+16)$ is always a positive integer for positive integral values of α. On the other hand, the ratio $\frac{1}{6}(\beta^2+45\beta+512)$ is only a positive integer if additionally $\beta \not\equiv 0 \pmod 3$. We shall show in Section 8 below that the first family of points in terms of α do in fact correspond to known VOAs — all but finitely many of the examples in Theorem 1 correspond to points in this family. In the remainder of this section we show that the family of points defined in terms of β does *not* correspond to any VOAs (save for some small values of β).

Consider now the values $(m,x,y)=\left(\frac{1}{6}(\beta^2+45\beta+512),\ \frac{1}{16}\beta,\ -\frac{3}{2}\right)$, where $\beta>24$ is not divisible by 3 and it is not divisible by 8. In this case we have

$$v_p(B_k)=c_p\left(-\tfrac{1}{48}(\beta+48),k\right)+c_p\left(-\tfrac{1}{48}(\beta+32),k\right)+c_p\left(-\tfrac{1}{48}(\beta+16),k\right)-c_p\left(-\tfrac{1}{16}(\beta+16),k\right)-c_p\left(\tfrac{1}{2},k\right).$$

Let $p>3$ be a prime divisor of $\beta+24$. The parameters above are congruent to the quantities mod p^2 given in Table 3.

Therefore, if $p>3$ is a prime divisor of $\beta+24$ we find that $v_p(B_{(p-1)/2})=-1$. Notice that since β is coprime to 3, $\beta+24$ is likewise coprime to 3. Therefore, $\beta+24$ can only fail to have an odd prime divisor $p>3$ if $\beta+24=2^u$ for some $u\geq 0$. If $u\geq 3$ then this violates that 8 does not divide β. We thus see that thanks to our hypotheses, there is always a prime $p>3$ that divides $\beta+24$.

It now remains to verify that, for such a prime p, the factor of p in the denominator of $B_{(p-1)/2}$ is not canceled upon multiplying the hypergeometric factor by the power $j^{\beta/48}$ and substituting $1728/j$ for the argument of ${}_3F_2$, as in the definition of f_0. This is a straightforward computation using the q-expansions for $1728/j$ and $j^{\beta/48}$, where the latter q-expansion is computed via the binomial theorem. Therefore, this family of points does not contribute any series f_0 with nonnegative integer coefficients for parameters, and hence there is no corresponding VOA for any of these choice of parameters.

In this way one can parametrize all possible rational solutions to (10) in the horizontal region in Figure 2 with $|y+1| \leq 1$ where f_0 has positive integral Fourier coefficients, and the first two coefficients of f_1 and f_2 are positive.

6.2. ***The diagonal fibers.*** It remains finally to treat the diagonal fibers in Figure 2. Thus suppose that $x - y = a$ for some $|a| \leq 1$. In fact, we may suppose that $y = x - a$ for $0 < a < 1$, since the cases where $a = 0, 1$ correspond to reducible monodromy, and we may assume $a > 0$ by making use of the (x, y) symmetry of (10). In this case,

$$v_p(B_k) = c_p\left(-\tfrac{2}{3}x + \tfrac{1}{6}(2a-9), k\right) + c_p\left(-\tfrac{2}{3}x + \tfrac{1}{6}(2a-7), k\right) + c_p\left(-\tfrac{2}{3}x + \tfrac{1}{6}(2a-5), k\right) - c_p(-x-1, k) - c_p(a-x-1, k).$$

By the classification of the possible monodromy representations of Section 3, we need only consider the cases where $a = \frac{1}{5}b$, $\frac{1}{7}c$ or $\frac{1}{16}d$, and then x must also be a rational number with denominator supported at the same prime. These three cases can be treated as we treated the horizontal fibers in the previous subsection, by choosing primes so that the zeroth p-adic coefficient of $-x-1$ is large relative to the other quantities appearing above. It turns out that no new solutions to (10) arise in this diagonal region (outside of the boxed area where $|x+1| \leq \frac{5}{2}$, $|y+1| \leq \frac{5}{2}$ which was treated separately by a finite computation), where f_0 has positive and integral Fourier coefficients. This concludes our discussion of how to describe a list, corresponding to one infinite family and a number of sporadic exceptions, of solutions to (10) that can be used to establish Theorem 1.

In Tables 4 and 5 on pages 1644–1645, we list all possible solutions to (10) such that f_0, f_1 and f_2 satisfy:

(1) The monodromy is irreducible with a congruence subgroup as kernel.

(2) The first thousand Fourier coefficients of f_0, f_1 and f_2 are all nonnegative.

(3) The first thousand Fourier coefficients of f_0 are integers.

We believe that (3) could be easily strengthened to show that f_0 is in fact positive integral in each case, but we have not gone to the trouble of doing so. This is because in all of the cases of interest for this paper, namely those corresponding to VOAs, integrality follows automatically since the Fourier coefficients count dimensions of finite-dimensional vector spaces.

We shall show that most of the entries in Tables 4 and 5 are *not* realized by a strongly regular VOA with exactly 3 nonisomorphic simple modules and irreducible monic monodromy. Presumably some of these sets of parameters are realized by VOAs V with a 3-dimensional space of characters $\mathfrak{ch}_V$ but more than 3 simple modules, and therefore we include the full dataset.

7. Trimming down to Theorem 1

Most of the potential examples that are tabulated in Tables 4 and 5 do not in fact correspond to a VOA satisfying the conditions of Theorem 1. In this section we explain how to trim these lists down to the statement of Theorem 1.

First we shall use the deep fact, which follows by Huang [2008], that $\rho(S)$ must be a *symmetric* matrix, with $\rho(T)$ diagonal. Since $\rho(T)$ has distinct eigenvalues (see Section 3), and the only invertible matrices that commute with a diagonal matrix with distinct eigenvalues are the diagonal matrices, the only remaining freedom in changing the basis is in conjugating by diagonal matrices. Since we wish to keep the coordinate f_0 fixed, this conjugation amounts to rescaling f_1 and f_2. Said differently, there is *at most one choice of scalars* A_1 and A_2 appearing in the definition of f_1 and f_2 such that $\rho(S)$ is symmetric. Since A_1 and A_2 must themselves be integers, we performed a numerical computation in all of the finitely many remaining cases $\left(\text{with } y \neq -\frac{1}{2}\right)$ to symmetrize $\rho(S)$ and compute exact values for A_1 and A_2. The idea was to numerically evaluate $F(\tau)$ and $F(-1/\tau)$ at random points τ near i, and by comparing the results we obtained a numerical expression for $\rho(S)$ to high enough precision to determine when the values of A_1 and A_2 were nonnegative integers. Note that the hypergeometric expression for $F(\tau)$ is very well-suited to this type of high precision computation. Also, since we used finite precision computation, we could only rule out exactly when A_1 and A_2 are *not* integers. The exact values for A_1 and A_2 that we report here are then justified since in each case we produce examples of VOAs realizing them.

After computing values for A_1 and A_2, we were then able to test the integrality of the first thousand coefficients of all three coordinates of $F(\tau)$, whereas previously we had only been able to make use of the integrality of the first coordinate f_0. This cut our list of possible character vectors down very dramatically. We then checked the remaining cases to verify that the Verlinde formula holds for $\rho(S)$. After all of this work, we found the following exhaustive list of sets of data that could possibly correspond to a VOA as in Theorem 1:

(1) examples corresponding to solutions of (10) with $y = -\frac{1}{2}$;

(2) $(m, c, h_1, h_2) = \left(0, -\frac{68}{7}, -\frac{2}{7}, -\frac{3}{7}\right)$ which is realized by $\mathrm{Vir}(c_{2,7})$;

(3) $(m, c, h_1, h_2) = \left(24, 4, \frac{2}{5}, \frac{3}{5}\right)$ which is realized by $A_{4,1}$;

(4) 11 exceptional cases with $y = \frac{1}{2}$, equivalently, $h_1 = \frac{3}{2}$.

Definition 25. The 11 exceptional examples with $h_1 = \frac{3}{2}$ comprise the *U-series*.

We shall discuss the U-series in greater detail in Section 9.

Table 6 on page 1646 lists data for the U-series, and Table 7 on page 1647 lists the first several Fourier coefficients of f_0, f_1 and f_2 for the examples in the U-series. For convenience we recall here the formulas for the character vector $F(\tau) = (f_0, f_1, f_2)^T$ in terms of the parameters $h_1 = x + 1$, $h_2 = y + 1$ and

$c = 8(h_1 + h_2) - 4$:

$$f_0 = j^{\frac{1}{24}c}\,{}_3F_2\Big(-\frac{c}{24}, \frac{8-c}{24}, \frac{16-c}{24}; 1-h_1, 1-h_2; \frac{1728}{j}\Big),$$

$$f_1 = A_1 j^{\frac{1}{6}(2h_2-4h_1-1)}\,{}_3F_2\Big(\frac{4h_1-2h_2+1}{6}, \frac{4h_1-2h_2+3}{6}, \frac{4h_1-2h_2+5}{6}; h_1, h_1-h_2; \frac{1728}{j}\Big),$$

$$f_2 = A_2 j^{\frac{1}{6}(2h_1-4h_2-1)}\,{}_3F_2\Big(\frac{4h_2-2h_1+1}{6}, \frac{4h_2-2h_1+3}{6}, \frac{4h_2-2h_1+5}{6}; h_2, h_2-h_1; \frac{1728}{j}\Big).$$

Further, for the examples in the U-series $\rho(T) = \exp\big(2\pi i \operatorname{diag}\big(-\frac{1}{24}c, \frac{1}{6}(4h_1-2h_2+1), \frac{1}{6}(4h_2-2h_1+1)\big)\big)$ and

$$\rho(S) = \frac{1}{2}\begin{pmatrix} 1 & 1 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 0 \end{pmatrix}.$$

In Section 8 we shall discuss the existence of VOAs for the infinite family of solutions to (10) with $y = -\frac{1}{2}$, and in Section 9 we provide some more detail about the U-series.

8. Solutions with $y = -\frac{1}{2}$

We turn now to the solutions of (10) with $y = -\frac{1}{2}$. Recall that (10) specializes to

$$x(128x^2 + 248x - m + 120) = 0,$$

and we can ignore the solutions $\big(m, 0, -\frac{1}{2}\big)$ since they correspond to reducible monodromy representations (see Section 3). Therefore we now study the solutions

$$(m, x, y) = \big(\tfrac{1}{2}s(s-1), \tfrac{1}{16}s - 1, -\tfrac{1}{2}\big)$$

where $s > 0$ is an integer that is not divisible by 8. The restriction $s > 0$ arises from the fact that $m = \dim V_1$ must be a nonnegative integer, and the restriction that 8 does not divide s is due to the irreducibility of the monodromy; see Section 3. The main result of this section, whose proof occupies the remainder of the section, classifies exactly what VOAs satisfying the restrictions of Theorem 1 correspond to these examples:

Theorem 26. *Suppose that* $(m, x, y) = \big(\frac{1}{2}s(s-1), \frac{1}{16}s - 1, -\frac{1}{2}\big)$ *for an integer* $s > 0$ *not divisible by* 8. *Then* V *is isomorphic to one of the following*:

$$B_{\ell,1}, \quad A_{1,2}, \quad \operatorname{Vir}(c_{3,4}).$$

Remembering that $c = 8\big(h_1 + h_2 - \frac{1}{2}\big) = 8\big(x + y + \frac{3}{2}\big)$ we have $c = \tilde{c} = \frac{1}{2}s$. In particular we have $c \in \frac{1}{2}\mathbf{Z}$, so that Corollary 5 applies. Our approach to the proof of Theorem 26 is to deal separately with each of the possibilities (a)–(c) of Corollary 5, although the arguments are similar in each case. We try to determine the structure of the Lie algebra V_1, or else prove that there is no choice of V_1 that is compatible with the data. A basic property [Dong and Mason 2004] is that V_1 is *reductive* and its Lie rank is denoted by ℓ (see Section 2.1). As for case (a), we will prove:

Proposition 27. *For (m, x, y) as in Theorem 26, Corollary 5(a)* ***cannot*** *hold.*

Proof. Until further notice we assume that the proposition is false.

