Algebra & Number Theory

Volume 14
2020
No. 10

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 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2575

Arithmetic of curves on moduli of local systems

Junho Peter Whang

We investigate the arithmetic of algebraic curves on coarse moduli spaces for special linear rank two local systems on surfaces with fixed boundary traces. We prove a structure theorem for morphisms from the affine line into the moduli space. We show that the set of integral points on any nondegenerate algebraic curve on the moduli space can be effectively determined.

1. Introduction

1A. This is a continuation of our Diophantine study [Whang 2017] of moduli spaces for local systems on surfaces and their mapping class group dynamics. Let Σ be a smooth compact oriented surface of genus g with n boundary curves satisfying $3g+n-3>0$. Let X_k be the coarse moduli space of $\mathrm{SL}_2(\mathbb{C})$-local systems on Σ with prescribed boundary traces $k \in \mathbb{A}^n(\mathbb{C})$. It is an irreducible complex affine algebraic variety of dimension $6g+2n-6$, and we showed in [Whang 2020] that it is log Calabi-Yau if the surface has nonempty boundary. For $k \in \mathbb{A}^n(\mathbb{Z})$, the variety X_k admits a natural model over $\mathbb{Z}$. The mapping class group Γ of the surface acts on X_k via pullback of local systems, and an associated theory of descent on the integral points $X_k(\mathbb{Z})$ was developed in [Whang 2017]. In this paper, we investigate the interplay between the dynamics of this action and the Diophantine geometry of algebraic curves on X_k.

1B. ***Main results.*** We describe the contents of this paper. Relevant background on surfaces and their moduli of local systems is given in Section 2, where we repeat material from [Whang 2017, Section 2]. As in [Whang 2017], let us say that a simple closed curve on Σ is *essential* if it cannot be continuously deformed into a point or a boundary curve on Σ. This paper is devoted to developing consequences of the following boundedness theorem [Whang 2017, Theorem 3] for nonarchimedean systoles of local systems.

Theorem [Whang 2017]. *Let $\mathcal{O}$ be a discrete valuation ring with fraction field F. Given any representation $\rho : \pi_1\Sigma \to \mathrm{SL}_2(F)$ whose boundary traces all take values in $\mathcal{O}$, there is an essential simple closed curve $a \subset \Sigma$ with $\operatorname{tr}\rho(a) \in \mathcal{O}$.*

In Section 3, we apply the above theorem to the field of rational functions and prove our first main result, which is a structure theorem for morphisms from the affine line $\mathbb{A}^1$ into the moduli space X_k. Following [Whang 2017], let us say that a possibly reducible algebraic variety Z is *parabolic* if it is covered by nonconstant morphisms $\mathbb{A}^1 \to Z$. We also define a subvariety of X_k to be *degenerate* if it is

MSC2010: primary 11G30; secondary 11G35, 57M50.
Keywords: surfaces, character variety, algebraic curves, integral points.

contained in a parabolic subvariety of X_k, and *nondegenerate* otherwise. The following theorem gives a modular characterization of the degenerate points of X_k.

Theorem 1.1. *A point $\rho \in X_k(\mathbb{C})$ is degenerate if and only if*

(1) (parabolic curve) *there is an essential simple closed curve $a \subset \Sigma$ such that* $\operatorname{tr}\rho(a) = \pm 2$*, or*

(2) (parabolic pants) $(g, n, k) \neq (1, 1, 2)$ *and there is a subsurface $\Sigma' \subset \Sigma$ of genus 0 with 3 boundary curves, each of which is an essential curve or a boundary curve of Σ, such that the restriction $\rho|\Sigma'$ is reducible.*

In particular, there is a parabolic proper closed subvariety Z of X_k such that every nonconstant morphism $\mathbb{A}^1 \to X_k$ over $\mathbb{C}$ is mapping class group equivalent to one with image in Z.

Theorem 1.1 is reminiscent of a result of Sterk [1985] that the automorphism group of a projective K3 surface acts on the set of its smooth rational curve classes with finitely many orbits. It also has an interesting consequence (Corollary 3.6) that any polynomial deformation of a Fuchsian representation of surface group preserving the boundary traces must be isotrivial. Finally, Theorem 1.1 is used in formulating the main Diophantine result of [Whang 2017] for the integral points of X_k.

In Section 4, we study the behavior of integral points on algebraic curves in X_k. For each curve $a \subset \Sigma$, let tr_a be the regular function on X_k given by monodromy trace of local systems along a. We define an algebraic curve $C \subset X_k$ to be *integrable* if there is a pants decomposition P of Σ (i.e., a maximal union of pairwise disjoint and nonisotopic essential simple closed curves) such that tr_a is constant on C for every curve $a \subset P$. Otherwise, C is *nonintegrable*. Given an algebraic curve $C \subset X_k$ and an arbitrary subset $A \subseteq \mathbb{C}$, let us denote

$$C(A) = \{\rho \in V(\mathbb{C}) : \operatorname{tr}_a(\rho) \in A \text{ for every essential simple closed curve } a \subset \Sigma\}.$$

We prove the following result, by applying the boundedness of nonarchimedean systoles on local systems to function fields of algebraic curves.

Theorem 1.2. *If $C \subset X_k$ is a geometrically irreducible nonintegrable algebraic curve, then $C(A)$ is finite for any closed discrete set $A \subset \mathbb{C}$.*

Moreover, our method will show that, given an embedding of C into affine space, the sizes of the coordinates of $C(A)$ from the theorem can be effectively determined. One application is the following. For each positive squarefree integer d, let $O_d \subset \mathbb{C}$ denote the ring of integers of the imaginary quadratic field $\mathbb{Q}(\sqrt{-d})$. Applying Theorem 1.2 with $A = \bigcup_{d>0} O_d$, we conclude that a nonintegrable curve in X_k has at most finitely many imaginary quadratic integral points. As a special case, this recovers the finiteness result of Long and Reid [2003] for imaginary quadratic integral points on character curves of one-cusped hyperbolic three-manifolds; see Section 4 for details. Our approach to finiteness of integral points on nonintegrable curves shares its basis with the so-called Runge's method, described in [Zannier 2009].

By combining Theorem 1.2 with an analysis of integrable algebraic curves using Baker's theory on linear forms in logarithms, we also obtain the following result in Section 4. Let us define an element in

the mapping class group of Σ to be a *multitwist* if it is given by a product of commuting Dehn twists (and their powers) along essential curves in a pants decomposition of Σ. By a 1-dimensional algebraic torus we shall mean an irreducible algebraic curve of genus 0 with 2 punctures.

Theorem 1.3. *Let $C \subset X_k$ be a geometrically irreducible nondegenerate algebraic curve over $\mathbb{Z}$. Then $C(\mathbb{Z})$ can be effectively determined, and*

(1) *$C(\mathbb{Z})$ is finite, or*

(2) *C is the image of a 1-dimensional algebraic torus preserved by a nontrivial multitwist, under which $C(\mathbb{Z})$ consists of finitely many orbits.*

If moreover C is not fixed pointwise by any nontrivial multitwist, the same result holds with $C(\mathbb{Z})$ replaced by the set of all imaginary quadratic integral points on C.

Theorem 1.3 gives a complete analysis of integral points on every nondegenerately embedded curve $C \subset X_k$ with no intrinsic restrictions, e.g., regarding the genus and number of punctures of the curve. The structure of Theorem 1.3 is strongly reminiscent of, and motivated by, classical Diophantine results on subvarieties of log Calabi-Yau varieties of linear type, such as algebraic tori and abelian varieties (cf. Skolem's approach to the Thue equations [Borevich and Shafarevich 1966], as well as [Vojta 1991; 1996; Faltings 1991]). Finally, for $(g, n) = (1, 1)$ or $(0, 4)$, the moduli space X_k has an explicit presentation as an affine cubic algebraic surface with equation of the form

$$x^2 + y^2 + z^2 + xyz = ax + by + cz + d \qquad (*)$$

for some constants a, b, c, d depending on k. Affine algebraic surfaces of this type were first studied in [Markoff 1880], which introduced a form of nonlinear descent which essentially coincides the mapping class group action. Our work therefore specializes to the following result, which may be proved elementarily (but still using the group action).

Corollary 1.4. *On an affine algebraic surface with an equation of the form $(*)$, the integral solutions to any Diophantine equation over $\mathbb{Z}$ can be effectively determined.*

Ghosh and Sarnak [2017] showed that, in the sense of proportions, almost all "admissible" Markoff type surfaces X_k for $(g, n) = (1, 1)$ have a Zariski dense set of integral points. Thus, Corollary 1.4 provides an infinite family of nontrivial ambient varieties of dimension two where every Diophantine equation over $\mathbb{Z}$ can be effectively solved.

2. Background

This section collects relevant background on surfaces and their moduli of local systems, repeating material from [Whang 2017, Section 2]. We also recall a boundedness result [Whang 2017, Theorem 1.3] for nonarchimedean systoles of local systems in Section 2D, which will prove instrumental in our Diophantine analysis.

2A. ***Surfaces.*** A *surface* is an oriented two-dimensional smooth manifold, which we assume to be compact with at most finitely many boundary components unless otherwise indicated. A connected surface is said to have type (g, n) if it has genus g and has n boundary components. A *curve* on a surface is an embedded copy of an unoriented circle, which we shall tacitly assume to be smooth in appropriate contexts. Given a surface Σ, we shall say that a curve $a \subset \Sigma$ is *nondegenerate* if it does not bound a disk, and *essential* if it is nondegenerate and is disjoint from, and not isotopic to, a boundary curve of Σ.

A *multicurve* on Σ is a finite union of disjoint curves on Σ. It is said to be *nondegenerate*, resp. *essential*, if each of its components is. Given a surface Σ and an essential multicurve $Q \subset \Sigma$, we denote by $\Sigma|Q$ the surface obtained by cutting Σ along the curves in Q. A *pants decomposition* of Σ is an essential multicurve P such that $\Sigma|P$ is a disjoint union of surfaces of type $(0, 3)$. Equivalently, a pants decomposition is a maximal (with respect to inclusion) essential multicurve whose components are pairwise nonisotopic. If Σ is a surface of type (g, n) with $3g + n - 3 > 0$, then any pants decomposition of Σ consists of $3g + n - 3$ essential curves. An essential curve $a \subset \Sigma$ is *separating* if the two boundary curves of $\Sigma|a$ corresponding to a are on different connected components, and *nonseparating* otherwise.