Lemma 28. *We have*

$$2\ell^2 + 3\ell + 1 \leq m.$$

m	h_1	h_2	c	$\tilde{c}$
0	$-\frac{1}{5}$	$-\frac{2}{5}$	$-\frac{44}{5}$	$\frac{4}{5}$
0	$\frac{12}{5}$	$\frac{11}{5}$	$\frac{164}{5}$	$\frac{164}{5}$
1	$\frac{1}{5}$	$-\frac{1}{5}$	-4	$\frac{4}{5}$
2	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{4}{5}$
3	$-\frac{2}{5}$	$-\frac{3}{5}$	-12	$\frac{12}{5}$
3	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{12}{5}$	$\frac{12}{5}$
10	$\frac{1}{5}$	$-\frac{2}{5}$	$-\frac{28}{5}$	4
24	$\frac{3}{5}$	$\frac{2}{5}$	4	4
27	$\frac{9}{5}$	$\frac{7}{5}$	$\frac{108}{5}$	$\frac{108}{5}$
28	$\frac{4}{5}$	$\frac{2}{5}$	$\frac{28}{5}$	$\frac{28}{5}$
58	$\frac{9}{5}$	$\frac{8}{5}$	$\frac{116}{5}$	$\frac{116}{5}$
92	$\frac{8}{5}$	$\frac{6}{5}$	$\frac{92}{5}$	$\frac{92}{5}$
104	$\frac{6}{5}$	$\frac{3}{5}$	$\frac{52}{5}$	$\frac{52}{5}$
105	$\frac{4}{5}$	$-\frac{3}{5}$	$-\frac{12}{5}$	12
120	$\frac{8}{5}$	$\frac{7}{5}$	20	20
136	$\frac{7}{5}$	$\frac{4}{5}$	$\frac{68}{5}$	$\frac{68}{5}$
144	$\frac{4}{5}$	$\frac{3}{5}$	$\frac{36}{5}$	$\frac{36}{5}$

m	h_1	h_2	c	$\tilde{c}$
156	$\frac{6}{5}$	$-\frac{2}{5}$	$\frac{12}{5}$	12
220	$\frac{6}{5}$	$\frac{2}{5}$	$\frac{44}{5}$	$\frac{44}{5}$
222	$\frac{7}{5}$	$\frac{3}{5}$	12	12
253	$\frac{7}{5}$	$\frac{1}{5}$	$\frac{44}{5}$	$\frac{44}{5}$
312	$\frac{11}{5}$	$-\frac{2}{5}$	$\frac{52}{5}$	20
336	$\frac{7}{5}$	$\frac{6}{5}$	$\frac{84}{5}$	$\frac{84}{5}$
374	$\frac{9}{5}$	$\frac{2}{5}$	$\frac{68}{5}$	$\frac{68}{5}$
380	$\frac{8}{5}$	$\frac{4}{5}$	$\frac{76}{5}$	$\frac{76}{5}$
437	$\frac{9}{5}$	$\frac{3}{5}$	$\frac{76}{5}$	$\frac{76}{5}$
534	$\frac{12}{5}$	$\frac{1}{5}$	$\frac{84}{5}$	$\frac{84}{5}$
690	$\frac{11}{5}$	$\frac{3}{5}$	$\frac{92}{5}$	$\frac{92}{5}$
860	$\frac{14}{5}$	$\frac{2}{5}$	$\frac{108}{5}$	$\frac{108}{5}$
1404	$\frac{12}{5}$	$\frac{4}{5}$	$\frac{108}{5}$	$\frac{108}{5}$
1536	$\frac{16}{5}$	$\frac{3}{5}$	$\frac{132}{5}$	$\frac{132}{5}$
1711	$\frac{13}{5}$	$\frac{4}{5}$	$\frac{116}{5}$	$\frac{116}{5}$
3612	$\frac{18}{5}$	$\frac{4}{5}$	$\frac{156}{5}$	$\frac{156}{5}$
13110	$\frac{33}{5}$	$\frac{4}{5}$	$\frac{276}{5}$	$\frac{276}{5}$

m	h_1	h_2	c	$\tilde{c}$
0	$-\frac{2}{7}$	$-\frac{3}{7}$	$-\frac{68}{7}$	$\frac{4}{7}$
0	$\frac{17}{7}$	$\frac{16}{7}$	$\frac{236}{7}$	$\frac{236}{7}$
1	$\frac{2}{7}$	$-\frac{1}{7}$	$-\frac{20}{7}$	$\frac{4}{7}$
1	$\frac{3}{7}$	$\frac{1}{7}$	$\frac{4}{7}$	$\frac{4}{7}$
6	$\frac{3}{7}$	$\frac{2}{7}$	$\frac{12}{7}$	$\frac{12}{7}$
41	$\frac{13}{7}$	$\frac{11}{7}$	$\frac{164}{7}$	$\frac{164}{7}$
78	$\frac{12}{7}$	$\frac{11}{7}$	$\frac{156}{7}$	$\frac{156}{7}$
88	$\frac{5}{7}$	$\frac{4}{7}$	$\frac{44}{7}$	$\frac{44}{7}$
156	$\frac{6}{7}$	$\frac{4}{7}$	$\frac{52}{7}$	$\frac{52}{7}$
210	$\frac{8}{7}$	$\frac{3}{7}$	$\frac{60}{7}$	$\frac{60}{7}$
221	$\frac{9}{7}$	$\frac{3}{7}$	$\frac{68}{7}$	$\frac{68}{7}$
248	$\frac{10}{7}$	$\frac{9}{7}$	$\frac{124}{7}$	$\frac{124}{7}$
325	$\frac{11}{7}$	$\frac{5}{7}$	$\frac{100}{7}$	$\frac{100}{7}$
348	$\frac{10}{7}$	$\frac{8}{7}$	$\frac{116}{7}$	$\frac{116}{7}$
378	$\frac{11}{7}$	$\frac{6}{7}$	$\frac{108}{7}$	$\frac{108}{7}$
380	$\frac{12}{7}$	$\frac{4}{7}$	$\frac{100}{7}$	$\frac{100}{7}$
456	$\frac{13}{7}$	$\frac{4}{7}$	$\frac{108}{7}$	$\frac{108}{7}$
1248	$\frac{18}{7}$	$\frac{5}{7}$	$\frac{156}{7}$	$\frac{156}{7}$

Table 4. Full dataset of parameters with f_0 nonnegative integral and f_1 and f_2 nonnegative, where parameters have denominators 5 and 7. Since we only used one-thousand Fourier coefficients to generate this data, some of these series could in fact fail to be integral, but certainly the integral list is a subset of ours.

m	h_1	h_2	c	$\tilde{c}$
0	$\frac{31}{16}$	$\frac{3}{2}$	$\frac{47}{2}$	$\frac{47}{2}$
1	$-\frac{7}{16}$	$-\frac{1}{2}$	$-\frac{23}{2}$	$\frac{1}{2}$
1	$-\frac{3}{8}$	$-\frac{1}{2}$	-11	1
1	$\frac{7}{16}$	$-\frac{1}{16}$	-1	$\frac{1}{2}$
2	$-\frac{1}{4}$	$-\frac{1}{2}$	-10	2
2	$\frac{3}{8}$	$-\frac{1}{8}$	-2	1
3	$\frac{5}{16}$	$-\frac{3}{16}$	-3	$\frac{3}{2}$
4	$\frac{1}{4}$	$-\frac{1}{4}$	-4	2
5	$\frac{3}{16}$	$-\frac{5}{16}$	-5	$\frac{5}{2}$
6	$\frac{1}{8}$	$-\frac{3}{8}$	-6	3
7	$\frac{1}{16}$	$-\frac{7}{16}$	-7	$\frac{7}{2}$
9	$-\frac{1}{16}$	$-\frac{9}{16}$	-9	$\frac{9}{2}$
10	$-\frac{1}{8}$	$-\frac{5}{8}$	-10	5
11	$-\frac{3}{16}$	$-\frac{11}{16}$	-11	$\frac{11}{2}$
12	$-\frac{1}{4}$	$-\frac{3}{4}$	-12	6
13	$-\frac{5}{16}$	$-\frac{13}{16}$	-13	$\frac{13}{2}$
14	$-\frac{3}{8}$	$-\frac{7}{8}$	-14	7
15	$-\frac{7}{16}$	$-\frac{15}{16}$	-15	$\frac{15}{2}$
17	$-\frac{9}{16}$	$-\frac{17}{16}$	-17	$\frac{17}{2}$
18	$-\frac{5}{8}$	$-\frac{9}{8}$	-18	9
19	$-\frac{11}{16}$	$-\frac{19}{16}$	-19	$\frac{19}{2}$
20	$-\frac{3}{4}$	$-\frac{5}{4}$	-20	10
21	$-\frac{13}{16}$	$-\frac{21}{16}$	-21	$\frac{21}{2}$
22	$-\frac{7}{8}$	$-\frac{11}{8}$	-22	11
23	$-\frac{15}{16}$	$-\frac{23}{16}$	-23	$\frac{23}{2}$
23	$\frac{15}{8}$	$\frac{3}{2}$	23	23
25	$-\frac{17}{16}$	$-\frac{25}{16}$	-25	$\frac{25}{2}$
26	$-\frac{9}{8}$	$-\frac{13}{8}$	-26	13
27	$-\frac{19}{16}$	$-\frac{27}{16}$	-27	$\frac{27}{2}$
28	$-\frac{5}{4}$	$-\frac{7}{4}$	-28	14

m	h_1	h_2	c	$\tilde{c}$
29	$-\frac{21}{16}$	$-\frac{29}{16}$	-29	$\frac{29}{2}$
30	$-\frac{11}{8}$	$-\frac{15}{8}$	-30	15
31	$-\frac{23}{16}$	$-\frac{31}{16}$	-31	$\frac{31}{2}$
33	$-\frac{25}{16}$	$-\frac{33}{16}$	-33	$\frac{33}{2}$
34	$-\frac{13}{8}$	$-\frac{17}{8}$	-34	17
35	$-\frac{27}{16}$	$-\frac{35}{16}$	-35	$\frac{35}{2}$
36	$-\frac{7}{4}$	$-\frac{9}{4}$	-36	18
37	$-\frac{29}{16}$	$-\frac{37}{16}$	-37	$\frac{37}{2}$
38	$-\frac{15}{8}$	$-\frac{19}{8}$	-38	19
39	$-\frac{31}{16}$	$-\frac{39}{16}$	-39	$\frac{39}{2}$
45	$\frac{29}{16}$	$\frac{3}{2}$	$\frac{45}{2}$	$\frac{45}{2}$
66	$\frac{7}{4}$	$\frac{3}{2}$	22	22
86	$\frac{27}{16}$	$\frac{3}{2}$	$\frac{43}{2}$	$\frac{43}{2}$
105	$\frac{13}{8}$	$\frac{3}{2}$	21	21
123	$\frac{25}{16}$	$\frac{3}{2}$	$\frac{41}{2}$	$\frac{41}{2}$
156	$\frac{3}{2}$	$\frac{23}{16}$	$\frac{39}{2}$	$\frac{39}{2}$
171	$\frac{3}{2}$	$-\frac{7}{16}$	$\frac{9}{2}$	15
171	$\frac{3}{2}$	$\frac{11}{8}$	19	19
185	$\frac{3}{2}$	$-\frac{3}{8}$	5	14
185	$\frac{3}{2}$	$\frac{21}{16}$	$\frac{37}{2}$	$\frac{37}{2}$
198	$\frac{3}{2}$	$-\frac{5}{16}$	$\frac{11}{2}$	13
198	$\frac{3}{2}$	$\frac{5}{4}$	18	18
210	$\frac{3}{2}$	$-\frac{1}{4}$	6	12
210	$\frac{3}{2}$	$\frac{19}{16}$	$\frac{35}{2}$	$\frac{35}{2}$
221	$\frac{3}{2}$	$-\frac{3}{16}$	$\frac{13}{2}$	11
221	$\frac{3}{2}$	$\frac{9}{8}$	17	17
231	$\frac{3}{2}$	$-\frac{1}{8}$	7	10
231	$\frac{3}{2}$	$\frac{17}{16}$	$\frac{33}{2}$	$\frac{33}{2}$
240	$\frac{3}{2}$	$-\frac{1}{16}$	$\frac{15}{2}$	9
248	$\frac{3}{2}$	$\frac{15}{16}$	$\frac{31}{2}$	$\frac{31}{2}$

m	h_1	h_2	c	$\tilde{c}$
255	$\frac{3}{2}$	$\frac{1}{16}$	$\frac{17}{2}$	$\frac{17}{2}$
255	$\frac{3}{2}$	$\frac{7}{8}$	15	15
261	$\frac{3}{2}$	$\frac{1}{8}$	9	9
261	$\frac{3}{2}$	$\frac{13}{16}$	$\frac{29}{2}$	$\frac{29}{2}$
266	$\frac{3}{2}$	$\frac{3}{16}$	$\frac{19}{2}$	$\frac{19}{2}$
266	$\frac{3}{2}$	$\frac{3}{4}$	14	14
270	$\frac{3}{2}$	$\frac{1}{4}$	10	10
270	$\frac{3}{2}$	$\frac{11}{16}$	$\frac{27}{2}$	$\frac{27}{2}$
273	$\frac{3}{2}$	$\frac{5}{16}$	$\frac{21}{2}$	$\frac{21}{2}$
273	$\frac{3}{2}$	$\frac{5}{8}$	13	13
275	$\frac{3}{2}$	$\frac{3}{8}$	11	11
275	$\frac{3}{2}$	$\frac{9}{16}$	$\frac{25}{2}$	$\frac{25}{2}$
276	$\frac{3}{2}$	$\frac{7}{16}$	$\frac{23}{2}$	$\frac{23}{2}$
496	$\frac{45}{16}$	$-\frac{5}{16}$	16	$\frac{47}{2}$
496	$\frac{23}{8}$	$-\frac{3}{8}$	16	25
496	$\frac{47}{16}$	$-\frac{7}{16}$	16	$\frac{53}{2}$
496	$\frac{49}{16}$	$-\frac{9}{16}$	16	$\frac{59}{2}$
496	$\frac{25}{8}$	$-\frac{5}{8}$	16	31
496	$\frac{51}{16}$	$-\frac{11}{16}$	16	$\frac{65}{2}$
598	$\frac{5}{2}$	$\frac{1}{4}$	18	18
1118	$\frac{7}{2}$	$\frac{1}{4}$	26	26
1194	$\frac{9}{2}$	$-\frac{1}{4}$	30	36
1298	$\frac{5}{2}$	$\frac{3}{4}$	22	22
1640	$\frac{5}{2}$	$\frac{13}{16}$	$\frac{45}{2}$	$\frac{45}{2}$
2323	$\frac{5}{2}$	$\frac{7}{8}$	23	23
2778	$\frac{7}{2}$	$\frac{3}{4}$	30	30
3599	$\frac{7}{2}$	$\frac{13}{16}$	$\frac{61}{2}$	$\frac{61}{2}$
4371	$\frac{5}{2}$	$\frac{15}{16}$	$\frac{47}{2}$	$\frac{47}{2}$
5239	$\frac{7}{2}$	$\frac{7}{8}$	31	31

Table 5. Full dataset for imprimitive representations. This does not include the one infinite family that we shall treat separately in Section 8. Also, the same remark on integrality as in Table 4 applies here.

m	h_1	h_2	c	$\tilde{c}$	V_1	A_1	A_2
0	$\frac{3}{2}$	$\frac{31}{16}$	$\frac{47}{2}$	$\frac{47}{2}$	0	4371	96256
45	$\frac{3}{2}$	$\frac{29}{16}$	$\frac{45}{2}$	$\frac{45}{2}$	D_5 *	4785	46080
86	$\frac{3}{2}$	$\frac{27}{16}$	$\frac{43}{2}$	$\frac{43}{2}$	$2B_4 \oplus G_2$ *	5031	22016
123	$\frac{3}{2}$	$\frac{25}{16}$	$\frac{41}{2}$	$\frac{41}{2}$	$A_1 \oplus A_{10}$ *	5125	10496
156	$\frac{3}{2}$	$\frac{23}{16}$	$\frac{39}{2}$	$\frac{39}{2}$	$2B_6$ *	5083	4992
185	$\frac{3}{2}$	$\frac{21}{16}$	$\frac{37}{2}$	$\frac{37}{2}$	$E_7 \oplus F_4$ *	4921	2368
210	$\frac{3}{2}$	$\frac{19}{16}$	$\frac{35}{2}$	$\frac{35}{2}$	B_{10} *	4655	1120
231	$\frac{3}{2}$	$\frac{17}{16}$	$\frac{33}{2}$	$\frac{33}{2}$	D_{11} *	4301	528
248	$\frac{3}{2}$	$\frac{15}{16}$	$\frac{31}{2}$	$\frac{31}{2}$	E_8 $A_1 \oplus D_{11} \oplus G_2$	3875	248
261	$\frac{3}{2}$	$\frac{13}{16}$	$\frac{29}{2}$	$\frac{29}{2}$	$A_2 \oplus B_{11}$ *	3393	116
270	$\frac{3}{2}$	$\frac{11}{16}$	$\frac{27}{2}$	$\frac{27}{2}$	$A_2 \oplus E_8 \oplus G_2$ $\mathbb{C} \oplus B_3 \oplus E_8$ $\mathbb{C} \oplus C_3 \oplus E_8$	2871	54

Table 6. Data for the U-series. The expressions for V_1 are far from unique in general. In the cases where there are at most three possibilities for V_1, we have listed all of them. In all other cases there are several (in some cases thousands) of possibilities and we have only listed one of them, along with an asterisk.

Proof. Because we are assuming that (a) of Corollary 5 holds, then $\tilde{c} \geq \ell + 1$. Therefore,

$$m = \tfrac{1}{2}s(s-1) = \tilde{c}(2\tilde{c}-1) \geq (\ell+1)(2\ell+1) = 2\ell^2 + 3\ell + 1. \qquad \square$$

As a reductive Lie algebra, V_1 has a direct sum decomposition

$$V_1 = \bigoplus_{i \geq 0} \mathfrak{g}_i, \tag{15}$$

where $\mathfrak{g}_0$ is abelian and each $\mathfrak{g}_i$ is a nonabelian simple Lie algebra ($i \geq 1$), say of Lie rank ℓ_i. Let $\ell_0 := \dim \mathfrak{g}_0$. Then the total Lie rank of V_1 is $\ell = \sum_{i \geq 0} \ell_i$.

The table of dimensions for simple Lie algebras compared with Lie rank is given in Table 8.

Now suppose that V_1 *only* has components $\mathfrak{g}_i$ that are classical (type $A_\ell, \ldots, D_\ell$) or of type G_2 or E_6. By Table 8 each of these satisfies $\dim \mathfrak{g}_i \leq 2\ell_i^2 + 3\ell_i$. Using Lemma 28 we have

$$2\ell^2 + 3\ell + 1 \leq \dim V_1 \leq \ell_0 + \sum_{i \geq 1} (2\ell_i^2 + 3\ell_i) \leq 3\ell + 2\sum_{i \geq 1} \ell_i^2$$

h_1	h_2	$a_0, a_1, a_2, \ldots$
$\frac{3}{2}$	$\frac{31}{16}$	1, 0, 96256, 9646891, 366845011, 8223700027, 130416170627, ... 4371, 1143745, 64680601, 1829005611, 33950840617, 470887671187, ... 96256, 10602496, 420831232, 9685952512, 156435924992, 1958810851328, ...
$\frac{3}{2}$	$\frac{29}{16}$	1, 45, 90225, 7671525, 260868780, 5354634636, 78809509455, ... 4785, 977184, 48445515, 1241925725, 21267996075, 275102618220, ... 46080, 5161984, 199388160, 4423680000, 68709350400, 827293870080, ...
$\frac{3}{2}$	$\frac{27}{16}$	1, 86, 82775, 5989341, 182136390, 3421630228, 46706033862, ... 5031, 819279, 35627220, 827820606, 13070793291, 157564970907, ... 22016, 2515456, 94360576, 2013605376, 30017759232, 346922095616, ...
$\frac{3}{2}$	$\frac{25}{16}$	1, 123, 74374, 4586752, 124739876, 2143484264, 27115530974, ... 5125, 673630, 25702490, 541136245, 7872255635, 88368399005, 816197168410, ... 10496, 1227008, 44597504, 913172992, 13037354496, 144348958464, ...
$\frac{3}{2}$	$\frac{23}{16}$	1, 156, 65442, 3442179, 83713890, 1314851889, 15401260043, 145567687044, ... 5083, 542685, 18172323, 346513193, 4640683320, 48464931804, 419554761418, ... 4992, 599168, 21046272, 412414080, 5625756032, 59548105344, 520893998976, ...
$\frac{3}{2}$	$\frac{21}{16}$	1, 185, 56351, 2528691, 54987069, 788715865, 8545883340, 75369712213, ... 4921, 427868, 12578261, 217080369, 2673896760, 25953557278, 210363766807 ... 2368, 292928, 9914816, 185395456, 2410143296, 24333700608, 203337098176 ...
$\frac{3}{2}$	$\frac{19}{16}$	1, 210, 47425, 1816325, 35302155, 461945596, 4624903605, 38016539200, ... 4655, 329707, 8512950, 132853700, 1503485200, 13547531620, 102694766167, ... 1120, 143392, 4661440, 82908000, 1024273600, 9839831680, 78373048544, ...
$\frac{3}{2}$	$\frac{17}{16}$	1, 231, 38940, 1274086, 22116963, 263714253, 2436524530, 18642807645, ... 4301, 247962, 5625708, 79296041, 823487514, 6879624345, 48709339624, ... 528, 70288, 2186448, 36857568, 431399936, 3932664912, 29784812640, ...
$\frac{3}{2}$	$\frac{15}{16}$	1, 248, 31124, 871627, 13496501, 146447007, 1246840863, 8867414995, ... 3875, 181753, 3623869, 46070247, 438436131, 3390992753, 22393107641, ... 248, 34504, 1022752, 16275496, 179862248, 1551303736, 11142792024, ...
$\frac{3}{2}$	$\frac{13}{16}$	1, 261, 24157, 580609, 8004754, 78925762, 618182705, 4079878514, ... 3393, 129688, 2270671, 25996789, 226351177, 1618088408, 9950251364, ... 116, 16964, 476876, 7131680, 74132236, 602971480, 4095721620, ...
$\frac{3}{2}$	$\frac{11}{16}$	1, 270, 18171, 375741, 4602852, 41167332, 296065548, 1809970083, ... 2871, 89991, 1380456, 14210922, 112987953, 745155153, 4259274975, ... 54, 8354, 221508, 3097278, 30156048, 230475996, 1475743590, 8240806224, ...

Table 7. Fourier coefficients for the characters of the U-series.

ℓ	1	2	3	4	5	6	7	8	9	10	ℓ
A_ℓ	3	8	15	24	35	48	63	80	99	120	$\ell^2+2\ell$
B_ℓ		10	21	36	55	78	105	136	171	220	$2\ell^2+\ell$
C_ℓ			21	36	55	78	105	136	171	220	$2\ell^2+\ell$
D_ℓ				28	45	66	91	120	153	190	$2\ell^2-\ell$
$\mathbb{C}$	1										
G_2		14									
F_4				52							
E_6						78					
E_7							133				
E_8								248			

Table 8. Ranks and dimensions of simple Lie algebras.

so that $\ell^2+\frac{1}{2}\leq\sum_{i\geq 1}\ell_i^2$, and this is impossible because each ℓ_i is a positive integer and ℓ is their sum. This shows, with a rather naive use of inequalities, that V_1 must have some component that is exceptional of type F_4, E_7 or E_8.

We will rework this argument. So essentially we backtrack because the inequalities can be improved as we gain more restrictions on the $\mathfrak{g}_i$. For any exceptional simple component $\mathfrak{g}_i$ of type F_4, E_7 or E_8 we write $\dim\mathfrak{g}_i=2\ell_i^2+\ell_i+e_i$ and let $1\leq i\leq e$ index such components. Note that $e_i\leq 14\ell_i$, with equality being met only if $\mathfrak{g}_1=E_8$. Then we have

$$\begin{aligned}2\ell^2+3\ell+1&\leq\ell_0+\sum_{i=1}^{e}(2\ell_i^2+\ell_i+e_i)+\sum_{i>e}\dim\mathfrak{g}_i\\&\leq\ell_0+\sum_{i=1}^{e}(2\ell_i^2+\ell_i+e_i)+\sum_{i>e}(2\ell_i^2+\ell_i)\\&=\ell+\sum_{i=1}^{e}e_i+2\sum_{i\geq 1}\ell_i^2\end{aligned}$$

It follows that

$$\begin{aligned}&2\ell^2+2\ell+1\leq\sum_{i=1}^{e}e_i+2\sum_{i\geq 1}\ell_i^2\\&\Rightarrow\ 2\ell_0^2+4\sum_{0\leq i<j}\ell_i\ell_j+2\sum_{i\geq 0}\ell_i+1\leq\sum_{i=1}^{e}e_i\leq 14\sum_{i=1}^{e}\ell_i\\&\Rightarrow\ (\ell_0^2+\ell_0)+2\sum_{0\leq i<j}\ell_i\ell_j<6\sum_{i=1}^{e}\ell_i\end{aligned}\tag{16}$$

Because the possible exceptional components are F_4, E_7 and E_8, and because there is at least one of them, the *minimum* of the ℓ_i $(1 \le i \le e)$ is *at least* 4 and *at most* 8. Then the previous inequality implies that

$$(\ell_0^2 + 9\ell_0) + 2\ell_1 \sum_{2\le j} \ell_j \quad + 2 \sum_{0\le i<j, i\ne 1} \ell_i \ell_j < 6\ell_1 + 6 \sum_{j=2}^{e} \ell_j$$

and so

$$(\ell_0^2 + 9\ell_0) + (2\ell_1 - 6)\Big(\sum_{j=2}^{e} \ell_j\Big) + 2\ell_1\Big(\sum_{j>e} \ell_j\Big) + 2 \sum_{0\le i<j, i\ne 1} \ell_i \ell_j < 6\ell_1. \tag{17}$$

From this we can deduce that

$$\sum_{j>e} \ell_j \le 2$$

If this is an equality then also

$$(2\ell_1 - 6) \sum_{j=2}^{e} \ell_j < 2\ell_1.$$

In this case we claim that $\sum_{j=2}^{e} \ell_j = 0$. To see this, denote the sum by Σ. Then $\ell_1(\Sigma - 1) < 3\Sigma$ so that $(\Sigma - 1) < \frac{3}{\ell_1}\Sigma$. But this is impossible if $\Sigma > 0$ because $\ell_1 \ge 4$ and $\Sigma \ge 4$.

By a very similar argument, suppose that $\sum_{j>e} \ell_j = 1$. This means that there is a unique component A_1 apart from those of types F_4, E_7, E_8. Moreover $(\Sigma - 2) < \frac{3}{\ell_1}\Sigma \le \frac{3}{4}\Sigma$, whence $\Sigma < 8$. And if $\ell_1 \ge 7$ then $\frac{4}{7}\Sigma < 2$, which is once again impossible unless $\Sigma = 0$. The conclusion is that if we have a component of type A_1 and two exceptional components then we must have $\ell_1 = 4$. Since ℓ_1 could have been chosen to be any of the exceptional Lie ranks, all of the exceptional components must be F_4.