2A1. *Optimal generators.* Let Σ be a surface of type (g, n), and choose a base point $x \in \Sigma$. We have the *standard presentation* of the fundamental group

$$\pi_1(\Sigma, x) = \langle \alpha_1, \beta_1', \ldots, \alpha_g, \beta_g', \gamma_1, \ldots, \gamma_n | [\alpha_1, \beta_1'] \cdots [\alpha_g, \beta_g'] \gamma_1 \cdots \gamma_n \rangle, \tag{1}$$

where in particular $\gamma_1, \ldots, \gamma_n$ correspond to loops around the boundary curves of Σ. For $i = 1, \ldots, g$, let β_i be the based loop traversing β_i' in the opposite direction. We can choose the sequence of generating loops $(\alpha_1, \beta_1, \ldots, \alpha_g, \beta_g, \gamma_1, \ldots, \gamma_n)$ so that it satisfies the following:

(1) Each loop in the sequence is simple.

(2) Any two distinct loops in the sequence intersect exactly once (at x).

(3) Every product of distinct elements in the sequence preserving the cyclic ordering can be represented by a simple loop in Σ.

Some examples of products alluded to in (3) are $\alpha_1\beta_g$, $\alpha_1\alpha_2\beta_2\beta_g$, and $\beta_g\gamma_n\alpha_1$. We refer to the sequence $(\alpha_1, \beta_1, \ldots, \alpha_g, \beta_g, \gamma_1, \ldots, \gamma_n)$ as an *optimal sequence of generators* for $\pi_1\Sigma$. See Figure 1 for an illustration of optimal generators for $(g, n) = (2, 1)$.

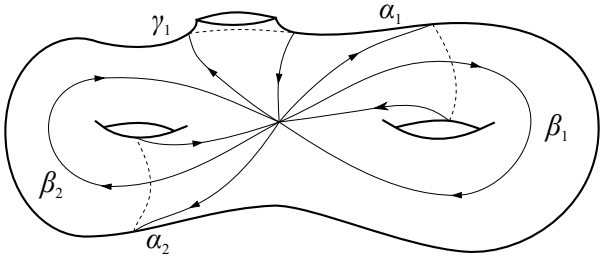


Figure 1. Optimal generators for $(g, n) = (2, 1)$.

2A2. *Mapping class group.* Given a surface Σ, let $\Gamma = \Gamma(\Sigma) = \pi_0 \operatorname{Diff}^+(\Sigma, \partial\Sigma)$ denote its mapping class group. By definition, it is the group of isotopy classes of orientation preserving diffeomorphisms of Σ fixing the boundary of Σ pointwise. Given a (simple closed) curve $a \subset \Sigma$ disjoint from $\partial\Sigma$, the associated (left) *Dehn twist* $\tau_a \in \Gamma$ on Σ is defined as follows. Let $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ be the unit circle. Let τ be the diffeomorphism from $S^1 \times [0, 1]$ to itself given by $(z, t) \mapsto (ze^{2\pi i\xi(t)}, t)$ where $\xi(t)$ is a smooth bump function of $t \in [0, 1]$ that is 0 on a neighborhood of 0 and 1 on a neighborhood of 1. Choose a closed tubular neighborhood N of a in Σ, and an orientation preserving diffeomorphism $f : N \to S^1 \times [0, 1]$. The Dehn twist τ_a is given by

$$\tau_a(x) = \begin{cases} f^{-1} \circ \tau \circ f(x) & \text{if } x \in N, \\ x & \text{otherwise.} \end{cases}$$

The class of τ_a in Γ is independent of the choices involved above, and depends only on the isotopy class of a. It is a standard fact that $\Gamma = \Gamma(\Sigma)$ is generated by Dehn twists along simple closed curves in Σ (see [Farb and Margalit 2012, Chapter 4]).

2B. ***Character varieties.*** Throughout this paper, an *algebraic variety* is a scheme of finite type over a field. Given an affine variety V over a given field k, we denote by $k[V]$ its coordinate ring over k. If moreover V is integral, then $k(V)$ denotes its function field over k. Given a commutative ring A with unity, the elements of A will be referred to as *regular functions* on the affine scheme $\operatorname{Spec} A$.

2B1. *Character varieties of groups.* Let π be a finitely generated group. Its (SL_2) *representation variety* $\operatorname{Rep}(\pi)$ is the affine scheme determined by the functor

$$A \mapsto \operatorname{Hom}(\pi, \mathrm{SL}_2(A))$$

for every commutative ring A. Given a sequence of generators of π with m elements, we have a presentation of $\operatorname{Rep}(\pi)$ as a closed subscheme of SL_2^m defined by equations coming from relations among the generators. For each $a \in \pi$, let tr_a be the regular function on $\operatorname{Rep}(\pi)$ given by $\rho \mapsto \operatorname{tr} \rho(a)$.

The (SL_2) *character variety* of π over $\mathbb{C}$ is the affine invariant theoretic quotient

$$X(\pi) = \operatorname{Rep}(\pi) /\!\!/ \mathrm{SL}_2 = \operatorname{Spec} \mathbb{C}[\operatorname{Rep}(\pi)]^{\mathrm{SL}_2(\mathbb{C})}$$

under the simultaneous conjugation action of SL_2. Note that the regular function tr_a for each $a \in \pi$ descends to a regular function on $X(\pi)$. Moreover, $X(\pi)$ has a natural model over $\mathbb{Z}$, defined as the spectrum of

$$R(\pi) = \mathbb{Z}[\mathrm{tr}_a : a \in \pi]/(\mathrm{tr}_1 - 2, \mathrm{tr}_a \mathrm{tr}_b - \mathrm{tr}_{ab} - \mathrm{tr}_{ab^{-1}}).$$

The relations in the above presentation arise from the fact that the trace of the 2×2 identity matrix is 2, and $\operatorname{tr}(A)\operatorname{tr}(B) = \operatorname{tr}(AB) + \operatorname{tr}(AB^{-1})$ for every $A, B \in \mathrm{SL}_2(\mathbb{C})$.

Given an integral domain A with fraction field F of characteristic zero, the A-points of $X(\pi)$ parametrize the Jordan equivalence classes of $\mathrm{SL}_2(\bar{F})$-representations of π having character valued in A (see [Simpson 1994, Proposition 6.1]). Here, following [Simpson 1994], we say that two finite-dimensional

linear representations of π are Jordan equivalent if they admit composition series with isomorphic graded representations. Since a semisimple finite-dimensional representation of a group over a field of characteristic zero is determined by its character [Lang 2002, Chapter XVII, Section 3, Corollary 3.8], we see in particular two representations $\rho : \pi \to \mathrm{SL}_2(\mathbb{C})$ are Jordan equivalent if and only if they have the same character. (It is not true in general that, for a reductive algebraic group $G \leq \mathrm{GL}_r$ over $\mathbb{C}$, the points of $\mathrm{Hom}(\pi, G) /\!\!/ G$ are determined by their characters.) We refer to [Horowitz 1972; Przytycki and Sikora 2000; Saito 1996] for further details on SL_2-character varieties.

Example 2.1. We refer to [Goldman 2009] for details of examples below. Let F_m denote the free group on $m \geq 1$ generators $a_1, \ldots, a_m$.

(1) We have $\mathrm{tr}_{a_1} : X(F_1) \simeq \mathbb{A}^1$.

(2) We have $(\mathrm{tr}_{a_1}, \mathrm{tr}_{a_2}, \mathrm{tr}_{a_1a_2}) : X(F_2) \simeq \mathbb{A}^3$ by Fricke [Goldman 2009, Section 2.2].

(3) The coordinate ring $\mathbb{Q}[X(F_3)]$ is the quotient of the polynomial ring

$$\mathbb{Q}[\mathrm{tr}_{a_1}, \mathrm{tr}_{a_2}, \mathrm{tr}_{a_3}, \mathrm{tr}_{a_1a_2}, \mathrm{tr}_{a_2a_3}, \mathrm{tr}_{a_1a_3}, \mathrm{tr}_{a_1a_2a_3}, \mathrm{tr}_{a_1a_3a_2}]$$

by the ideal generated by two elements

$$\mathrm{tr}_{a_1a_2a_3} + \mathrm{tr}_{a_1a_3a_2} - (\mathrm{tr}_{a_1a_2}\,\mathrm{tr}_{a_3} + \mathrm{tr}_{a_1a_3}\,\mathrm{tr}_{a_2} + \mathrm{tr}_{a_2a_3}\,\mathrm{tr}_{a_1} - \mathrm{tr}_{a_1}\,\mathrm{tr}_{a_2}\,\mathrm{tr}_{a_3})$$

and

$$\mathrm{tr}_{a_1a_2a_3}\,\mathrm{tr}_{a_1a_3a_2} - \big\{(\mathrm{tr}_{a_1}^2 + \mathrm{tr}_{a_2}^2 + \mathrm{tr}_{a_3}^2) + (\mathrm{tr}_{a_1a_2}^2 + \mathrm{tr}_{a_2a_3}^2 + \mathrm{tr}_{a_1a_3}^2) - (\mathrm{tr}_{a_1}\,\mathrm{tr}_{a_2}\,\mathrm{tr}_{a_1a_2} + \mathrm{tr}_{a_2}\,\mathrm{tr}_{a_3}\,\mathrm{tr}_{a_2a_3} + \mathrm{tr}_{a_1}\,\mathrm{tr}_{a_3}\,\mathrm{tr}_{a_1a_3}) + \mathrm{tr}_{a_1a_2}\,\mathrm{tr}_{a_2a_3}\,\mathrm{tr}_{a_1a_3} - 4\big\}.$$

We record the following, which is attributed by Goldman [2009] to Vogt [1889].

Lemma 2.2. *Given a finitely generated group π and $a_1, a_2, a_3, a_4 \in \pi$, we have*

$$\begin{aligned}2\mathrm{tr}_{a_1a_2a_3a_4} = {} & \mathrm{tr}_{a_1}\mathrm{tr}_{a_2}\mathrm{tr}_{a_3}\mathrm{tr}_{a_4} + \mathrm{tr}_{a_1}\mathrm{tr}_{a_2a_3a_4} + \mathrm{tr}_{a_2}\mathrm{tr}_{a_3a_4a_1} + \mathrm{tr}_{a_3}\mathrm{tr}_{a_4a_1a_2} \\ & + \mathrm{tr}_{a_4}\mathrm{tr}_{a_1a_2a_3} + \mathrm{tr}_{a_1a_2}\mathrm{tr}_{a_3a_4} + \mathrm{tr}_{a_4a_1}\mathrm{tr}_{a_2a_3} - \mathrm{tr}_{a_1a_3}\mathrm{tr}_{a_2a_4} \\ & - \mathrm{tr}_{a_1}\mathrm{tr}_{a_2}\mathrm{tr}_{a_3a_4} - \mathrm{tr}_{a_3}\mathrm{tr}_{a_4}\mathrm{tr}_{a_1a_2} - \mathrm{tr}_{a_4}\mathrm{tr}_{a_1}\mathrm{tr}_{a_2a_3} - \mathrm{tr}_{a_2}\mathrm{tr}_{a_3}\mathrm{tr}_{a_4a_1}.\end{aligned}$$

The above computation implies the following fact, which forms a special case of Procesi's theorem [1976] that rings of invariants of tuples of $N \times N$ matrices under simultaneous conjugation are (finitely) generated by the trace functions of products of matrices.