We can argue similarly if all components are exceptional. In this case the main inequality (17) reads

$$\tfrac{1}{2}(\ell_0^2 + 9\ell_0) + (\ell_1 - 3)\Big(\sum_{j=2}^{e} \ell_j\Big) + \sum_{0\le i<j, i\ne 1} \ell_i \ell_j < 3\ell_1 \tag{18}$$

and all of the l_i are equal to 4, 7 or 8. So if there are at least three components then $\sum_{j\ge 2} \ell_j \ge 8$ and we can deduce that $8(\ell_1 - 3) + 16 < 3\ell_1$, i.e., $5\ell_1 < 8$, a contradiction. Similarly if there are two components and the second is *not* F_4 we obtain $7(\ell_1 - 3) < 3\ell_1$, whence $\ell_1 \le 5$ and the first component is F_4. So either way one of the two components must be F_4.

To summarize, so far, we've shown that one of the following must hold for the semisimple part, that is the *Levi factor* L of V_1:

- $L = \mathfrak{g}_1$,
- $L = F_4 \oplus \mathfrak{g}_1$,
- $L = A_1 \oplus \mathfrak{g}_1$,
- $L = A_1 \oplus F_4 \oplus F_4$,
- $L = A_1 \oplus A_1 \oplus \mathfrak{g}_1$,
- $L = A_2 \oplus \mathfrak{g}_1$.

In all cases $\mathfrak{g}_1$ is one of F_4, E_7 or E_8.

Now let's assume that there is no exceptional component of type E_8. Then $e_i \leq 4\ell_i$ and $\sum_{i=1}^{e} \ell_i \leq 11$. Going back to (16) we obtain

$$2\ell_0^2 + 4\sum_{0\leq i<j} \ell_i\ell_j + 2\sum_{i\geq 0}\ell_i + 1 \leq \sum_{i=1}^{e} e_i \leq 4\sum_{i=1}^{e}\ell_i$$

$$\Rightarrow\ (1+\ell_0^2+\ell_0) + 2\sum_{0\leq i<j}\ell_i\ell_j + \sum_{i>e}\ell_i < \sum_{i=1}^{e}\ell_i \leq 11$$

$$\Rightarrow\ (\ell_0^2+\ell_0) + 2\ell_0\sum_{1\leq i}\ell_i + 2\ell_1\Big(\sum_{i\geq 2}\ell_i\Big) + \sum_{i>e}\ell_i < 10.$$

Therefore $\ell_1\sum_{i\geq 2}\ell_i \leq 4$, which can only happen if $\ell_1 = 4$ and $\sum_{i\geq 2}\ell_i = 1$, or if $\sum_{i\geq 2}\ell_i = 0$. The latter equation means that L is simple. The first conditions mean that the first exceptional component is F_4, it is the only exceptional component, and if there are nonexceptional components they must comprise a single A_1.

In the simple case (not E_8) we have $(\ell_0^2+\ell_0) + 2\ell_0\ell_1 < 10$ and because $\ell_1 \geq 4$ then $\ell_0 = 0$. Observe, too, that if $V_1 = E_8$ then $\ell = \tilde{c} = c$, an impossibility because we are assuming that $\tilde{c} > \ell$. If E_8 is the only component then the very first inequality $2\ell^2 + 3\ell + 1 \leq \dim V_1 = \ell_0 + 248$ together with $\ell = \ell_0 + 8$ readily implies that $\ell_0 \leq 2$.

This allows us to refine the list of possibilities for V_1:

- $V_1 = F_4$,
- $V_1 = E_7$,
- $V_1 = \mathbb{C}^k \oplus E_8$, $(1 \leq k \leq 2)$
- $L = F_4 \oplus E_8$,
- $L = A_1 \oplus F_4$,
- $L = A_1 \oplus E_8$,
- $L = A_1 \oplus A_1 \oplus E_8$,
- $L = A_2 \oplus E_8$.

Here's another trick. We have $\dim V_1 = \frac{1}{2}s(s-1)$ for a positive integer s. This eliminates all possibilities when L is simple. Now we are obliged to look more closely at ℓ_0. In the absence of an E_8 component the possibilities are $L = A_1 \oplus F_4$, so $\ell_1 = 4$, $\ell_2 = 1$ and the last displayed inequality implies that $(\ell_0^2+\ell_0) + 10\ell_0 + 8 + 1 < 10$, in which case $\ell_0 = 0$ and $\dim V_1 = 3 + 52 = \frac{1}{2}s(s-1)$ with $s = 11$. Then $\tilde{c} = \frac{11}{2}$, $\ell = 5$. But we are assuming that $\tilde{c} - \ell \geq 1$, a contradiction. Now we're reduced to the following possibilities with an E_8 component:

- $L = F_4 \oplus E_8$,
- $L = A_1 \oplus E_8$,
- $L = A_1 \oplus A_1 \oplus E_8$,
- $L = A_2 \oplus E_8$.

In the first case we may apply (18). We have $\ell_1 = 4$, $\ell_2 = 8$, so $\frac{1}{2}(\ell_0^2 + 9\ell_0) + 8 + 8\ell_0 < 12$ is enough to force $\ell_0 = 0$. Therefore $V_1 = F_4 \oplus E_8$ has dimension $52 + 248 = 300$, so $s = 25$ and $c = \tilde{c} = \frac{25}{2}$, $\ell = 12$. Once again this is outside of the scope of the case under consideration, so this case does not occur.

In the second case we utilize (16) together with $\ell_1 = 8$, $\ell_2 = 1$, $e_1 = 112$ to find that $2\ell_0^2 + 38\ell_0 \leq 61$ and thus $\ell_0 \leq 1$. Then $\dim V_1 = 251$ or 252 and neither integer has the required form $\frac{1}{2}s(s-1)$. So this case does not occur.

The fourth case is similar, except that $\ell_2 = 2$. Just as before this leads to $\ell_0 = 0$, so $\dim V_1 = 256$, which does not conform to $\frac{1}{2}s(s-1)$, so this case does not occur.

For the third and final case we proceed similarly, but now with $\ell_1 = 8$, $\ell_2 = \ell_3 = 1$. As before this leads to $\ell_0 = 0$, $\dim V_1 = 254$ which once again is not of the form $\frac{1}{2}s(s-1)$. This completes the proof of Proposition 27. □

Proposition 29. *Let (m, x, y) be as in Theorem 26. Then Corollary 5(c) cannot hold.*

Proof. We are assuming here that $c = \tilde{c} = \ell = \frac{1}{2}s$ so by a theorem of Dong and Mason (see Theorem 4), we know that $V \cong V_L$ is a lattice theory for some even lattice L. Let the root system of L be denoted by L_2. The q-character of V is then the quotient of modular forms

$$f_0(\tau) = \frac{\theta_L(\tau)}{\eta(\tau)^\ell} = q^{-\ell/24}(1 + (\ell + |L_2|)q + \cdots).$$

We have $m = \frac{1}{2}s(s-1) = 2\ell^2 - \ell$. Therefore $|L_2| = 2\ell^2 - 2\ell$. Because L is an even lattice then its root system is the direct sum of simple root systems of types ADE. Let $\mathfrak{g}_i$ $(1 \le i \le N)$ be the nonabelian simple Lie algebra components of V_1, and let Φ_i be the root system of $\mathfrak{g}_i$, say of rank ℓ_i. Then we have

$$\begin{aligned} 2\Big(\sum_{r=1}^{N} \ell_r\Big)^2 - 2\sum_{r=1}^{N} \ell_r = 2\ell^2 - 2\ell &= \sum_{i=1}^{N} |\Phi_i| = \sum_i |\Phi_i| + \sum_j |\Phi_j| + \sum_k |\Phi_k| \\ &= \sum_i (\ell_i^2 + \ell_i) + \sum_j (2\ell_j^2 - 2\ell_j) + \sum_k (2\ell_k^2 - 2\ell_k + f_k) \\ &= \sum_i (-\ell_i^2 + 3\ell_i) + \sum_k f_k - 2\sum_{r=1}^{N} \ell_r + 2\sum_{r=1}^{N} \ell_r^2, \end{aligned}$$

where $|\Phi_i| = \ell_i^2 + \ell_i$, $2\ell_i^2 - 2\ell_i$, 72, 126, 240 for $\mathfrak{g}_i$ of type A_{ℓ_i}, D_{ℓ_i}, E_6, E_7, E_8, respectively, and where we use i, j, k to index the occurring root systems of type A, D, E respectively. We also have $f_k := 18, 42, 128$ for E_6, E_7, E_8 respectively. Note that $f_k < 16\ell_k$.

This begins to look like what we faced in the course of the proof of Proposition 27, where we first made a relatively naive estimate, then backtracked. The previous displayed equality yields

$$4 \sum_{1 \le r < s \le N} \ell_r \ell_s = -2\ell + \sum_i (-\ell_i^2 + 3\ell_i) + \sum_k f_k$$

Therefore, *if there is* a component of type E then for some ℓ_k, say with $k = N$, we have $\ell_N \ge 6$ and

$$\begin{aligned} 4\ell_N \sum_{r<N} \ell_r &< \sum_i (-\ell_i^2 + 3\ell_i) + \sum_k f_k \\ &\Rightarrow (4\ell_N - 3)\sum_i \ell_i + \sum_{k<N} (4\ell_N \ell_k - f_k) < f_N - \sum_i \ell_i^2. \end{aligned}$$

Because the sum over $k < N$ is nonnegative we then have

$$(4\ell_N - 3)\sum_i \ell_i + \sum_i \ell_i^2 < f_N$$

and so

$$19\sum_i \ell_i + \Big(\sum_i \ell_i\Big)^2 < f_N.$$

Now we find that if $f_N = 128$, then $\sum_i \ell_i \leq 5$; if $f_N = 42$ then $\sum_i \ell_i \leq 1$; and if $f_N = 18$ then $\sum_i \ell_i = 0$.

If $\sum_i \ell_i = 0$ then there are no type A components, and we then have

$$4\sum_{1\leq r<s\leq N} \ell_r\ell_s < \sum_k f_k \;\Rightarrow\; 4f_N \sum_{1\leq r<N} \ell_r < \sum_k f_k,$$

from which it follows easily that there is at most one nonzero type E component. And if there are any of type D, then $4f_N\sum_j \ell_j < f_N$, which is a contradiction. So we are reduced to the possibility that there is a single component, of type E. Then $240 = 2\ell^2 - 2\ell$, an impossibility. This shows that some $\ell_i > 0$.

We have therefore shown that *if* there is a type E component, then there must also be at least one type A component. Suppose there is a unique type E component. Then

$$4\ell_N \sum_i \ell_i < \sum_i(-\ell_i^2 + 3\ell_i) + f_N,$$

and so $\sum_i(\ell_i^2 + 21\ell_i) < f_N$, forcing $f_N = 42$ or 128. If $f_N = 42$ then necessarily $\{\ell_i\} = \{1\}$, i.e., there is a unique type A component and it is A_1. Then $L_2 = A_1 \oplus E_7$ and $|L_2| = 128 \neq 2\ell^2 - 2\ell$. Suppose that $f_N = 128$. Then $\sum_i \ell_i \leq 4$ and $|L_2| \in \{240, 242, 244, 246, 248, 250, 254, 260\}$, none of which are $2\ell^2 - 2\ell$. This shows that there are at least two type E components. In this case we have

$$4\ell_N \sum_i \ell_i + 4\sum_{k<k'} \ell_k\ell_{k'} < \sum_i(-\ell_i^2 + 3\ell_i) + \sum_k f_k$$

and so

$$\sum_i(\ell_i^2 + 21\ell_i) + 4\sum_{k<k'} \ell_k\ell_{k'} < \sum_k f_k,$$

and since each $\ell_k \geq 6$, and $4\ell_k\ell_{k'} > f_k$, then there can be no more than two type E components. Moreover they are both of type E_8, whence $\sum_i(\ell_i^2 + 21\ell_i) = 0$. Hence $L_2 = E_8 \oplus E_8$, $|L_2| = 480$, and $\ell = 16$. Now $L = E_8 \oplus E_8$, in which case $V = V_L$ is *holomorphic*, a contradiction.

We have finally shown that V_1 has *no* components of type E. So we have

$$\begin{aligned} 2\Big(\sum_{r=1}^N \ell_r\Big)^2 - 2\sum_{r=1}^N \ell_r = 2\ell^2 - 2\ell &= \sum_i |\Phi_i| + \sum_j |\Phi_j| \\ &= \sum_i(\ell_i^2 + \ell_i) + \sum_j(2\ell_j^2 - 2\ell_j) \\ &= \sum_i(-\ell_i^2 + 3\ell_i) - 2\sum_{r=1}^N \ell_r + 2\sum_{r=1}^N \ell_r^2, \end{aligned}$$

so that

$$2\Big(\sum_{r=1}^{N}\ell_r\Big)^2=\sum_i(-\ell_i^2+3\ell_i)+2\sum_{r=1}^{N}\ell_r^2$$

and $4\sum_{1\le r<s}\ell_r\ell_s=\sum_i(-\ell_i^2+3\ell_i)$. If there are no type A component then the right-hand side of this inequality vanishes, whence so does the left-hand side, meaning that there is a unique component, and it has type D. Here, then, we have $V\cong D_{\ell,1}$. However, this VOA has 4 simple modules if $l\ge 5$ and 2 if $l=4$. Thus this example does not occur. Suppose there are some type A components. Then the last displayed inequality implies that such a type A component is unique, call it $\mathfrak{g}_1$. Then

$$0=4\ell_1\sum_{2\le r}\ell_r=-\ell_1^2+3\ell_1,$$

so $\ell_1=3$ and $V=V_L\cong A_{3,1}$. Once again, this VOA has 4 simple modules so it does not occur. This completes the proof of Proposition 29. □

The final case is:

Proposition 30. *Assume that (m,x,y) is as in Theorem 26 and that Corollary 5(b) holds. Then $V\cong B_{\ell,1}$, $A_{1,2}$ or $\mathrm{Vir}(c_{3,4})$.*

Proof. In this case we have $c=\tilde{c}=\frac{1}{2}s$, $\ell=\tilde{c}-\frac{1}{2}=\frac{1}{2}(s-1)$, $m=\frac{1}{2}(s(s-1))=2\ell^2+\ell$.

Now we have seen that the q-character of V (and that of its simple modules, too) is uniquely determined by this data. It follows that the q-character of V is equal to that of one of the VOAs in the statement of the proposition.

Suppose first that $\ell=0$. Then $s=1$, $\tilde{c}=c=\frac{1}{2}$, and by [Mason 2014, Theorem 8], it follows that V contains the Virasoro VOA $\mathrm{Vir}(c_{3,4})$ as a subVOA. However from the last paragraph V has the same q-character as this Virasoro VOA and therefore they are equal. This proves the proposition if $\ell=0$. Thus from now on we may, and shall, assume that $V_1\ne 0$. We would like to then show that V_1 is isomorphic to B_ℓ, or A_1.

Suppose that $\ell=1$. Then $m=3$ and $V_1\cong A_1$. By [Dong and Mason 2006] the subVOA $U:=\langle V_1\rangle$ generated by V_1 is isomorphic to an affine algebra $A_{1,k}$ of some positive integral level k. Now we can use the *majorizing theorem* in Appendix B to see that because the q-character of V is the same as that for $A_{1,2}$ by the first paragraph, then $k\le 2$, and if $k=2$ then $U=V\cong A_{1,2}$. Suppose that $k=1$. Then the commutant C of U has central charge $\frac{1}{2}$. Now consider $U\otimes C$: it is a subVOA of V and from what we have said it majorizes $A_{1,1}\otimes\mathrm{Vir}(c_{3,4})$ or is equal to it. But this latter VOA itself majorizes $A_{1,2}$ as one sees by a direct check of q-expansions, and this shows that the case $k=1$ does *not* occur.

Now suppose that $\ell\ge 2$. By the first paragraph V has the same q-character as $B_{\ell,1}$. If we can show that $V_1\cong B_\ell$ then the same arguments used in the previous paragraph show that $V\cong B_{\ell,1}$, and the proposition will be proved.

We can attack this much as we did in the proofs of Propositions 27 and 29. Let V_1 have Levi decomposition (15). Then $\ell = \ell_0 + \sum_i \ell_i$. Let Φ_i be the root system of $\mathfrak{g}_i$. Then

$$\begin{aligned} 2\ell^2 + \ell = 2\Big(\sum_{i\geq 0}\ell_i\Big)^2 + \Big(\sum_{i\geq 0}\ell_i\Big) = \dim V_1 = \ell_0 + \sum_{i\geq 1}(\ell_i + |\Phi_i|) \\ \Rightarrow\ 2\Big(\sum_{i\geq 0}\ell_i\Big)^2 = \sum_{i\geq 1}|\Phi_i| \\ \Rightarrow\ 2\ell_0^2 + 4\ell_0\sum_{i\geq 1}\ell_i + 4\sum_{1\leq i<j}\ell_i\ell_j = \sum_{i\geq 1}(|\Phi_i| - 2\ell_i^2). \end{aligned} \tag{19}$$

Now $|\Phi_i| - 2\ell_i^2 = \ell_i - \ell_i^2$; 0; $-2\ell_i$; 6; 20; 6; 35; 120 for types A_{ℓ_i}; B_{ℓ_i} or C_{ℓ_i}; D_{ℓ_i}; G_2; F_4; E_6, E_7, E_8, respectively.

Suppose that the left-hand side of (19) is 0. Then $\ell_0 = 0$, V_1 has a unique component, and it has type B or C. If the type is B then $V \cong B_{\ell,1}$ as we have already explained, so we are done in this case. If the type is *not* B then $V_1 \cong C_\ell$ with $\ell \geq 3$. By Theorem 1.1 of [Dong and Mason 2006] it follows that the subVOA $U := \langle V_1\rangle$ generated by V_1 is isomorphic to $C_{\ell,k}$ for some positive integral level k. Since U is generated by weight 1 states, a consideration of the conformal subVOA $U \otimes C$, where C is the commutant of U, shows that $\dim U_2 \leq \dim V_2$ and hence that $\dim(C_{\ell,k})_2 \leq \dim(B_{\ell,1})_2$. However this contradicts Theorem 41 in Appendix B. This proves Proposition 30 if the left-hand side of (19) is 0.

This reduces us to consideration of the case that the left-hand side of (19) is *positive*, so the right side is too. So there must be at least one *exceptional* component

Suppose there are k components of type E_8, and r exceptional components *not* of type E_8. The right side of (19) is at most $120k + 35r$, whereas the left side of (19) is at least $4\big(64\binom{k}{2} + 16kr + 4\binom{r}{2}\big)$. Therefore

$$\begin{aligned} 16(k^2 - k) + 8kr + (r^2 - r) \leq 15k + \tfrac{35}{8}r \\ \Rightarrow 16k^2 - 31k + 8kr + (r^2 - r) \leq \tfrac{35}{8}r \end{aligned}$$

It follows easily that $k \leq 1$, and if $k = 1$ then $r^2 - 15 \leq -\frac{29}{8}r \Rightarrow r \leq 2$. Again with $k = 1$ we can argue more precisely that if the exceptional components are $\mathfrak{g}_1 = E_8$, $\mathfrak{g}_2$, $\mathfrak{g}_3$ then

$$4(8\ell_2 + 8\ell_3 + \ell_2\ell_3) \leq 120 + (|\Phi_2| - 2\ell_2^2) + (|\Phi_3| - 2\ell_3^2)$$

and the two terms on the right-hand side are among $\{6, 20, 6, 35\}$, and ℓ_2, ℓ_3 are each one of $\{2, 4, 6, 7\}$. We see that this can never hold.

This shows that $k = 0$, i.e., there are *no* components of type E_8. Repeating the argument if there are k' components of type E_7 and r' other exceptional components, then

$$\begin{aligned}
&4\left(49\binom{k'}{2}+4\binom{r'}{2}+14k'r'\right) \le 35k'+20r' \\
&\Rightarrow\ 98k'(k'-1)+8r'(r'-1)+56k'r' \le 35k'+20r' \\
&\Rightarrow\ \tfrac{49}{2}k'^2+4r'^2+14k'r' \le \tfrac{133}{4}k'+7r' \\
&\Rightarrow\ \tfrac{49}{2}\left(k'-\tfrac{19}{4\cdot 7}\right)^2+4\left(r'-\tfrac{7}{8}\right)^2+14k'r' \le \tfrac{1}{2}\left(\tfrac{19}{4}\right)^2+\tfrac{49}{16} < \tfrac{361}{32}+3\tfrac{1}{16} < 15.
\end{aligned}$$

We readily deduce that at least one of k' or r' is 0. Thus if there are any exceptional components then either there is an E_7 and no other exceptional component, or else there are no exceptional components of type E_8 or E_7. In the former case, if $V_1 = E_7$ then the right side of (19) is odd, while the left side is even, a contradiction. If there are no E_8, E_7 components, then as before we have in case there are t exceptional components that $8t(t-1) \le 20t$, which implies $(t^2-4t) \le 0$, so $t \le 2$. But if $t = 2$ we get equality, meaning two F_4 components and $l\ell_1 = \ell_2 = 4$, impossible.

Thus $t = 1$, i.e., there is a unique exceptional component, and $2\ell_0^2 + 8\ell_0 \le 20$. Then $\ell_0 \le 1$ and $\dim V_1 = 2\ell^2+\ell = 14(15), 52(53), 78(79)$ (parentheses denotes the case $\ell_0 = 1$), which can only occur when $\ell = 6$, $\ell_0 = 0$ and $V_1 = E_6$. Furthermore $c = \tilde{c} = \frac{13}{2}$ and the commutant of $U := \langle V_1\rangle$ is isomorphic to $\mathrm{Vir}_{c_{3,4}}$. Note that $U \cong E_{6,k}$ for some positive integral k by [Dong and Mason 2006]. But it now follows that U has central charge 6. Since $E_{6,k}$ has $c = 78k/(k+12)$ we must have $k = 1$.

Now $m = 78 = s(s-1)/2$, so $s = 13$. Therefore

$$h_1 = x+1 = \left(\frac{s}{16}-1\right)+1 = \frac{13}{16} \quad \text{and} \quad h_2 = y+1 = \frac{1}{2}$$

and the conformal weights of V are $\{0, \frac{1}{2}, \frac{13}{16}\}$. Now the conformal weights for the simple $E_{6,1}$-modules are $\{0, \frac{2}{3}\}$ while those for $\mathrm{Vir}(c_{3,4})$ are $\{0, \frac{1}{2}, \frac{1}{16}\}$. Since $E_{6,1} \otimes \mathrm{Vir}(c_{3,4})$ is a conformal subVOA of V, it is impossible to reconcile the conformal weights of the tensor product with those for V. Thus this case cannot occur. This finally completes the proof of Proposition 30. □

With these propositions in hand we have completed the proof of Theorem 26. There remain two outstanding cases, enumerated as (2) and (3) on page 1642. As noted, there are examples of VOAs with the relevant numerical data in both cases, namely $\mathrm{Vir}(c_{2,7})$ and $A_{4,1}$. In the first case we have $m = 0$ and here we may appeal to the main result of [Arike et al. 2017] to immediately conclude that indeed $V \cong \mathrm{Vir}(c_{2,7})$.

For the sake of brevity we sketch how to prove nonexistence in example (3). First note that inasmuch as the data determines the character vector of V it follows in particular that the character of V coincides with that of $A_{4,1}$. Now we may proceed much as in the proofs of the three propositions, although here it is much easier because we already know that $m = 24$. We have $c = \tilde{c} = 4$ and $\ell \le c$. If $\ell = c$ then V is a lattice theory and as before we find that $V_1 = A_4$ and then that $V \cong A_{4,1}$. However this VOA has more than three simple modules so it cannot occur. If $\ell < c$ we obtain a contradiction as in the propositions.

Alternatively, we first identify the Lie algebra $V_1 = A_4$ then conclude that $\langle V_1 \rangle \cong A_{4,k}$ for some integral level k. Now apply the majorization argument and knowledge of the character to get $V \cong A_{4,1}$, and hence a contradiction as before.

This finally completes our proof of Theorem 1.

9. The U-series

By their very definition, potential VOAs that belong to the U-series have exactly three simple modules and survive all of the numerical tests that we have so far applied. From an arithmetic perspective they are exquisitely balanced.

In this section we discuss further properties of these VOAs, especially the question of whether they actually *exist*. We shall present some results that render it very likely that there are 15 VOAs in the U-series. See Remark 2. Two of these examples are well-known in the literature, namely $E_{8,2}$ and Höhn's baby monster VOA $\mathrm{VB}^{\natural}_{(0)}$ [1996]. The remaining examples come about by an application of the results of [Gaberdiel et al. 2016; Lin 2017]. These works are applicable on the basis of an apparent and surprising connection between VOAs in the U-series and VOAs X on the Schellekens list [1993] of holomorphic VOAs of central charge $c = 24$. Indeed, we propose Hypothesis S below, which is a natural assumption about gluing VOAs and which leads to the identification of the U-series VOAs with certain commutants of subalgebras for various choices of X.

9.1. ***Connections with the Schellekens list.*** Let us record some of the properties of a VOA V that lies in the U-series:

(i) V is strongly regular and has just 3 simple modules $M_0 = V$, M_1, M_2.

(ii) The q-characters $f_j(\tau)$ of the M_j are each congruence modular functions of weight 0 with nonnegative integral Fourier coefficients described explicitly in Table 7.

(iii) The character vector $F = (f_0, f_1, f_2)^T$ is a vector-valued modular form whose associated MLDE is monic with irreducible monodromy ρ.

(iv) There is an integer p in the range $5 \le p \le 15$ such that the central charge c, the dimension m of the Lie algebra on V_1, and the *conformal weights* h_j of the M_j are as follows (see Table 6):

$$c = p + \tfrac{17}{2}, \quad m = (15-p)(2p+17), \quad h_0 = 0, \quad h_1 = \tfrac{3}{2}, \quad h_2 = \tfrac{1}{16}(2p+1).$$

The formula for m derives from that for the elliptic surface (10).