Fact 2.3. *If π is a group generated by $a_1, \ldots, a_m$, then $\mathbb{Q}[X(\pi)]$ is generated as a $\mathbb{Q}$-algebra by the collection $\{\mathrm{tr}_{a_{i_1}\cdots a_{i_k}} : 1 \leq i_1 < \cdots < i_k \leq m\}_{1 \leq k \leq 3}$.*

2B2. *Moduli of local systems on manifolds.* Given a connected smooth (compact) manifold M, the coarse moduli space of local systems on M that we shall study is the character variety $X(M) = X(\pi_1 M)$ of its fundamental group. The complex points of $X(M)$ parametrize the Jordan equivalence classes of $\mathrm{SL}_2(\mathbb{C})$-local systems on M. More generally, given a smooth manifold $M = M_1 \sqcup \cdots \sqcup M_m$ with finitely many connected components M_i, we define

$$X(M) = X(M_1) \times \cdots \times X(M_m).$$

The construction of the moduli space $X(M)$ is functorial in the manifold M. Any smooth map $f : M \to N$ of manifolds induces a morphism $f^* : X(N) \to X(M)$, depending only on the homotopy class of f, given by pullback of local systems.

Let Σ be a surface. For each curve $a \subset \Sigma$, there is a well-defined regular function $\mathrm{tr}_a : X(\Sigma) \to X(a) \simeq \mathbb{A}^1$, which agrees with tr_α for any $\alpha \in \pi_1 \Sigma$ represented by a path freely homotopic to a parametrization of a. Implicit here is the observation that tr_α is independent of the choice of an orientation for a since $\mathrm{tr}(A) = \mathrm{tr}(A^{-1})$ for any matrix A in SL_2. The boundary curves $\partial\Sigma$ of Σ induce a natural morphism

$$\mathrm{tr}_{\partial\Sigma} : X(\Sigma) \to X(\partial\Sigma) \simeq \mathbb{A}^n,$$

where the latter isomorphism is given by a choice of ordering $\partial\Sigma = c_1 \sqcup \cdots \sqcup c_n$ of the boundary curves c_i. The fibers of $\mathrm{tr}_{\partial\Sigma}$ for $k \in \mathbb{A}^n$ will be denoted $X_k = X_k(\Sigma)$. Each X_k is often referred to as a *relative character variety* in the literature. If Σ is a surface of type (g, n) satisfying $3g + n - 3 > 0$, the relative character variety $X_k(\Sigma)$ is an irreducible algebraic variety of dimension $6g + 2n - 6$.

We shall often simplify our notation by combining parentheses where applicable, e.g., $X_k(\Sigma, \mathbb{Z}) = X_k(\Sigma)(\mathbb{Z})$. Given a fixed surface Σ, a subset $K \subseteq X(\partial\Sigma, \mathbb{C})$, and a subset $A \subseteq \mathbb{C}$, we shall denote by

$$X_K(A) = X_K(\Sigma, A)$$

the set of all $\rho \in X(\Sigma, \mathbb{C})$ such that $\mathrm{tr}_{\partial\Sigma}(\rho) \in K$ and $\mathrm{tr}_a(\rho) \in A$ for every essential curve $a \subset \Sigma$. The following lemma shows that there is no risk of ambiguity with this notation.

Lemma 2.4. *If A is a subring of $\mathbb{C}$ and $k \in A^n$, then X_k has a model over A and $X_k(A)$ recovers the set of A-valued points of X_k in the sense of algebraic geometry.*

Proof. Let A and $k \in A^n$ be as above. We have a model of X_k over A with coordinate ring $\mathrm{Spec}\, R(\pi_1\Sigma) \otimes_{\mathbb{Z}} A$. It is clear that an A-valued point in the sense of algebraic geometry corresponds to a point in $X_k(A)$. The converse follows from the observation, using the identity $\mathrm{tr}_a \mathrm{tr}_b = \mathrm{tr}_{ab} + \mathrm{tr}_{ab^{-1}}$, that tr_b for every $b \in \pi_1\Sigma$ can be written as a $\mathbb{Z}$-linear combination of products of traces tr_a for nondegenerate curves $a \subset \Sigma$. □

Similarly, given $k \in X(\partial\Sigma, \mathbb{C})$, a subvariety $V \subseteq X_k(\Sigma)$, and a subset $A \subseteq \mathbb{C}$, we shall denote

$$V(A) = \{\rho \in V(\mathbb{C}) : \mathrm{tr}\, \rho(a) \in A \text{ for every essential curve } a \subset \Sigma\}.$$

Given an immersion $\Sigma' \to \Sigma$ of surfaces, we have the associated restriction

$$(-)|_{\Sigma'} : X(\Sigma) \to X(\Sigma').$$

The mapping class group $\Gamma(\Sigma)$ acts naturally on $X(\Sigma)$ by pullback of local systems, preserving the integral structure as well as each relative character variety $X_k(\Sigma)$ and the sets $X_K(\Sigma, A)$ defined above. The dynamical aspects of this action on the complex points of $X(\Sigma)$ are not fully understood, but have been studied on certain special subloci. These include the locus of SU(2)-local systems (see [Goldman 1997]) on X, and the Teichmüller locus parametrizing *Fuchsian representations* associated to marked hyperbolic structures on Σ with geodesic boundary. This paper is largely concerned with the descent properties of the dynamics on $X(\mathbb{C})$ beyond the classical setting.

2B3. *Reconstruction.* Let Σ be a surface of type (g, n) with $3g+n-3>0$, and let $a \subset \Sigma$ be an essential curve. Let $x \in \Sigma$ be a base point lying on a, and let α be a simple based loop parametrizing a. We shall summarize the reconstruction of a representation $\rho : \pi_1(\Sigma, x) \to \mathrm{SL}_2(\mathbb{C})$ from representations on connected components of $\Sigma|a$, as well as associated lifts of Dehn twists. Our main reference is [Goldman and Xia 2011]. There are two cases to consider, according to whether a is separating or nonseparating.

Nonseparating curves. Suppose that a is nonseparating, so $\Sigma|a$ is connected. Let a_1 and a_2 be the boundary curves of $\Sigma|a$ corresponding to a, and let (x_i, α_i) be the lifts of (x, α) to each a_i. We shall assume that we have chosen the numberings so that the interior of $\Sigma|a$ lies to the left as one travels along α_1. Let β be a simple loop on Σ based at x, intersecting the curve a once transversely at the base point, such that β lifts to a path β' in $\Sigma|a$ from x_2 to x_1. Let us denote by α_2' the loop based at x_1 given by the path $\alpha_2' = (\beta')^{-1}\alpha_2\beta'$, where $(\beta')^{-1}$ refers to the path β' traversed in the opposite direction. The immersion $\Sigma|a \to \Sigma$ induces an embedding $\pi_1(\Sigma|a, x_1) \to \pi_1(\Sigma, x)$, giving us the isomorphism

$$\pi_1(\Sigma, x) = (\pi_1(\Sigma|a, x_1) \vee \langle\beta\rangle)/(\alpha_2' = \beta^{-1}\alpha_1\beta).$$

Thus, any representation $\rho : \pi_1(\Sigma, x) \to \mathrm{SL}_2(\mathbb{C})$ is determined uniquely by a pair (ρ', B), where $\rho' : \pi_1(\Sigma|a, x_1) \to \mathrm{SL}_2(\mathbb{C})$ is a representation and $B \in \mathrm{SL}_2(\mathbb{C})$ is an element such that $\rho'(\alpha_2') = B^{-1}\rho'(\alpha_1)B$, with the correspondence

$$\rho \mapsto (\rho', B) = (\rho|_{\pi_1(\Sigma|a, x_1)}, \rho(\beta)).$$

We define an automorphism τ_α of $\mathrm{Hom}(\pi_1(\Sigma, x), \mathrm{SL}_2)$ as follows. Given $\rho = (\rho_a, B)$, we set $\tau_\alpha(\rho_a, B) = (\rho_a, B')$ where $B' = \rho(\alpha)B$. This descends to the action τ_a of the left Dehn twist action along a on the moduli space $X(\Sigma)$.

Separating curves. Suppose that a is separating, so we have $\Sigma|a = \Sigma_1 \sqcup \Sigma_2$ with each Σ_i of type (g_i, n_i) satisfying $2g_i + n_i - 2 > 0$. Let a_i be the boundary curve of Σ_i corresponding to a. Let (x_i, α_i) be the lift of (x, α) to a_i. We shall assume that we have chosen the numberings so that the interior of Σ_1 lies to the left as one travels along α_1. The immersions $\Sigma_i \hookrightarrow \Sigma$ of the surfaces induce embeddings $\pi_1(\Sigma_i, x_i) \to \pi_1(\Sigma, x)$ of fundamental groups, and we have an isomorphism

$$\pi_1(\Sigma, x) \simeq (\pi_1(\Sigma_1, x_1) \vee \pi_1(\Sigma_2, x_2))/(\alpha_1 = \alpha_2).$$

Thus, any representation $\rho : \pi_1(\Sigma, x) \to \mathrm{SL}_2(\mathbb{C})$ is determined uniquely by a pair (ρ_1, ρ_2) of representations $\rho_i : \pi_1(\Sigma_i, x_i) \to \mathrm{SL}_2(\mathbb{C})$ such that $\rho_1(\alpha_1) = \rho_2(\alpha_2)$, with the correspondence

$$\rho \mapsto (\rho_1, \rho_2) = (\rho|_{\pi_1(\Sigma_1, x_1)}, \rho|_{\pi_1(\Sigma_2, x_2)}).$$

We define an automorphism τ_α of $\mathrm{Hom}(\pi_1(\Sigma, x), \mathrm{SL}_2)$ as follows. For a representation $\rho = (\rho_1, \rho_2)$, we set $\tau_\alpha(\rho_1, \rho_2) = (\rho_1, \rho_2')$, where

$$\rho_2'(\gamma) = \rho(\alpha)\rho_2(\gamma)\rho(\alpha)^{-1}$$

for every $\gamma \in \pi_1(\Sigma_2, x_2)$. This descends to the action τ_a of the left Dehn twist along a on the moduli space $X(\Sigma)$.