(v) The S-matrix is

$$\rho(S) = \frac{1}{2}\begin{pmatrix} 1 & 1 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 0 \end{pmatrix} \tag{20}$$

with lexicographic ordering. In particular the fusion rules for V are the same as the Ising model $\mathrm{Vir}(c_{3,4})$. Especially, it follows that M_1 has quantum dimension 1 and is a *simple current*.

Now let k be a nonnegative integer. We define a family of VOAs $V^{(k)}$ as follows:

$$V^{(k)} := \begin{cases} \mathrm{Vir}(c_{3,4}) & \text{for } k = 0, \\ A_{1,2} & \text{for } k = 1, \\ B_{k,1} & \text{for } k \geq 2. \end{cases}$$

As a reminder, from Table 1 we see that, like VOAs in the U-series, $V^{(k)}$ is a simple VOA with just three simple modules. Denote these by $V^{(k)}$, M_1', M_2', say with conformal weights 0, $h_1' = \frac{1}{2}$ and $h_2' = \frac{1}{16}(2k+1)$, respectively. The central charge of $V^{(k)}$ is equal to $c_k := \frac{1}{2}(2k+1)$.

Now choose any VOA in the U-series with parameter p as before, and denote this VOA by $W^{(p)}$ and choose $k := 15 - p$, so that $0 \leq k \leq 10$. For this choice of k the tensor product VOA

$$T^k := W^{(p)} \otimes V^{(k)}$$

is a simple VOA with central charge equal to $p + \frac{17}{2} + \frac{1}{2}(2k+1) = 24$. Let us consider the T^k-module

$$X := (W^{(p)} \otimes V^{(k)}) \oplus (M_1 \otimes M_1') \oplus (M_2 \otimes M_2'). \tag{21}$$

Höhn calls this procedure *gluing* $W^{(p)}$ and $V^{(k)}$. Each $M_j \otimes M_j'$ is a simple module for T^k, $j = 1, 2$. The next result is very useful.

Lemma 31. *The conformal weights of $M_j \otimes M_j'$ for $j = 1, 2$ are both equal to 2. In particular the conformal grading on X is* ***integral****.*

Proof. We have $h_1 + h_1' = \frac{3}{2} + \frac{1}{2} = 2$ and $h_2 + h_2' = \frac{1}{16}(2p+1) + \frac{1}{16}(2k+1) = 2$. The lemma follows. □

Corollary 32. *The conformal weight 1 piece X_1 of X satisfies*

$$X_1 = T_1^k = W_1^{(p)} \oplus V_1^{(k)}.$$

Let $\chi := \chi_X = \mathrm{Tr}_X\, q^{L(0)-1}$ be the q-character of X. It follows from Lemma 31 that

$$\chi \in q^{-1}\mathbb{Z}[[q]]. \tag{22}$$

Lemma 33. *χ is the modular function of level 1 and weight 0 given by*

$$\chi = J(q) + 48k.$$

where $J(q) := q^{-1} + 196884q + \cdots$ is the absolute modular invariant with constant term 0.

Proof. After (22), χ is invariant under the T-action $\tau \mapsto \tau + 1$. So to prove that χ is modular of level 1 it suffices to establish invariance under the action of S. This will follow directly by a formal calculation based on the nature of $\rho(S)$ (20). For the VOAs $W^{(p)}$ and $V^{(k)}$ have identical S-matrices. Therefore if we formally let $\{e_j\}$, $\{f_j\}$ $(j = 1, 2, 3)$ index bases with respect to which the two S-matrices are written then

$$S \otimes S : \sum_i e_i \otimes f_i \mapsto \tfrac{1}{4}\big\{2(e_1 \otimes f_1 + e_2 \otimes f_2 + 2e_3 \otimes f_3) + 2(e_1 \otimes f_1 + e_2 \otimes f_2)\big\} = \sum_i e_i \otimes f_i,$$

which is the required S-invariance.

It is well-known [Zhu 1996] that the q-characters of simple modules for strongly regular VOAs are holomorphic in the complex upper-half plane. Therefore χ is modular of level 1 with a simple pole at ∞ and no other poles, and leading coefficient 1. It follows that $\chi = J(\tau) + \kappa$ for a constant κ.

To compute the constant κ, which is equal to $\dim X_1$, use Corollary 32 to see that

$$\dim X_1 = \dim W_1^{(p)} + \dim V_1^{(k)} = m + \dim B_k = (15-p)(2p+17) + (2k^2+k) = 48k.$$

This completes the proof. □

Lemma 33 naturally suggests:

Hypothesis S. X carries the structure of a holomorphic VOA containing T^k as a subVOA. X is therefore a holomorphic VOA of central charge 24, that is, it is on the Schellekens list.

Hypothesis S is completely analogous to [Höhn 1996, Vermutung 3.2.1]. It suggests where we should look to find VOAs in the U-series. We consider this option in the next subsection.

9.2. ***Existence of U-series VOAs.*** Throughout this subsection, and for the sake of comparison, we generally use notation similar to that of the previous subsection. In particular, we now fix X to be a VOA on the Schellekens list. For a recent survey on the status of the VOAs in the Schellekens list, we refer the reader to [Lam and Shimakura 2019]. In particular, the Schellekens list VOAs intervening in Table 9 exist and they are unique. Let $V^{(k)} \subseteq X$ be a subVOA isomorphic to an affine algebra as in the previous subsection such that the weight 1 piece $V_1^{(k)}$ is a simple Lie algebra component of X_1 isomorphic to either $A_{1,2}$ or $B_{k,1}$.

This assumption involves some *exclusions*. First, the case $k = 0$ and $V^k = \mathrm{Vir}(c_{3,4})$ does not occur. This case is somewhat exceptional and was, in any case, handled by Höhn. Secondly, the cases $k = 7, 9, 10$ do not occur either, but for a different reason. Namely because there is no X with such a subalgebra (see Table 9).

Lemma 34. *We have $V^{(k)} = C(C(V^{(k)}))$, i.e., $V^{(k)}$ coincides with its double commutant in X.*

Proof. Let $D := C(C(V^{(k)}))$ be the double commutant in question. It is a subVOA of X that contains $V^{(k)}$, and we have $D_1 = V_1^{(k)}$. Furthermore D and $V^{(k)}$ share the *same Virasoro element*. On the other hand, of the three simple modules for $V^{(k)}$, the adjoint module is the only one that has integral conformal grading. Now the equality $D = V^{(k)}$ follows. □

Continuing earlier notation, we set

$$W^{(p)} := C(V^{(k)}). \tag{23}$$

Note the difference, however. Our earlier $W^{(p)}$ was the hypothesized U-series VOA, whereas now there is no question about its existence. What *is* in doubt is whether $W^{(p)}$ as defined in (23) is in the U-series. Basically, this comes down to the question, does $W^{(p)}$ have exactly three simple modules? In the next few paragraphs we will state and prove what we know about this. We will show on the basis of the results of [Gaberdiel et al. 2016] that $W^{(p)}$ is indeed in the U-series. This work, while undoubtedly correct, was not developed on an axiomatic basis.

We will need to use the following standard conjecture concerning commutants in a strongly regular VOA. In the present context it says:

Hypothesis C. $W^{(p)}$ is a strongly regular VOA.

Proceeding on the basis of Hypothesis C (actually, we only need $W^{(p)}$ to be rational and C_2-cofinite) we first prove:

Lemma 35. *The decomposition (21) holds, where M_1', M_2' are the nonadjoint simple modules for $V^{(k)}$, labeled according to the S-matrix (20) and M_1, M_2 are simple modules for $W^{(p)}$.*

Proof. Because we are assuming Hypothesis C, this follows from results of Lin [2017, (1.1)]. □

Furthermore we have:

Proposition 36. *The following hold*:

(a) *M_1 and M_1' are simple currents for $W^{(p)}$ and $V^{(k)}$ respectively.*

(b) *$V^{(k)} \oplus M_1'$ and $W^{(p)} \oplus M_1$ are both rational super VOAs.*

(c) *$(W^{(p)} \otimes V^{(k)}) \oplus (M_1 \otimes M_1')$ is a conformal subVOA of X.*

Proof. That M_1' is a simple current for $V^{(k)}$ was already pointed out following (20). Indeed, these simple currents for affine algebras are well-known. That $V^{(k)} \oplus M_1'$ is a rational super VOA is proved in [Dong et al. 1996, Examples 5.11 and 5.12]. As for M_1, that it is a simple current for $W^{(p)}$ follows from the existence of M_1' and the duality between module subcategories proved by Lin [2017]. Now it follows that the subspace $(W^{(p)} \otimes V^{(k)}) \oplus (M_1 \otimes M_1') \subseteq X$ is closed with respect to products; hence it is a subVOA. Therefore $V^{(k)} \oplus M_1'$ is itself a super VOA because we have already seen that $W^{(p)} \oplus M_1$ is. This completes the proof. □

Our main result is the following:

Theorem 37. *$W^{(p)}$ is a VOA in the U-series.*

Proof. The main point in the proof is the work of Gaberdiel, Hampapura, and Mukhi [2016]. These authors consider the properties of commutants of affine algebras such as $V^{(k)}$ from the perspective of MLDEs. They are able to show, in the framework that we are working, that the commutant $W^{(p)}$, while not necessarily having exactly 3 simple modules (which is what we need), at least satisfies $\dim \mathfrak{ch}_{W^{(p)}} = 3$. That is, the space of q-characters for the simple $W^{(p)}$-modules is 3-dimensional. Note that we know the q-characters of the simple $W^{(p)}$-modules that are contained in X and that they furnish an irreducible representation ρ of Γ. As a check, [Gaberdiel et al. 2016] describes the MLDE satisfied by these characters and one can check from their tables that the conformal weights of these modules are precisely those of the U-series that we have already calculated from a completely different perspective.

#	X_1	k
5	$(A_{1,2}^{15}) \oplus A_{1,2}$	1
7	$(A_{3,4}^{3})A_{1,2}$	1
8	$(A_{5,6}C_{2,3}) \oplus A_{1,2}$	1
10	$(D_{5,8}) \oplus A_{1,2}$	1
25	$(D_{4,2}^2 B_{2,1}^3) \oplus B_{2,1}$	2
26	$(A_{5,2}^2 A_{2,1}^2) \oplus B_{2,1}$	2
28	$(E_{6,4}A_{2,1}) \oplus B_{2,1}$	2
39	$(D_{6,2}C_{4,1}B_{3,1}) \oplus B_{3,1}$	3
40	$(A_{9,2}A_{4,1}) \oplus B_{3,1}$	3
47	$(D_{8,2}B_{4,1}) \oplus B_{4,1}$	4
48	$(C_{6,1}^2) \oplus B_{4,1}$	4
53	$(E_{7,2}F_{4,1}) \oplus B_{5,1}$	5
56	$(C_{10,1}) \oplus B_{6,1}$	6
62	$(E_{8,2}) \oplus B_{8,1}$	8

Table 9. VOAs on the Schellekens list with X_1 having a summand $B_{k,1}$ or $A_{1,2}$. Columns give the Schellekens list number, the structure of the Lie algebra X_1 with levels, and the k-value.

From these comments it follows that we can organize the simple modules for $W^{(p)}$ into sets of 3 so that the corresponding matrix representation of Γ on the q-characters looks like

$$\begin{pmatrix} \rho & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \rho \end{pmatrix}$$

In particular, the S-matrix has a similar block diagonal decomposition. However, at least in its action on the 1-point functions for $W^{(p)}$ (genus 1 conformal block), the S-matrix has first row with only nonzero entries. It follows immediately from the displayed block diagonal matrix that this can only happen if it is a 1×1 block matrix. That is, there are only 3 simple modules for $W^{(p)}$. This completes the proof. □

Example 38. When $k = 8$ we see from Table 9 that $W_1^{(p)} = E_8$ and $W^{(p)} = E_{8,2}$. This equality holds because $E_{8,2}$ itself has only three simple modules. Thus this U-series VOA is well-known, as is its simple current M_1 and the super VOA $E_{8,2}\oplus M_1$ (see [Dong et al. 1996]). This case was first handled by Höhn [1996].

We obtain 14 different VOAs in the U-series, including the know $E_{8,2}$, as we see from Table 9. The other 13 are probably new.

In the exceptional case when $k = 0$ we have $X_1 = 0$, so it was not considered in [Gaberdiel et al. 2016]. In any case Höhn already proved, under the natural assumption that $X = V^\natural$ is the Frenkel–Lepowsky–Meurman moonshine module, that the commutant of $V^{(0)}$ is the baby monster VOA $\mathrm{VB}^\natural_{(0)}$. So this example also falls into the U-series.

Appendix A: Primes in progressions

In Section 6 we cut down the possible character vectors for VOAs occurring in Theorem 1 by making use of hypergeometric formulas for the character vector. To prove nonintegrality of the vectors *not* contributing to Theorem 1, we produced nontrivial denominators in all but finitely many cases. Our argument relies on the existence of primes in progressions that lie in specific intervals. In this short appendix we explain how to use effective versions of the prime number theorem for primes in arithmetic progressions to prove what we need. These sorts of results, which go back to Bertrand's postulate that there is always a prime between x and $2x$, are well-known to analytic number theorists. A recent paper [Bennett et al. 2018], which establishes effective versions of prime number theorems for arithmetic progressions, enables us to get the precise results necessary for our application to the problem of classifying VOAs as in Theorem 1. To treat solutions (m, x, y) to (10) with $y = a/5$ where a is an integer coprime to 5, we make use in Section 6 of the following result:

Theorem 39. *If $X > 6496$ then the interval $\left[X, \frac{28}{27}X\right]$ contains at least one prime from each congruence class $a \pmod{30}$ with $\gcd(a, 30) = 1$.*

Proof. Let $\pi(X; q, a)$ denote the prime counting function for primes $p \equiv a \pmod q$. By Theorem 1.3 of [Bennett et al. 2018]

$$\left|\pi(X; q, a) - \frac{\mathrm{Li}(X)}{\phi(q)}\right| < c_\pi(q)\frac{X}{(\log X)^2}$$

for all $X \geq x_\pi(q)$ for explicit constants $c_\pi(q)$ and $x_\pi(q)$ that are independent of a. We are interested in the function

$$F(X) = \pi(28X/27; q, a) - \pi(X; q, a).$$

We must prove that there exists an N such that $F(X) \geq 1$ for all $x > N$. Notice that if $X \geq x_\pi(q)$,

$$\pi(28X/27; q, a) > \frac{\mathrm{Li}(28X/27)}{\phi(q)} - \frac{28c_\pi(q)}{27}\frac{X}{(\log X + \log(28/27))^2}$$
$$-\pi(X; q, a) > -\frac{\mathrm{Li}(X)}{\phi(q)} - c_\pi(q)\frac{X}{(\log X)^2}$$

Therefore for $X \geq x_\pi(q)$,

$$|F(X)| > \frac{\mathrm{Li}(28X/27) - \mathrm{Li}(X)}{\phi(q)} - c_\pi(q)X\left(\frac{28}{27}\frac{1}{(\log X + \log(28/27))^2} - \frac{1}{(\log X)^2}\right)$$

Taking $q = 30$, [Bennett et al. 2018] gives $x_\pi(30) = 789693271$ and $c_\pi(30) = 0.0005661$. One sees that for $X \geq x_\pi(30)$, $|F(X)|$ is much larger than 1. To prove the theorem for $6496 < X < x_\pi(30)$ we used a computer to verify it in the remaining cases. □

Other solutions to (10) can be treated in a similar manner, where the relevant moduli are $6 \cdot 7 = 42$ (primitive fibers) and $6 \cdot 16 = 96$ (imprimitive fibers with $y \neq -\frac{1}{2}$); both of these moduli are treated in [Bennett et al. 2018].

Appendix B: Affine algebras

Let $\mathcal{G}$ be a finite-dimensional simple Lie algebra of type A, B, C, D, E, F or G and Lie rank ℓ (dimension of a Cartan subalgebra). In this appendix we discuss some properties of the universal vertex algebra $V(\mathcal{G}, k)$ of level k and its simple quotient VOA $\mathcal{G}_{\ell,k}$, which is often called a WZW model when k is a positive integer. For convenience, the constructions of these VOAs will be recalled below. For additional background, see [Kac 1990; Lepowsky and Li 2004].

B.1: ***Statement of the main results.*** There are two main results that we intend to prove in this appendix, both having to do with the conformal grading of WZW models. The first one we call the *majorization theorem*. As a referee has pointed out, this result may not be new, but we are unaware of a good reference:

Theorem 40 (majorization). *Fix the type $\mathcal{G}$ and Lie rank ℓ. Regarding $\mathcal{G}_{\ell,k}$ as a linear space equipped with its conformal $\mathbb{Z}$-grading, there are* ***surjective*** *$\mathbb{Z}$-graded morphisms*

$$\mathcal{G}_{\ell,k'} \to \mathcal{G}_{\ell,k}$$

for all positive integral $k' \geq k$.

The second result is more specialized:

Theorem 41. *For all positive integers k, ℓ, we have*

$$\dim(C_{\ell,k})_2 \geq \dim(B_{\ell,1})_2$$

with equality only if $\ell = 2$.

Remark 42. The proof will show that

$$\dim(C_{\ell,k})_2 - \dim(B_{\ell,1})_2 \geq 2\ell - 4.$$

B.2: ***The universal affine VOA $V(\mathcal{G}, k)$.*** Let $\mathcal{G}$ be a finite-dimensional nonabelian simple Lie algebra with Killing form $\langle\ ,\ \rangle$. The affine algebra associated to $\mathcal{G}$ is the Lie algebra defined by

$$\mathcal{G} \otimes \mathbb{C}[t, t^{-1}] \oplus \mathbb{C}K,$$

where K is a central element and the nontrivial brackets are

$$[a \otimes t^m, b \otimes t^n] := [a, b] \otimes t^{m+n} + m\delta_{m+n,0}\langle a, b\rangle K$$

for $a, b \in \mathcal{G}$. There is a natural triangular decomposition

$$\widehat{\mathcal{G}} := \widehat{\mathcal{G}}^+ \oplus \widehat{\mathcal{G}}_0 \oplus \widehat{\mathcal{G}}^-$$

with

$$\widehat{\mathcal{G}}^\pm := \mathcal{G} \otimes t^{\pm 1}\mathbb{C}[t^{\pm 1}], \qquad \widehat{\mathcal{G}}_0 := \mathcal{G} \otimes t^0 \oplus \mathbb{C}K \cong \mathcal{G} \oplus \mathbb{C}.$$

$\widehat{\mathcal{G}}$ is also naturally $\mathbb{Z}$-graded by

$$\widehat{\mathcal{G}} = \bigoplus_{n \in \mathbf{Z}} \widehat{\mathcal{G}}_n, \quad \widehat{\mathcal{G}}_n := \mathcal{G} \otimes t^{-n} \qquad (n \neq 0),$$

so that $[\widehat{\mathcal{G}}_m, \widehat{\mathcal{G}}_n] \subseteq \widehat{\mathcal{G}}_{m+n}$.

Choose any scalar (the *level*) $k \in \mathbb{C}$, and let $\mathbb{C}_k$ denote the 1-dimensional $(\mathcal{G}^+ \oplus \widehat{\mathcal{G}}_0)$-module defined as follows: $\mathcal{G}^+$ acts as 0; $\mathcal{G} = \mathcal{G} \otimes t^0$ acts as 0; K acts as multiplication by the level k. The corresponding *Verma-module* is the induced module

$$V = V(\mathcal{G}, k) := \mathcal{U}(\widehat{\mathcal{G}}) \otimes_{\mathcal{U}(\widehat{\mathcal{G}}^+ \oplus \widehat{\mathcal{G}}_0)} \mathbb{C}_k,$$

where, here and below, $\mathcal{U}$ denotes universal enveloping algebra. Using the PBW theorem and the triangular decomposition for $\widehat{\mathcal{G}}$, one sees that V is linearly isomorphic to the symmetric algebra $S(\widehat{\mathcal{G}}^-)$. The *conformal grading* on the symmetric algebra is related to the grading on $\widehat{\mathcal{G}}$ in which $a \otimes t^{-n}$ ($n \geq 1$, $a \in \mathcal{G}$) has weight (i.e., degree) n and the *vacuum element* $\mathbf{1} := 1 \otimes 1$ has weight 0.

$$V = V(\mathcal{G}, k) \cong S(\widehat{\mathcal{G}}^-) = \oplus_{n \geq 0} S(\widehat{\mathcal{G}}^-)_n, \tag{24}$$

where

$$S(\widehat{\mathcal{G}}^-)_0 = \mathbb{C}\mathbf{1}, \qquad S(\widehat{\mathcal{G}}^-)_1 = \mathcal{G} \otimes t^{-1}.$$

As long as k is a positive integer (the only case that we care about) then V carries the structure of a VOA, and the grading on V induced by the $L(0)$-operator of the Virasoro element is the conformal grading we just described. An obvious — though important — point is that this is *independent of the level* k.

B.3: ***The quotient VOA $L(\mathcal{G}, k)$ and the majorization theorem.*** We continue to discuss the VOAs $V := V(\mathcal{G}, k)$, always with k a positive integer. Up to scalars, V admits a unique nonzero, invariant, bilinear form b_V by a theorem of Li [1994], however b_V is always degenerate for the values of k under consideration. The *radical* of b_V is the unique maximal 2-sided ideal in V, and we denote the simple quotient VOA by

$$L(\mathcal{G}, k) := V(\mathcal{G}, k)/\operatorname{Rad}(b_V).$$

If $\mathcal{G}$ has type $A, B, \ldots, F, G$ and Lie rank ℓ we will often denote this VOA by $\mathcal{G}_{\ell,k}$.

A fundamental theorem for us is the determination of $\operatorname{Rad}(b_V)$ by Kac [1990]. See also [Lepowsky and Li 2004, Proposition 6.6.17]. To state the result concisely, we need some notation. Let Φ be the root system of $\mathcal{G}$ and let $\theta \in \Phi$ be the (unique) positive root of maximal height. Let $S_\theta \subseteq \mathcal{G}$ be a fundamental $\mathfrak{sl}_2$-subalgebra determined by θ having a Chevalley basis $\{e_\theta, f_\theta, h_\theta\}$.

Proposition 43 (Kac). *We have*

$$\mathrm{Rad}(b_V) = \mathcal{U}(\widehat{\mathcal{G}})e_\theta(-1)^{k+1}\mathbf{1}. \qquad \square$$

We are now ready for:

Proof of majorization Theorem 40. We have seen that, considered as just a $\mathbb{Z}$-graded linear space, $V(\mathcal{G}, k)$ coincides with the graded symmetric algebra (24) which does not depend on k. From Kac's theorem it is clear that the radical ideals $R_{\ell,k} := b_{V(\mathcal{G},k)}$ are graded subspaces that satisfy

$$R_{\ell,k'} \subseteq R_{\ell,k}$$

for $k' \geq k$. These containments induce surjections of graded linear spaces

$$L(\mathcal{G}, k') \to L(\mathcal{G}, k)$$

and this is the statement of Theorem 40. $\square$

B.4: ***Proof of Theorem 41.*** In order to prove Theorem 41 we may assume that $\ell \geq 3$, and we shall do this. Furthermore, by applying Theorem 40 we are reduced to proving Theorem 41 in the case $k = 1$, and we shall from now on also assume that this is the case.

Thus we must compare the dimensions of the weight 2 pieces of the VOAs $C_{\ell,1}$ and $B_{\ell,1}$. According to Proposition 43 these are given by the weight 2 pieces of the graded quotients

$$C_{\ell,1} = V(C_\ell, 1)/\mathcal{U}(\widehat{\mathcal{C}_\ell})e_{\theta C}(-1)^2\mathbf{1},$$
$$B_{\ell,1} = V(B_\ell, 1)/\mathcal{U}(\widehat{\mathcal{B}_\ell})e_{\theta B}(-1)^2\mathbf{1},$$

where θB and θC are the highest roots for the root systems of type B_ℓ and C_ℓ, respectively.

From the description of the underlying $\mathbb{Z}$-graded space of $V(\mathcal{G}, 1)$ as a graded symmetric algebra presented in Section B.2, and because B_ℓ, C_ℓ are Lie algebras of equal dimension, it follows that the degree 2 pieces of $V(B_\ell, 1)$ and $V(C_\ell, 1)$ are also equal. Therefore, in order to prove Theorem 41 we must compare the dimensions of the degree 2 pieces $(\mathcal{U}(\widehat{\mathcal{B}_\ell})e_{\theta B}(-1)^2\mathbf{1})_2$ and $(\mathcal{U}(\widehat{\mathcal{C}_\ell})e_{\theta C}(-1)^2\mathbf{1})_2$. Indeed, we shall prove the next result (and Remark 42 also follows from this):

Lemma 44. *We have*

$$\dim(\mathcal{U}(\widehat{\mathcal{B}_\ell})e_{\theta B}(-1)^2\mathbf{1})_2 - \dim(\mathcal{U}(\widehat{\mathcal{C}_\ell})e_{\theta C}(-1)^2\mathbf{1})_2 = 2\ell - 4.$$

The remainder of this appendix proceeds with the proof of this lemma. It amounts to a fairly elaborate computation of the dimensions of the 2 graded spaces in question.