2C. ***Markoff type cubic surfaces.*** Here, we give a description of the moduli spaces $X_k(\Sigma)$ and their mapping class group dynamics for $(g, n) = (1, 1)$ and $(0, 4)$. These cases are distinguished by the fact that each X_k is an affine cubic algebraic surface with an explicit equation.

2C1. *Case* $(g, n) = (1, 1)$. Let Σ be a surface of type $(g, n) = (1, 1)$, i.e., a one holed torus. Let (α, β, γ) be an optimal sequence of generators for $\pi_1\Sigma$. By Example 2.1(2), we have an identification $(\mathrm{tr}_\alpha, \mathrm{tr}_\beta, \mathrm{tr}_{\alpha\beta}) : X(\Sigma) \simeq \mathbb{A}^3$. From the trace relations in Section 2B, we obtain

$$\begin{aligned}\mathrm{tr}_\gamma &= \mathrm{tr}_{\alpha\beta\alpha^{-1}\beta^{-1}} = \mathrm{tr}_{\alpha\beta\alpha^{-1}}\,\mathrm{tr}_{\beta^{-1}} - \mathrm{tr}_{\alpha\beta\alpha^{-1}\beta} \\ &= \mathrm{tr}_\beta^2 - \mathrm{tr}_{\alpha\beta}\,\mathrm{tr}_{\alpha^{-1}\beta} + \mathrm{tr}_{\alpha\alpha} = \mathrm{tr}_\beta^2 - \mathrm{tr}_{\alpha\beta}(\mathrm{tr}_{\alpha^{-1}}\,\mathrm{tr}_\beta - \mathrm{tr}_{\alpha\beta}) + \mathrm{tr}_\alpha^2 - \mathrm{tr}_1 \\ &= \mathrm{tr}_\alpha^2 + \mathrm{tr}_\beta^2 + \mathrm{tr}_{\alpha\beta}^2 - \mathrm{tr}_\alpha\,\mathrm{tr}_\beta\,\mathrm{tr}_{\alpha\beta} - 2.\end{aligned}$$

Writing $(x, y, z) = (\mathrm{tr}_\alpha, \mathrm{tr}_\beta, \mathrm{tr}_{\alpha\beta})$ so that each of the variables x, y, and z corresponds to an essential curve on Σ as depicted in Figure 2, the moduli space $X_k \subset X$ has an explicit presentation as an affine cubic algebraic surface in $\mathbb{A}^3_{x,y,z}$ with equation

$$x^2 + y^2 + z^2 - xyz - 2 = k.$$

2C2. *Case* $(g, n) = (0, 4)$. Let Σ be a surface of type $(0, 4)$, i.e., a four holed sphere. Let $(\gamma_1, \ldots, \gamma_4)$ be an optimal sequence of generators for $\pi_1\Sigma$. Let us set

$$(x, y, z) = (\mathrm{tr}_{\gamma_1\gamma_2}, \mathrm{tr}_{\gamma_2\gamma_3}, \mathrm{tr}_{\gamma_1\gamma_3}),$$

so that each of the variables corresponds to an essential curve on Σ as depicted in Figure 3. By Example 2.1(3), for $k = (k_1, k_2, k_3, k_4) \in \mathbb{A}^4(\mathbb{C})$ the relative character variety $X_k = X_k(\Sigma)$ is an affine cubic algebraic surface in $\mathbb{A}^3_{x,y,z}$ given by

$$x^2 + y^2 + z^2 + xyz = Ax + By + Cz + D,$$

with

$$\begin{cases} A = k_1k_2 + k_3k_4 \\ B = k_1k_4 + k_2k_3 \\ C = k_1k_3 + k_2k_4 \end{cases} \quad \text{and} \quad D = 4 - \sum_{i=1}^4 k_i^2 - \prod_{i=1}^4 k_i.$$

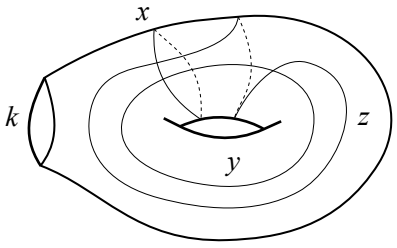


Figure 2. Curves on a surface of type (1, 1) with corresponding functions.

2D. ***Nonarchimedean systoles.*** In [Whang 2017], we proved the following result.

Theorem 2.5. *Let $\mathcal{O}$ be a discrete valuation ring with fraction field F. Given any representation $\rho : \pi_1\Sigma \to \mathrm{SL}_2(F)$ whose boundary traces all take values in $\mathcal{O}$, there is an essential curve $a \subset \Sigma$ with* $\operatorname{tr}\rho(a) \in \mathcal{O}$.

Corollary 2.6. *Let $\mathcal{O}$ and F be as above. If F has characteristic zero, then for any $\rho \in X_k(F)$ with $k \in \mathcal{O}^n$ there is an essential curve $a \subset \Sigma$ such that* $\operatorname{tr}\rho(a) \in \mathcal{O}$.

Proof. Given $\rho \in X_k(F)$, since F has characteristic zero there exists a finite field extension F'/F such that ρ is the class of a representation $\rho' : \pi_1\Sigma \to \mathrm{SL}_2(F')$. Choosing an extension of the valuation on F to F', let $\mathcal{O}' \subset F'$ be the associated valuation ring. It follows from Theorem 1.1 and our hypothesis $k \in \mathcal{O}^n \subseteq (\mathcal{O}')^n$ that there is an essential curve $a \subset \Sigma$ with $\operatorname{tr}\rho'(a) \in \mathcal{O}'$. This then implies that $\operatorname{tr}\rho(a) = \operatorname{tr}\rho'(a) \in \mathcal{O}' \cap F = \mathcal{O}$, which is the desired result. □

Below, we record a special case of Corollary 2.6, which plays a crucial role when we analyze the structure of morphisms from the affine line to X_k in Section 3.

Lemma 2.7. *Given any morphism $f : \mathbb{A}^1 \to X_k$ over $\mathbb{C}$, there is an essential curve $a \subset \Sigma$ such that* $\operatorname{tr}_a \circ f : \mathbb{A}^1 \to \mathbb{A}^1$ *is constant.*

Proof. A morphism $f : \mathbb{A}^1 \to X_k$ corresponds to a $\mathbb{C}[t]$-valued point of X_k, giving rise to a $\mathbb{C}(t)$-valued point $\rho_f \in X_k(\mathbb{C}(t))$. Applying Corollary 2.6 with $F = \mathbb{C}(t)$, with discrete valuation given by the order of vanishing at ∞, we deduce that there is an essential curve $a \subset \Sigma$ such that $\operatorname{tr}\rho_f(a) = \operatorname{tr}_a \circ f : \mathbb{A}^1 \to \mathbb{A}^1$ has no pole at ∞, which implies that it must be constant, as desired. □

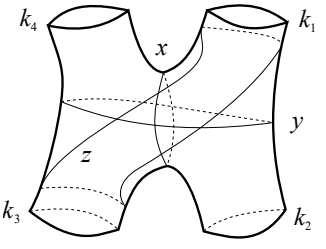


Figure 3. Curves on surfaces of type (0, 4) with corresponding functions.

3. Parabolic subvarieties

Let Σ be a surface of type (g, n) satisfying $3g+n-3>0$, and let $X_k = X_k(\Sigma)$ for $k \in X(\partial\Sigma, \mathbb{C})$ be a relative character variety of Σ. Given a pants decomposition P of Σ, the immersion $P \to \Sigma$ induces a morphism

$$\mathrm{tr}_P : X_k \to X(P) \simeq \mathbb{A}^{3g+n-3}$$

whose fibers for $t \in X(P, \mathbb{C})$ will be denoted $X^P_{k,t} = \mathrm{tr}_P^{-1}(t)$. Let

$$(-)|_{\Sigma|P} : X_k \to X(\Sigma|P) = X(\Sigma_1) \times \cdots \times X(\Sigma_{2g+n-2})$$

be the morphism induced by the immersion $\Sigma|P \to \Sigma$, where the product on the right-hand side is taken over the connected components Σ_i of $\Sigma|P$. Since $\pi_1\Sigma$ is a free group of rank 2 and a point of $X(\Sigma_i) \simeq \mathbb{A}^3$ (see Example 2.1(2)) is determined by the value of its traces along the boundary curves of Σ_i, it follows that $(-)|_{\Sigma|P}$ is constant along each fiber $X^P_{k,t}$. We make the following definition.

Definition 3.1. Let (P, t) be as above. The fiber $X^P_{k,t}$ is *perfect* if

(1) $\mathrm{tr}_a(X^P_{k,t}) \neq \pm 2$ for every curve $a \subseteq P$, and

(2) each factor of $(X^P_{k,t})|_{\Sigma|P}$ is irreducible, or $(g, n, k) = (1, 1, 2)$.

The first part of condition (2) in the definition above means that, for each connected component Σ_i of $\Sigma|P$, the point $(X^P_{k,t})|_{\Sigma_i}$ is represented by an irreducible local system on Σ_i, or an irreducible representation $\pi_1\Sigma_i \to \mathrm{SL}_2(\mathbb{C})$.

The above definition is motivated by the following theorem, which is the main result of this section. Recall from Section 1 that an algebraic variety Z over $\mathbb{C}$ is said to be *parabolic* if every closed point of Z lies in the image of some nonconstant morphism $\mathbb{A}^1 \to Z$. In the following, a pair (P, t) will denote a pants decomposition P of Σ and an element $t \in X(P, \mathbb{C})$.

Theorem 3.2. (A) *For each (P, t), the fiber $X^P_{k,t}$ is either perfect or parabolic.*

(B) *For any nonconstant morphism $f : \mathbb{A}^1 \to X_k$, there is a parabolic fiber $X^P_{k,t}$ for some (P, t) containing the image of f.*

The remainder of this section is organized as follows. In Section 3A, we give a proof of Theorem 3.2 in the cases $(g, n) = (1, 1)$ and $(0, 4)$. The moduli space X_k in these cases is an explicitly defined algebraic surface, making the proof easier. We prove the general case of Theorem 3.2 in Section 3B. In Section 3C, we describe the consequences of Theorem 3.2, including Theorem 1.1 as well as a rigidity result (Corollary 3.6) for certain polynomial deformations of Fuchsian representations of surface groups.

3A. ***Base cases.*** In this subsection, we give a proof of Theorem 3.2 in the cases $(g, n) = (1, 1)$ and $(0, 4)$. We shall refer the reader to Section 2C for explicit presentations of X_k in these cases. First, it is useful to record an elementary lemma.

Lemma 3.3. *If Σ is a surface of type $(0, 3)$, then $k = (k_1, k_2, k_3) \in X(\Sigma) \simeq \mathbb{A}^3$ is reducible if and only if $k_1^2 + k_2^2 + k_3^2 - k_1k_2k_3 - 2 = 2$.*

Proof. This follows by combining the observation that two matrices $A, B \in \mathrm{SL}_2(\mathbb{C})$ share an eigenvector if and only if $\mathrm{tr}(ABA^{-1}B^{-1}) = 2$ with the expression for this trace in terms of $\mathrm{tr}(A)$, $\mathrm{tr}(B)$, and $\mathrm{tr}(AB)$, derived for instance in Section 2C1. □

3A1. *Surfaces of type* (1, 1). Suppose Σ is of type (1, 1), and let (α, β, γ) be optimal generators of $\pi_1\Sigma$. The moduli space X_k can be presented as an affine cubic algebraic surface, with equation

$$x^2 + y^2 + z^2 - xyz - 2 = k,$$

where the variables (x, y, z) correspond to the monodromy traces $(\mathrm{tr}_\alpha, \mathrm{tr}_\beta, \mathrm{tr}_{\alpha\beta})$ and $\mathrm{tr}_\gamma \equiv k$. Now, let $a \subset \Sigma$ be the essential curve underlying α. Since any essential curve of Σ is, up to isotopy, mapping class group equivalent to a, it suffices to prove Theorem 3.2(A) for $(g, n) = (1, 1)$ where $P = a$. Let $t \in X(a, \mathbb{C}) \simeq \mathbb{C}$. (We are following the notation of Section 2B2; in effect, this means we are fixing the trace t of monodromy along a.) Since $X^P_{k,t} = X^a_{k,t}$ is a conic section

$$y^2 - tyz + z^2 + t^2 - 2 - k = 0$$

in the (y, z)-plane, elementary geometry shows that $X^a_{k,t}$ is parabolic if and only if $t = \pm 2$ or the conic section is degenerate. Let us assume $t \neq \pm 2$; the latter condition states that the equation for $X^a_{k,t}$ factors as

$$y^2 - tyz + z^2 + t^2 - 2 - k = (y - \lambda z + m_1)(y - \lambda^{-1}z + m_2) = 0$$

for some $\lambda \in \mathbb{C}^*$ and $m_i \in \mathbb{C}$. Expanding and comparing coefficients, we must have

$$\lambda + \lambda^{-1} = t.$$

We must also have

$$\begin{bmatrix} 1 & 1 \\ -\lambda & -\lambda^{-1} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and hence $m_1 = m_2 = 0$, and therefore

$$m_1 m_2 = t^2 - 2 - k = 0,$$

and in particular $k \neq 2$.

Thus, we see that $X^a_{k,t}$ is parabolic if and only if

(1) $t = \pm 2$, or

(2) $t \neq \pm 2$, $(g, n, k) \neq (1, 1, 2)$, and $(X^a_{k,t})|_{\Sigma|a} = (t, t, t^2 - 2)$.

Here, we made an identification $X(\Sigma|a) \simeq \mathbb{A}^3$. By Lemma 3.3, the last condition in (2) is equivalent to saying that $(X^a_{k,t})|_{\Sigma|a}$ is reducible. This concludes the proof of Theorem 3.2(A) for $(g, n) = (1, 1)$. Combining this with Lemma 2.7, we immediately obtain Theorem 3.2(B) in this case.

3A2. *Surfaces of type* $(0,4)$. Suppose Σ is of type $(0,4)$, and let $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ be optimal generators of $\pi_1\Sigma$. The moduli space X_k is an affine cubic algebraic surface with equation

$$x^2 + y^2 + z^2 + xyz - Ax - By - Cz - D = 0,$$

where the variables (x, y, z) correspond to the traces $(\mathrm{tr}_{\gamma_1\gamma_2}, \mathrm{tr}_{\gamma_2\gamma_3}, \mathrm{tr}_{\gamma_1\gamma_3})$ and

$$\begin{cases} A = k_1k_2 + k_3k_4 \\ B = k_2k_3 + k_1k_4 \\ C = k_1k_3 + k_2k_4 \end{cases} \quad \text{and} \quad D = 4 - k_1^2 - k_2^2 - k_3^2 - k_4^2 - k_1k_2k_3k_4$$

with $k_i \equiv \mathrm{tr}\,\gamma_i$. Now, let $a, b, c \subset \Sigma$ be essential curves lying in the free homotopy classes of $\gamma_1\gamma_2$, $\gamma_2\gamma_3$, and $\gamma_1\gamma_3$, respectively. Since any essential curve of Σ is, up to isotopy, mapping class group equivalent to one of the curves a, b, and c, it suffices to prove Theorem 3.2(A) for $(g,n) = (0,4)$ where P consists of one of the curves a, b, and c. We treat the case $P = a$ in what follows; the remaining cases will proceed similarly. Let $t \in X(a, \mathbb{C}) \simeq \mathbb{C}$. Again by elementary geometry, we see that $X^a_{k,t}$ is parabolic if and only if $t = \pm 2$ or $X^a_{k,t}$ is a degenerate conic in the (y,z)-plane. Let us assume $t \neq \pm 2$; the latter condition states that the equation for $X^a_{k,t}$ factors as

$$t^2 + y^2 + z^2 + tyz - At - By - Cz - D = (y + \lambda z + m_1)(y + \lambda^{-1}z + m_2) = 0$$

for some $\lambda \in \mathbb{C}^*$ and $m_i \in \mathbb{C}$. Expanding and comparing coefficients, we see that

$$\lambda + \lambda^{-1} = t, \quad \begin{bmatrix} 1 & 1 \\ \lambda & \lambda^{-1} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} -B \\ -C \end{bmatrix} \quad \text{and} \quad m_1m_2 = -At - D.$$

This is equivalent to

$$-\frac{B^2 - tBC + C^2}{t^2 - 4} = \frac{1}{(\lambda^{-1} - \lambda)^2}(\lambda^{-1}B - C)(-\lambda B + C) = -At - D$$

or in other words $(t^2 - 4)(-Ax - D) + (B^2 - tBC + C^2) = 0$. Upon rearranging, this is seen to be equivalent to

$$(k_1^2 + k_2^2 + x^2 - tk_1k_2 - 4)(k_3^2 + k_4^2 + x^2 - tk_3k_4 - 4) = 0,$$

which in turn is equivalent to saying that at least one factor of $(X^a_{k,t})|_{\Sigma|a}$ is reducible, by Lemma 3.3. This proves Theorem 3.2(A) for $(g,n) = (0,4)$, and Theorem 3.2(B) follows by Lemma 2.7.

3B. ***General case.*** Let Σ be a surface of type (g,n) with $3g + n - 3 > 0$, and let $X_k = X_k(\Sigma)$ be a relative character variety of Σ. We shall first prove the following claim, by induction on (g,n). Claim 3.4 proves Theorem 3.2(B), conditional on Theorem 3.2(A).

Claim 3.4. *For any nonconstant morphism $f : \mathbb{A}^1 \to X_k$, there exists an imperfect fiber $X^P_{k,t}$ for some (P, t) containing the image of f.*

Proof. We have already proved Theorem 3.2 for $(g, n) = (1, 1)$ and $(0, 4)$, and these will provide the base cases for our induction. Let $f : \mathbb{A}^1 \to X_k$ be a nonconstant morphism. By Lemma 2.7, there is an essential curve $a \subset \Sigma$ such that $\mathrm{tr}_a \circ f \equiv t_0$ is constant. Suppose first that the restriction

$$f|_{\Sigma'} : \mathbb{A}^1 \to X_k(\Sigma) \xrightarrow{(-)|_{\Sigma'}} X(\Sigma')$$

of f to some connected component Σ' of $\Sigma|a$ is nonconstant. We observe that the image of $f|_{\Sigma'}$ lies in $X_{k'}(\Sigma')$ for some $k' \in X(\partial\Sigma', \mathbb{C})$ determined by k and t_0. Hence, by inductive hypothesis, there is a pants decomposition P' of Σ' and an element $t' \in X(P', \mathbb{C})$ such that $X^{P'}_{k',t'}(\Sigma')$ is an imperfect fiber containing the image of $f|_{\Sigma'}$. If a is nonseparating, then taking

$$(P, t) = (P' \sqcup a, (t', t_0))$$

we see that $X^P_{k,t}$ is an imperfect fiber containing the image of f, as desired. If f is separating and $\Sigma|a = \Sigma' \sqcup \Sigma''$, then again $f|_{\Sigma''}$ has image lying in some $X_{k''}(\Sigma'')$ for some $k'' \in X(\partial\Sigma'', \mathbb{C})$ determined by k and t_0, and by a repeated application of Lemma 2.7 we see that there is a pants decomposition P'' of Σ'' and $t'' \in X(P'', \mathbb{C})$ such that $X^{P''}_{k'',t''}$ contains the image of $f|_{\Sigma''}$. Taking

$$(P, t) = (P' \sqcup a \sqcup P'', (t', t_0, t'')),$$

we again see that $X^P_{k,t}$ is an imperfect fiber containing the image of f.

To complete our proof of the claim, it remains only to consider the case where $f|_{\Sigma|a}$ is constant.

We first consider the case where a is nonseparating. Let $\alpha_1, \ldots, \alpha_{2g+n}$ be a sequence of optimal generators for $\pi_1\Sigma$. Up to mapping class group action, we may assume that a is the essential curve underlying α_1. By Fact 2.3, the coordinate ring of the fiber $X_k \to X(\Sigma|a)$ above $f|_{\Sigma|a}$ is generated by functions of the form

$$\mathrm{tr}_{\alpha_2}, \mathrm{tr}_{\alpha_2\alpha_i}, \mathrm{tr}_{\alpha_2\alpha_i\alpha_j}, \mathrm{tr}_{\alpha_1\alpha_2}, \mathrm{tr}_{\alpha_1\alpha_2\alpha_i}$$

for $3 \le i < j \le 2g + n$. Thus, since f is nonconstant, the composition of f with at least one of the above coordinate functions must be nonconstant. Let us consider the case where $\mathrm{tr}_{\alpha_2} \circ f$ is nonconstant; the other cases will follow similarly. There is a surface $\Sigma' \subset \Sigma$ of type $(1, 1)$ containing the loops α_1 and α_2 in its interior. By our hypothesis, we see that $f|_{\Sigma'}$ is nonconstant, and the image of $f|_{\Sigma'}$ lies in $X_{k'}(\Sigma)$ for some $k' \in X(\partial\Sigma', \mathbb{C})$ determined by the (constant) value of $f|_{\Sigma|a}$; indeed, the boundary of Σ' lies in $\Sigma|a$. Thus, by the case of Theorem 3.2 for $(g, n) = (1, 1)$ proved above, there is an essential curve $a' \subset \Sigma'$ and $t' \in X(a', \mathbb{C})$ such that $X^{a'}_{k',t'}(\Sigma')$ is an imperfect fiber containing the image of $f|_{\Sigma'}$. Completing $a' \sqcup \partial\Sigma' \subset \Sigma$ to a pants decomposition P of Σ, we see that there is some $t \in X(P, \mathbb{C})$ such that $X^P_{k,t}$ is an imperfect fiber containing the image of f, as desired.