We begin with any simple Lie algebra $\mathcal{G}$. $\mathcal{U}(\widehat{\mathcal{G}})$ is spanned by elements of the form

$$\{a^1(-m_1)\cdots a^r(-m_r)b^1(0)\cdots b^s(0)c^1(n_1)\cdots c^t(n_t) \mid m_i, n_i \geq 1\},$$

where the Lie algebra elements a^i, b^i, c^i span $\mathcal{G}$, the m_i and n_i are decreasing sequences of integers, $c^i(n_i)$ is the operator induced by $c^i \otimes t^{n_i}$, etc. Now it is well-known that the radical spaces $(\mathcal{U}(\widehat{\mathcal{G}})e_\theta(-1)^2\mathbf{1})$

contain *no* nonzero elements of degree less than 2. Thus the weight 2 piece is the lowest nonzero part. Because the operators $c^i(n_i)$ are *lowering* operators for $n_i > 0$ they must annihilate $e_\theta(-1)^2\mathbf{1}$ (a result that can be checked directly). Similarly, the $b^i(0)$ are weight 0 operators and the $a^i(-m_i)$ are *raising* operators for $m_i > 0$. The upshot is that we have

$$(\mathcal{U}(\widehat{\mathcal{G}})e_\theta(-1)^2\mathbf{1})_2 = \widehat{\mathcal{G}}_0 e_\theta(-1)^2\mathbf{1}.$$

For $b \in \mathcal{G}$ we also have

$$b(0)e_\theta(-1)^2\mathbf{1} = 2[b, e_\theta](-1)e_\theta$$

and as b ranges over $\mathcal{G}$ we generate in this way $e_\theta(-1)^2$ as well as $e_\gamma(-1)e_\theta$ for positive roots α, γ such that $\gamma + \alpha = \theta$. Let the number of such positive roots γ be denoted by $N = N_{\mathcal{G}}$. This argument shows that

$$\dim(\mathcal{U}(\widehat{\mathcal{G}})e_\theta(-1)^2\mathbf{1})_2 = 1 + N_{\mathcal{G}}.$$

There is a representation-theoretic meaning of the integer N. Recall the $\mathfrak{sl}_2$ Lie algebra $\mathcal{S} := \mathcal{S}_\theta = \langle e_\theta, f_\theta, h_\theta\rangle \subseteq \mathcal{G}$, and decompose the adjoint representation as a direct sum of irreducible $\mathcal{S}$-modules. We assert that

$$\mathcal{G} = C_{\mathcal{G}}(\mathcal{S}) \oplus \mathcal{S} \oplus_{i=1}^{N} V_i, \tag{25}$$

where $C_{\mathcal{G}}(\mathcal{S})$ is the *centralizer* of $\mathcal{S}$ and each V_i is 2-dimensional. Indeed, the 1-dimensional summands are all contained in the centralizer and there is at least one 3-dimensional summand, namely, $\mathcal{S}$ itself. Note that a Cartan subalgebra $\mathcal{H}$ is contained in the sum of these two modules. Let $V_i \subseteq \mathcal{G}$ be any other nonzero irreducible $\mathcal{S}$-submodule. On one hand V_i is spanned by root vectors because $\mathcal{G}$ is, and on the other hand it contains a unique highest weight vector for e_θ. Because θ is the highest root for $\mathcal{G}$ then *every* root vector v_γ ($\gamma \in \Phi^+$) is annihilated by e_θ, and this means that V_i contains *exactly one* positive root vector, call it v_γ, and exactly one negative root vector, which must be $v_{\gamma-\theta}$. Setting $\beta := \theta - \gamma \in \Phi^+$ we have $\alpha + \beta = \theta$.

This argument shows that $\dim V_i = 2$, thereby confirming the decomposition (25). Note that we obtain such a V_i whenever $\theta = \alpha + \beta$ is decomposed into a sum of two positive roots, so that the number of 2-dimensional summands in (25) is indeed equal to N.

We now find that

$$N_{\mathcal{G}} = \tfrac{1}{2}(\dim \mathcal{G} - \dim C_{\mathcal{G}}(\mathcal{S}) - 3). \tag{26}$$

Lemma 45. (1) *If $\mathcal{G} = B_\ell$ then $C_{\mathcal{G}}(\mathcal{S}) \cong A_1 \oplus B_{\ell-2}$.*

(2) *If $\mathcal{G} = C_\ell$ then $C_{\mathcal{G}}(\mathcal{S}) \cong C_{\ell-1}$.*

Proof. We tackle the case C_ℓ first. In standard notation (see [Humphreys 1972, Section 12]) we choose an orthonormal basis $\{e_i\}$ in Euclidean space $\mathbb{R}^\ell$. A root system of type C_ℓ may then be chosen to consist of $\{\pm e_i \pm e_j \mid i \neq j\} \cup \{\pm 2e_i\}$, and we have $\theta = 2e_1$. Then all roots with indices i, j greater than 1 correspond to elements of $C_{\mathcal{G}}(\mathcal{S})$, and these form a root system of type $C_{\ell-1}$. The assertion of the lemma in this case follows immediately.

Similarly, a root system of type B_ℓ may be taken to be $\{\pm e_i \pm e_j \mid i \neq j\} \cup \{\pm e_i\}$ and in this case $\theta = e_1 + e_2$. Here, all roots with indices greater than 2 together with $e_1 - e_2$ correspond to elements in the centralizer, and the conclusion is that $C_{\mathcal{G}}(\mathcal{S}) \cong A_1 \oplus B_{\ell-2}$. □

At last we can compute the needed dimensions using Lemma 45 and (26). We find that

$$N_{B_\ell} = \tfrac{1}{2}((2\ell^2+\ell) - (2(\ell-2)^2 + (\ell-2) + 3) - 3) = 4\ell - 6$$
$$N_{C_\ell} = \tfrac{1}{2}((2\ell^2+\ell) - (2(\ell-1)^2 + (\ell-1)) - 3) = (-(-4\ell + 2 + (-1)) - 3) = 2\ell - 2$$

Therefore $N_{B_\ell} - N_{C_\ell} = 2\ell - 4$, and this completes the proof of Lemma 44 and thereby that of Theorem 41.

Acknowledgements

We are indebted to the following individuals for helpful discussions, supplying references, and for answering our questions: Chongying Dong, Gerald Höhn, Ching Hung Lam, Sunil Mukhi, Kiyokazu Nagatomo, and Ivan Penkov. We also thank the referee for detailed comments.

References

[Arike et al. 2016] Y. Arike, M. Kaneko, K. Nagatomo, and Y. Sakai, "Affine vertex operator algebras and modular linear differential equations", *Lett. Math. Phys.* **106**:5 (2016), 693–718. MR Zbl

[Arike et al. 2017] Y. Arike, K. Nagatomo, and Y. Sakai, "Characterization of the simple Virasoro vertex operator algebras with 2 and 3-dimensional space of characters", pp. 175–204 in *Lie algebras, vertex operator algebras, and related topics*, edited by K. Barron et al., Contemp. Math. **695**, Amer. Math. Soc., Providence, RI, 2017. MR Zbl

[Bennett et al. 2018] M. A. Bennett, G. Martin, K. O'Bryant, and A. Rechnitzer, "Explicit bounds for primes in arithmetic progressions", *Illinois J. Math.* **62**:1-4 (2018), 427–532. MR Zbl

[Beukers and Heckman 1989] F. Beukers and G. Heckman, "Monodromy for the hypergeometric function ${}_nF_{n-1}$", *Invent. Math.* **95**:2 (1989), 325–354. MR Zbl

[Bombieri and Gubler 2006] E. Bombieri and W. Gubler, *Heights in Diophantine geometry*, New Math. Monogr. **4**, Cambridge Univ. Press, 2006. MR Zbl

[Dong and Mason 2004] C. Dong and G. Mason, "Rational vertex operator algebras and the effective central charge", *Int. Math. Res. Not.* **2004**:56 (2004), 2989–3008. MR Zbl

[Dong and Mason 2006] C. Dong and G. Mason, "Integrability of C_2-cofinite vertex operator algebras", *Int. Math. Res. Not.* **2006** (2006), art. id. 80468. MR Zbl

[Dong et al. 1996] C. Dong, H. Li, and G. Mason, "Simple currents and extensions of vertex operator algebras", *Comm. Math. Phys.* **180**:3 (1996), 671–707. MR Zbl

[Dong et al. 2000] C. Dong, H. Li, and G. Mason, "Modular-invariance of trace functions in orbifold theory and generalized Moonshine", *Comm. Math. Phys.* **214**:1 (2000), 1–56. MR Zbl

[Dwork 1990] B. Dwork, *Generalized hypergeometric functions*, Oxford Univ. Press, 1990. MR Zbl

[Elkies 2007] N. D. Elkies, "Three lectures on elliptic surfaces and curves of high rank", preprint, 2007. arXiv

[Franc and Mason 2014] C. Franc and G. Mason, "Fourier coefficients of vector-valued modular forms of dimension 2", *Canad. Math. Bull.* **57**:3 (2014), 485–494. MR Zbl

[Franc and Mason 2016a] C. Franc and G. Mason, "Hypergeometric series, modular linear differential equations and vector-valued modular forms", *Ramanujan J.* **41**:1-3 (2016), 233–267. MR Zbl

[Franc and Mason 2016b] C. Franc and G. Mason, "Three-dimensional imprimitive representations of the modular group and their associated modular forms", *J. Number Theory* **160** (2016), 186–214. MR Zbl

[Franc et al. 2018] C. Franc, T. Gannon, and G. Mason, "On unbounded denominators and hypergeometric series", *J. Number Theory* **192** (2018), 197–220. MR Zbl

[Gaberdiel et al. 2016] M. R. Gaberdiel, H. R. Hampapura, and S. Mukhi, "Cosets of meromorphic CFTs and modular differential equations", *J. High Energy Phys.* **2016**:4 (2016), art. id. 156. Zbl

[Grady and Tener 2018] J. C. Grady and J. E. Tener, "Classification of extremal vertex operator algebras with two simple modules", 2018. To appear in *J. Math. Phys.* arXiv

[Hampapura and Mukhi 2016] H. R. Hampapura and S. Mukhi, "Two-dimensional RCFT's without Kac–Moody symmetry", *J. High Energy Phys.* **2016**:7 (2016), art. id. 138. MR Zbl

[Höhn 1996] G. Höhn, *Selbstduale Vertexoperatorsuperalgebren und das Babymonster*, Bonner Math. Schriften **286**, Universität Bonn, 1996. MR Zbl

[Huang 2008] Y.-Z. Huang, "Vertex operator algebras and the Verlinde conjecture", *Commun. Contemp. Math.* **10**:1 (2008), 103–154. MR Zbl

[Humphreys 1972] J. E. Humphreys, *Introduction to Lie algebras and representation theory*, Grad. Texts in Math. **9**, Springer, 1972. MR Zbl

[Kac 1990] V. G. Kac, *Infinite-dimensional Lie algebras*, 3rd ed., Cambridge Univ. Press, 1990. MR Zbl

[Knopp et al. 1965] M. I. Knopp, J. Lehner, and M. Newman, "A bounded automorphic form of dimension zero is constant", *Duke Math. J.* **32** (1965), 457–460. MR Zbl

[Lam and Shimakura 2019] C. H. Lam and H. Shimakura, "71 holomorphic vertex operator algebras of central charge 24", *Bull. Inst. Math. Acad. Sin.* (*N.S.*) **14**:1 (2019), 87–118. MR Zbl

[Lepowsky and Li 2004] J. Lepowsky and H. Li, *Introduction to vertex operator algebras and their representations*, Progr. Math. **227**, Birkhäuser, Boston, 2004. MR Zbl

[Li 1994] H. Li, "Symmetric invariant bilinear forms on vertex operator algebras", *J. Pure Appl. Algebra* **96**:3 (1994), 279–297. MR Zbl

[Lin 2017] X. Lin, "Mirror extensions of rational vertex operator algebras", *Trans. Amer. Math. Soc.* **369**:6 (2017), 3821–3840. MR Zbl

[Mason 2014] G. Mason, "Lattice subalgebras of strongly regular vertex operator algebras", pp. 31–53 in *Conformal field theory, automorphic forms and related topics* (Heidelberg, 2011), edited by W. Kohnen and R. Weissauer, Contrib. Math. Comput. Sci. **8**, Springer, 2014. MR Zbl

[Mason 2020] G. Mason, "Five not-so-easy pieces: open problems about vertex rings", pp. 213–232 in *Vertex operator algebras, number theory and related topics*, edited by M. Krauel et al., Contemp. Math. **753**, Amer. Math. Soc., Providence, RI, 2020.

[Mason et al. 2018] G. Mason, K. Nagatomo, and Y. Sakai, "Vertex operator algebras with two simple modules: the Mathur–Mukhi–Sen theorem revisited", preprint, 2018. arXiv

[Mathur et al. 1988] S. D. Mathur, S. Mukhi, and A. Sen, "On the classification of rational conformal field theories", *Phys. Lett. B* **213**:3 (1988), 303–308. MR

[Mathur et al. 1989] S. D. Mathur, S. Mukhi, and A. Sen, "Reconstruction of conformal field theories from modular geometry on the torus", *Nuclear Phys. B* **318**:2 (1989), 483–540. MR

[Moore and Seiberg 1989] G. Moore and N. Seiberg, "Classical and quantum conformal field theory", *Comm. Math. Phys.* **123**:2 (1989), 177–254. MR Zbl

[Schellekens 1993] A. N. Schellekens, "Meromorphic $c = 24$ conformal field theories", *Comm. Math. Phys.* **153**:1 (1993), 159–185. MR Zbl

[Tuba and Wenzl 2001] I. Tuba and H. Wenzl, "Representations of the braid group B_3 and of $\mathrm{SL}(2, \mathbb{Z})$", *Pacific J. Math.* **197**:2 (2001), 491–510. MR Zbl

[Zhu 1996] Y. Zhu, "Modular invariance of characters of vertex operator algebras", *J. Amer. Math. Soc.* **9**:1 (1996), 237–302. MR Zbl

Communicated by Susan Montgomery
Received 2019-05-28 Revised 2019-11-05 Accepted 2020-02-10

franc@math.usask.ca *University of Saskatchewan, Saskatoon, SK, Canada*
gem@ucsc.edu *University of California at Santa Cruz, Santa Cruz, CA, 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 14 No. 6 2020