The case where a is separating is very similar, by appropriately invoking the case of Theorem 3.2 for $(g, n) = (0, 4)$. □

Suppose that $P = a_1 \sqcup \cdots \sqcup a_{3g+n-3}$ is a pants decomposition of Σ, and suppose further that $t = (t_1, \ldots, t_{3g+n-3}) \in X(P, \mathbb{C}) \simeq \mathbb{C}^{3g+n-3}$ is chosen so that $X^P_{k,t}$ is a perfect fiber. Let us consider the morphism

$$X^P_{k,t}(\Sigma) \to \prod_{i=1}^{3g+n-3} X^{a_i}_{k_i,t_i}(\Sigma_i),$$

where each Σ_i is the surface of type $(0, 4)$ or $(1, 1)$ obtained by gluing together the two boundary curves on $\Sigma|P$ corresponding to a_i, and the boundary traces k_i are appropriately determined from k, P, t, and Σ_i. As a consequence of Proposition 4.3 proved in Section 4C, the above morphism is finite at the level of complex points (cf. proof of Corollary 4.4). Combining this with the results from Section 3A, we deduce that there cannot be a nonconstant morphism from $\mathbb{A}^1$ to such a perfect fiber, proving one half of Theorem 3.2(A).

We shall henceforth assume that $(g, n) \neq (1, 1), (0, 4)$, to simplify our remaining argument. To complete the proof of Theorem 3.2, it suffices to prove the following.

Claim 3.5. *Assume that $(g, n) \neq (1, 1), (0, 4)$. Given a semisimple representation $\rho : \pi_1\Sigma \to \mathrm{SL}_2(\mathbb{C})$ whose image in the character variety $X(\Sigma)$ lies in an imperfect fiber $X^P_{k,t}$, there is a one-parameter polynomial family*

$$\rho_T : \pi_1\Sigma \to \mathrm{SL}_2(\mathbb{C})$$

of representations with nonconstant images all lying in $X^P_{k,t}$, so that we have $\rho = \rho_{T_0}$ for some $T_0 \in \mathbb{C}$.

Proof. Let ρ and $X^P_{k,t}$ be as above. We shall argue by division into several cases. For the benefit of the reader, we list the cases broadly considered and their hypotheses.

(1) **Parabolic curve.** There is a curve $a \subset P$ with $\mathrm{tr}_a(X^P_{k,t}) = \pm 2$.
   Case 1: The curve a is separating.
   Case 2: The curve a is nonseparating.

(2) **Parabolic pants.** There is no curve $a \subset P$ with trace ± 2, but there is a component Σ' of $\Sigma|P$ such that $X^P_{k,t}|\Sigma'$ is reducible.
   Case 1: The image of Σ' in Σ is a surface of type $(1, 1)$.
   Case 2: The image of Σ' in Σ is a surface of type $(0, 3)$.

We now begin our proof.

Parabolic curve. Let us first consider the case where there is a curve $a \subseteq P$ with $\mathrm{tr}_a(X^P_{k,t}) = \pm 2$. We may assume to have fixed the base point of Σ to lie on a. Let α be a smooth simple loop parametrizing a. Up to global conjugation, we may assume that

$$\rho(\alpha) = s \begin{bmatrix} 1 & u \\ 0 & 1 \end{bmatrix} \qquad (*)$$

for $s \in \{\pm 1\}$ and $u \in \{0, 1\}$. There are several elementary cases to consider.

Case 1: The curve a is separating. Let us write $\Sigma|a = \Sigma_1 \sqcup \Sigma_2$. Up to conjugation, we may assume that on top of $(*)$ the following conditions hold:

(1) $\rho|\Sigma_1$ is irreducible or upper triangular.

(2) $\rho|\Sigma_2$ is irreducible or lower triangular.

(3) If $\rho|\Sigma_1$ and $\rho|\Sigma_2$ are both reducible, then they are either both nondiagonal or both diagonal.

Indeed, if one of $\rho|\Sigma_1$ and $\rho|\Sigma_2$ is irreducible, then up to relabeling Σ_1 and Σ_2 we may assume $\rho|\Sigma_2$ is irreducible, and $\rho|\Sigma_1$ must be irreducible or upper triangular up to conjugation. So suppose both $\rho|\Sigma_1$ and $\rho|\Sigma_2$ are reducible. This implies that $u = 0$ in $(*)$ above, since otherwise ρ must be upper triangular and nondiagonal, contradicting the hypothesis that ρ is semisimple. Unless ρ is reducible (whence diagonal), there is a basis v_1, v_2 of $\mathbb{C}^2$ such that each v_i is a common eigenvector for $\rho|\Sigma_i$. Up to conjugation $M^{-1}\rho M$ of ρ by the invertible matrix $M = [v_1, v_2]$, we may thus assume $\rho|\Sigma_1$ is upper triangular and $\rho|\Sigma_2$ is lower triangular. For convenience, we shall denote $\rho_i = \rho|\Sigma_i$ so that we may write $\rho = (\rho_1, \rho_2)$ using the notation of the second part of Section 2B3.

Subcase 1A: ρ_2 is nondiagonal. Let us consider the family of representations $\rho^T = (\rho_1, u_T \rho_2 u_T^{-1})$ for $T \in \mathbb{C}$, where

$$u_T = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}.$$

Note that $\rho_1(\alpha) = u_T \rho_2(\alpha) u_T^{-1}$ so the representation ρ^T is well-defined. For any $\beta \in \pi_1\Sigma_1$ and $\gamma \in \pi_1\Sigma_2$ with

$$\rho(\beta) = \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix} \quad \text{and} \quad \rho(\gamma) = \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix},$$

we have

$$\begin{aligned} \operatorname{tr} \rho^T(\beta\gamma) &= \operatorname{tr}\left(\begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix} \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix} \begin{bmatrix} 1 & -T \\ 0 & 1 \end{bmatrix} \right) \\ &= b_1c_2 + b_2c_3 + b_3c_2 + b_4c_4 + (b_1c_3 - b_3c_1 - b_4c_3 + b_3c_4)T - b_3c_3T^2. \end{aligned}$$

Since ρ_2 is nondiagonal while irreducible or lower triangular, we may choose γ as above so that $c_3 \neq 0$. We have the following possibilities.

(a) Suppose ρ_1 is irreducible. We can choose β as above with $b_3 \neq 0$, so that $\operatorname{tr}_{\beta\gamma}(\rho^T)$ is a nonconstant function of T.

(b) Suppose ρ_1 is upper triangular (so $b_3 = 0$ for any choice of β), and there exists $\beta \in \pi_1\Sigma_1$ such that $\operatorname{tr} \rho(\beta) \neq \pm 2$. Choosing such β we find that $\operatorname{tr}_{\beta\gamma}(\rho^T)$ is a nonconstant function of T since $(b_1 - b_4)c_3 \neq 0$.

(c) Consider the case where ρ_1 is upper triangular and $\operatorname{tr}\rho(\beta) = \pm 2$ for any $\beta \in \pi_1\Sigma_1$. This implies that the image of ρ_1 is abelian. Suppose Σ_1 is of type (h, m), and let $S = (\alpha_1, \beta_1, \ldots, \alpha_h, \beta_h, \gamma_1, \ldots, \gamma_m)$ be an optimal sequence of generators for $\pi_1\Sigma_1$ such that $\gamma_m = \alpha$ (up to homotopy). Let us define a one-parameter family ρ_1^T of upper triangular deformations of $\rho_1 = \rho_1^0$ given by setting

$$\begin{cases} \rho_1^T(\alpha_1) = \rho_1(\alpha_1)u_T, & \text{and} \\ \rho_1^T(\ell) = \rho_1(\ell) & \text{for any other } \ell \in S \end{cases}$$

if $h \geq 1$, and setting

$$\begin{cases} \rho_1^T(\gamma_1) = \rho_1(\gamma_1)u_T, & \\ \rho_1^T(\gamma_2) = u_T^{-1}\rho_1(\gamma_2), & \text{and} \\ \rho_1^T(\ell) = \rho_1(\ell) & \text{for any other } \ell \in S \end{cases}$$

if $h = 0$ so that $m \geq 3$. Then choosing $\beta = \alpha_1$ (resp. $\beta = \gamma_1$) if $h \geq 1$ (resp. $h = 0$) we find that

$$\operatorname{tr}_{\beta\gamma}(\rho^T) = \operatorname{tr}\rho(\beta\gamma) \pm c_3 T$$

which is a nonconstant function of T.

Thus, in each of the cases the morphism $\mathbb{A}^1 \to X_{k,t}^P$ given by $T \mapsto \rho^T$ is nonconstant.

Subcase 1B: ρ_1 is nondiagonal. This case is established by the same argument as in Subcase 1A. The only difference is that, instead of the matrices u_T, we consider in appropriate places of our argument the matrices

$$l_T = \begin{bmatrix} 1 & 0 \\ T & 1 \end{bmatrix}.$$

Subcase 1C: Both ρ_1 and ρ_2 are diagonal. We shall first construct a nontrivial family of upper triangular representations ρ_1^T of $\pi_1\Sigma_1$ with $\rho_1^0 = \rho_1$ and $\rho_1^T(\alpha) = \rho_1(\alpha)$ for all T as follows.

- If there exists $\beta \in \pi_1\Sigma_1$ such that $\operatorname{tr}\rho(\beta) \neq \pm 2$, then let $\rho_1^T = u_T\rho_1 u_T^{-1}$.
- If $\operatorname{tr}\rho_1(\beta) = \pm 2$ for all $\beta \in \pi_1\Sigma$, then define ρ_1^T as in the treatment of possibility (c) in Subcase 1A.

Note that, in both cases, there exists $\beta \in \pi_1\Sigma_1$ such that the upper right corner entry of $\rho_1^T(\beta)$ is a nonconstant polynomial function of T. Similarly, let us construct a nontrivial family of lower triangular representations ρ_2^T of $\pi_1\Sigma_2$ with $\rho_2^0 = \rho_2$ and $\rho_2^T(\alpha) = \rho_2(\alpha)$ for all T, in such a way that there exists $\gamma \in \pi_1\Sigma_2$ such that the lower left corner entry of $\rho_2^T(\gamma)$ is a nonconstant polynomial of T.

Finally, let us define the representation $\rho^T = (\rho_1^T, \rho_2^T)$, which makes sense since we have $\rho_1^T(\alpha) = \rho(\alpha) = \rho_2^T(\alpha)$ for all T by construction. For β and γ chosen as above, we see that $\operatorname{tr}_{\beta\gamma}(\rho^T)$ is a nonconstant polynomial in T. Thus the morphism $\mathbb{A}^1 \to X_{k,t}^P$ defined by $T \mapsto \rho^T$ is nonconstant, passes through ρ.

Case 2: the curve a is nonseparating. We shall write $\rho = (\rho|(\Sigma|a), \rho(\beta)) = (\rho', B)$ using the notation of first part of Section 2B3, with a choice of simple loop β intersecting α exactly once. Up to conjugation, we may assume that the representation ρ' is irreducible or upper triangular.

Subcase 2A: ρ' is irreducible or B is not upper triangular. Let us consider the family of representations

$$\rho^T = (\rho', B_T),$$

where

$$B_T = u_T B = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} B.$$

Note that ρ^T is well-defined since $B^{-1}u_T^{-1}\rho(\alpha_1)u_T B = B^{-1}\rho(\alpha_1)B = \rho(\alpha_2)$ in the notation of Section 2B3. Now, consider the morphism $f : \mathbb{A}^1 \to X^P_{k,t}$ given by $T \mapsto \rho^T$. Note that we have $\rho^0 = \rho$. We claim that this morphism is nonconstant. To see this, it suffices to show that there is some element $\gamma \in \pi_1(\Sigma)$ where $\mathrm{tr}_\gamma \circ f : \mathbb{A}^1 \to \mathbb{A}^1$ is nonconstant. We have two possibilities.

- If B is not upper triangular, then $\gamma = \beta$ suffices.
- If B is upper triangular but ρ' is irreducible, then there exists $\delta \in \pi_1(\Sigma|a)$ which is not upper triangular. It suffices to choose $\gamma = \beta\delta$.

Subcase 2B: ρ' and B are both upper triangular, so that ρ is reducible. Since ρ is semisimple, ρ must be diagonal and in particular $\rho(\alpha) = \pm\mathbf{1}$. Let $\Sigma_1 \subset \Sigma$ be the subsurface of type $(1, 1)$ obtained by taking a closed tubular neighborhood of $a \cup b$, where b is the curve underlying β. Let c be the boundary curve of Σ_1, and write $\Sigma|c = \Sigma_1 \sqcup \Sigma_2$, where Σ_2 is a surface of type $(g-1, n+1)$. For convenience, we shall denote $\rho_i = \rho|\Sigma_i$.

Let us write $\rho_1 = (\rho_1', B)$ in the notation of Section 2B3 (with the same choice of α and β as before), and consider the family of lower triangular representations $\rho_1^T = (\rho_1', B_T)$, where

$$B_T = l_T B = \begin{bmatrix} 1 & 0 \\ T & 1 \end{bmatrix} B.$$

Note that the lower left entry of B_T is a nonconstant function of T. Now, without loss of generality, we may assume that our new basepoint lies on c. Let γ be a simple loop parametrizing c, so that $\rho(\gamma) = \mathbf{1}$ and $\rho_1^T(\gamma)$ is constant for all T. Proceeding as in Case 1, we can construct a nonconstant upper triangular deformation ρ_2^T of ρ_2 with $\rho_2^0 = \rho_2$ such that $\rho_2^T(\gamma) = \mathbf{1}$ for all T:

- If there exists $\gamma \in \pi_1\Sigma_2$ such that $\mathrm{tr}\,\rho(\gamma) \neq \pm 2$, then $\rho_2^T = u_T \rho u_T^{-1}$.
- If $\mathrm{tr}\,\rho_2(\beta) = \pm 2$ for all $\beta \in \pi_1\Sigma$, then define ρ_2^T as in the treatment of possibility (c) in Subcase 1A.

Then the representation $\rho^T = (\rho_1^T, \rho_2^T)$ (in the notation of second part of Section 2B3) is well-defined and has the property that $\rho^0 = \rho$. Since there exists $\gamma \in \pi_1\Sigma_2$ such that the upper right entry of $\rho^T(\gamma)$ is nonconstant, the nonconstancy of consideration of $\mathrm{tr}_{\beta\gamma}(\rho^T)$ shows that the morphism $f : \mathbb{A}^1 \to X_k$ given by $T \mapsto \rho^T$ is nonconstant. Moreover, since $\rho^T|(\Sigma|a)$ remains upper triangular, we see that the image of f lies in $X^P_{k,t}$, as desired.

Parabolic pants. We now consider the case where the following conditions hold:

(1) $\mathrm{tr}_a(X^P_{k,t}) \neq \pm 2$ for every curve $a \subseteq P$.

(2) $(g, n, k) \neq (1, 1, 2)$.

(3) $(X^P_{k,t})|_{\Sigma'}$ is reducible for some connected component Σ' of $\Sigma|P$.

We may assume for convenience that the base point $x \in \Sigma$ lies on Σ'. Let $\gamma_1, \gamma_2, \gamma_3$ be optimal generators for $\pi_1\Sigma'$ corresponding to the boundary curves c_1, c_2, c_3 of Σ'. We shall write $\rho|_{\Sigma'} = (t_1, t_2, t_3) \in X(\Sigma') \simeq \mathbb{A}^3$ (see Example 2.1(2)). By relabeling the boundary curves of Σ' if necessary, we may assume that c_1 corresponds to a curve of P. In particular, $t_1 \neq \pm 2$ by our hypothesis above. We further assume that, if the image of Σ' in Σ is a surface of type $(1, 1)$, then c_1 and c_2 map to the same curve in P.

It will be convenient for us to introduce distinguished families of representations $\pi_1\Sigma' \to \mathrm{SL}_2(\mathbb{C})$ which are reducible. For each $T \in \mathbb{C}$ and $s \in \{\pm 1\}$, let $\rho^s_T : \pi_1\Sigma' \to \mathrm{SL}_2(\mathbb{C})$ be the representation determined by

$$\rho^s_T(\gamma_1) = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{bmatrix}, \quad \rho^s_T(\gamma_2) = \begin{bmatrix} \mu & T \\ 0 & \mu^{-1} \end{bmatrix},$$

where $\lambda \in \mathbb{C}^\times \setminus \{\pm 1\}$ and $\mu \in \mathbb{C}^\times$ are such that

$$\lambda^s \in \{z \in \mathbb{C}^\times : \Im(z) \geq 0, z \notin [-1, 1]\}$$

and

$$\begin{cases} t_1 = \lambda + \lambda^{-1}, \\ t_2 = \mu + \mu^{-1}, \\ t_3 = \lambda\mu + \lambda^{-1}\mu^{-1}. \end{cases}$$

(The sign s is put there simply to remove ambiguities; they do not play a significant role in the proof.) Note that the Jordan equivalence class of each ρ^s_T in $X(\Sigma')$ is equal to that of $\rho|_{\Sigma'}$ since a point of $X(\Sigma') \simeq \mathbb{A}^3$ is determined by its traces along the boundary curves of Σ'. Up to global conjugation, we may assume that $\rho|_{\Sigma'} = \rho^s_{T_0}$ for some $T_0 \in \mathbb{C}$ and $s \in \{\pm 1\}$. Below, we proceed with a fixed choice of $s \in \{\pm 1\}$, as it will not make a difference to the argument. We shall write also $\rho^s_T = \rho'_T$ for easier notation.

We must consider different cases, according to the relative position of Σ' in Σ.

Case 1: the image of Σ' in Σ is a surface of type $(1, 1)$. By our hypothesis, the boundary curves c_1 and c_2 map to the same curve $a \subset P$, while c_3 maps to a separating curve $c \subset \Sigma$. Let us write $\Sigma|c = \Sigma'' \sqcup \Sigma'''$ with Σ'' being the image of Σ'. Without loss of generality, we may assume that the implicit base point $x \in \Sigma$ lies on c. Let (α, β, γ) be a sequence of optimal generators for $\pi_1\Sigma''$, such that under the immersion $\Sigma' \to \Sigma''$ we have

$$\gamma_1 \mapsto \alpha, \quad \gamma_2 \mapsto \beta^{-1}\alpha^{-1}\beta, \quad \gamma_3 \mapsto \gamma.$$

Let us write

$$\rho(\beta) = \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}.$$

The condition $\rho(\gamma_2) = \rho(\beta^{-1}\alpha^{-1}\beta)$ is then

$$\begin{bmatrix} \lambda & T_0 \\ 0 & \lambda^{-1} \end{bmatrix} = \begin{bmatrix} B_4 & -B_2 \\ -B_3 & B_1 \end{bmatrix} \begin{bmatrix} \lambda^{-1} & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix}$$
$$= \begin{bmatrix} \lambda + (\lambda^{-1} - \lambda) B_1 B_4 & (\lambda^{-1} - \lambda) B_2 B_4 \\ (\lambda - \lambda^{-1}) B_1 B_3 & \lambda^{-1} + (\lambda - \lambda^{-1}) B_1 B_4 \end{bmatrix},$$

where we have $\lambda - \lambda^{-1} \neq 0$ by our hypothesis on ρ that $\operatorname{tr} \rho(\alpha) \neq \pm 2$. This shows that we must have

$$\rho(\beta) = \begin{bmatrix} 0 & B_2 \\ -B_2^{-1} & \operatorname{tr} \rho(\beta) \end{bmatrix},$$

where furthermore $(\lambda^{-1} - \lambda) B_2 \operatorname{tr} \rho(\beta) = T_0$. Up to global conjugation of ρ by a diagonal matrix (which also results in a suitable adjustment of T_0), we may further assume that $B_2 = 1$, and hence $\operatorname{tr} \rho(\beta) = T_0/(\lambda^{-1} - \lambda)$. Let $\rho''_T : \pi_1 \Sigma'' \to \mathrm{SL}_2(\mathbb{C})$ be the representation determined by

$$\rho''_T(\alpha) = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{bmatrix}, \quad \rho''_T(\beta) = \begin{bmatrix} 0 & 1 \\ -1 & \frac{T}{\lambda^{-1} - \lambda} \end{bmatrix}.$$

The preceding observations show that $\rho''_{T_0} = \rho|_{\Sigma''}$ and $\rho''_T|_{\Sigma'} = \rho'_T$ for every $T \in \mathbb{C}$. Note that we have

$$\rho''_T(\gamma) = \rho'_T(\gamma_3) = \begin{bmatrix} \lambda^{-2} & -\lambda T \\ 0 & \lambda^2 \end{bmatrix}.$$

For $T \in \mathbb{C}$, let $\rho_T : \pi_1 \Sigma \to \mathrm{SL}_2(\mathbb{C})$ be the representation such that $\rho_T|\Sigma''' = \rho$ and

$$\rho_T|\Sigma'' = \begin{bmatrix} 1 & \lambda \frac{T - T_0}{\lambda^2 - \lambda^{-2}} \\ 0 & 1 \end{bmatrix} \rho''_T \begin{bmatrix} 1 & \lambda \frac{T - T_0}{\lambda^2 - \lambda^{-2}} \\ 0 & 1 \end{bmatrix}^{-1}.$$

Here, $\lambda^2 - \lambda^{-2} \neq 0$ since otherwise $\operatorname{tr} \rho(\gamma) = \lambda^{-2} + \lambda^2 = \pm 2$, which was precluded. It can be directly verified that

$$\rho(\gamma) = \begin{bmatrix} 1 & \lambda \frac{T - T_0}{\lambda^2 - \lambda^{-2}} \\ 0 & 1 \end{bmatrix} \rho''_T(\gamma) \begin{bmatrix} 1 & \lambda \frac{T - T_0}{\lambda^2 - \lambda^{-2}} \\ 0 & 1 \end{bmatrix}^{-1}$$

so ρ_T is well-defined by our discussion in Section 2B3. We have $\rho = \rho_{T_0}$ by construction, and we see that the morphism $\mathbb{A}^1 \to X_k$ given by $T \mapsto \rho_T$ lies in the fiber $X^P_{k,t}$. The fact that this morphism is nonconstant can be deduced from the observation that

$$\operatorname{tr} \rho_T(\beta) = \frac{T}{\lambda^{-1} - \lambda}$$

is a nonconstant function of T. Therefore, this proves the claim when the image of Σ' in Σ is a surface of type $(1, 1)$.

Case 2: the boundary curves of Σ' map to three distinct curves in Σ under the immersion $\Sigma' \to \Sigma$ (which we shall also denote c_1, c_2, c_3 for simplicity). Now, let us write

$$\begin{cases} \Sigma | c_3 = \Sigma_3 \sqcup \Sigma_3^\circ \\ \Sigma_3 | c_2 = \Sigma_2 \sqcup \Sigma_2^\circ \\ \Sigma_2 | c_1 = \Sigma_1 \sqcup \Sigma_1^\circ \end{cases}$$

where Σ_3 is the connected component of $\Sigma|c_3$ containing Σ' (here, Σ_3° is empty if c_3 is nonseparating or is a boundary curve in Σ), Σ_2 is the connected component of $\Sigma_3|c_2$ containing Σ', and finally $\Sigma_1 = \Sigma'$.

Let us extend the polynomial family of representations $\rho'_T : \pi_1\Sigma' \to \mathrm{SL}_2(\mathbb{C})$ to the polynomial family $\rho''_T : \pi_1\Sigma_2 \to \mathrm{SL}_2(\mathbb{C})$ given by

$$\rho''_T|_{\Sigma'} = \rho'_T, \quad \rho''_T|_{\Sigma_1^\circ} = \rho|_{\Sigma_1^\circ}.$$

This is well-defined by our discussion in Section 2B3. Note that $\rho''_{T_0} = \rho|_{\Sigma_2}$. Our next step is to extend ρ''_T to a family $\rho'''_T : \pi_1\Sigma_3 \to \mathrm{SL}_2(\mathbb{C})$ such that $\rho'''_{T_0} = \rho|\Sigma_3$. We need to consider three cases.

(1) Suppose c_2 is a boundary curve on Σ_3, so that $\Sigma_2 = \Sigma_3$. Let $\rho'''_T = \rho''_T$.

(2) Suppose c_2 is a separating curve on Σ_3. We define ρ'''_T by requiring

$$\rho'''_T|_{\Sigma_2} = \rho''_T, \quad \rho'''_T|_{\Sigma_2^\circ} = \begin{bmatrix} 1 & \frac{T-T_0}{\mu^{-1}-\mu} \\ 0 & 1 \end{bmatrix} \rho|_{\Sigma_2^\circ} \begin{bmatrix} 1 & \frac{T-T_0}{\mu^{-1}-\mu} \\ 0 & 1 \end{bmatrix}^{-1}.$$

Note that we must have $\mu^{-1} - \mu \neq 0$ since otherwise $\operatorname{tr}\rho(c_2) = \pm 2$, which was precluded. By construction, we have

$$\rho''_T(\gamma_2) = \begin{bmatrix} 1 & \frac{T-T_0}{\mu^{-1}-\mu} \\ 0 & 1 \end{bmatrix} \rho(\gamma_2) \begin{bmatrix} 1 & \frac{T-T_0}{\mu^{-1}-\mu} \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \mu & T \\ 0 & \mu^{-1} \end{bmatrix},$$

so that ρ'''_T is well defined by our discussion in Section 2B3.

(3) Suppose c_2 is a nonseparating curve on Σ_3. Let δ be a simple based loop on Σ_3 intersecting c_2 exactly once transversally, oriented as in the first part of Section 2B3 (with the pair (γ_2, δ) playing the role of (α, β) there). Let γ'_2 be the based loop on Σ_2 (like α'_2 in Section 2B3) whose image in Σ_3 lies in the homotopy class of $\delta^{-1}\gamma_2\delta$. We define the representation ρ'''_T by specifying the pair $(\rho'''_T|_{\Sigma_2}, \rho'''_T(\delta))$ as in the discussion of Section 2B3, where

$$\rho'''_T|_{\Sigma_2} = \rho''_T, \quad \rho'''_T(\delta) = \begin{bmatrix} 1 & \frac{T-T_0}{\mu^{-1}-\mu} \\ 0 & 1 \end{bmatrix} \rho(\delta).$$

We have $\rho'''_T(\delta^{-1})\rho'''_T(\gamma_2)\rho'''_T(\delta) = \rho(\delta^{-1})\rho(\gamma_2)\rho(\delta) = \rho(\gamma'_2) = \rho'''_T(\gamma'_2)$ by our construction, so that ρ'''_T is well defined.

Finally, by the same procedure, we define $\rho_T : \pi_1\Sigma \to \mathrm{SL}_2(\mathbb{C})$ extending ρ_T''' such that $\rho_{T_0} = \rho$. The important point here is that, at each stage of the above "gluing" process, the monodromy matrices along the two curves being glued are conjugate by a matrix of the form

$$\begin{bmatrix} 1 & P(T) \\ 0 & 1 \end{bmatrix},$$

where $P(T) \in \mathbb{C}[T]$ is a polynomial in T. By testing the trace of the representations ρ^T along loops we see that the resulting morphism $\mathbb{A}^1 \to X_{k,t}^P$ given by $T \mapsto \rho_T$ must be nonconstant, unless $\rho = \rho_{T_0}$ is reducible.

It therefore remains to treat the case where ρ is reducible. Since ρ is semisimple by hypothesis, it is diagonal. Given $\Sigma' \subset \Sigma|P$ as above, let us choose a different component $\Sigma'' \subset \Sigma|P$ whose image in Σ has at least one boundary curve (say $c \subset P$) in common with the image of Σ' in Σ. Let Σ_1 be the surface of type $(0, 4)$ obtained by gluing together Σ' and Σ'' along the boundary curves corresponding to c. Let us choose our base point of Σ to be on c, and lift it to Σ_1. We have a one-parameter family of representations $\rho_T' : \pi_1\Sigma_1 \to \mathrm{SL}_2(\mathbb{C})$ whose monodromy along c is constant and is such that $\rho_T'|\Sigma' = \rho_T'$ (with suitable labeling of loops) and $\rho_T'|\Sigma''$ is a similarly constructed polynomial family of lower triangular representations. Note that the morphism $\mathbb{A}^1 \to X(\Sigma_1)$ given by $T \mapsto \rho_T'$ is nonconstant since $\mathrm{tr}_{\beta\gamma}(\rho_T)$ is nonconstant for any choice of boundary loops $\beta \in \pi_1\Sigma'$ and $\gamma \in \pi_1\Sigma''$ which remain boundary loops in Σ_1.

We then proceed as before with the "gluing" procedure to produce a family of representations $\rho_T : \pi_1\Sigma \to \mathrm{SL}_2(\mathbb{C})$ extending ρ_T', the important point being that, at each stage, the monodromy matrices along two curves being glued are conjugate by a matrix which is given by a product of matrices of the form

$$\begin{bmatrix} 1 & P(T) \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 \\ Q(T) & 1 \end{bmatrix}$$

with $P, Q \in \mathbb{C}[T]$. By construction, the morphism $\mathbb{A}^1 \to X_{k,t}^P$ given by $T \mapsto \rho_T$ will be nonconstant. This concludes the proof of Claim 3.5 □

Claim 3.5 implies that, if $X_{k,t}^P$ is imperfect, then it is parabolic. This implies the remaining half of Theorem 3.2(A), and concludes the proof of Theorem 3.2.

3C. ***Applications.*** We prove Theorem 1.1 as a corollary of Theorem 3.2.

Theorem 1.1. *A point $\rho \in X_k(\mathbb{C})$ is degenerate if and only if*

(1) (parabolic curve) *there is an essential simple closed curve $a \subset \Sigma$ such that* $\mathrm{tr}\,\rho(a) = \pm 2$*, or*

(2) (parabolic pants) $(g, n, k) \neq (1, 1, 2)$ *and there is a subsurface $\Sigma' \subset \Sigma$ of genus* 0 *with* 3 *boundary curves, each of which is an essential curve or a boundary curve of Σ, such that the restriction $\rho|\Sigma'$ is reducible.*

In particular, there is a parabolic proper closed subvariety Z of X_k such that every nonconstant morphism $\mathbb{A}^1 \to X_k$ over $\mathbb{C}$ is mapping class group equivalent to one with image in Z.

Proof. The first sentence of the theorem follows directly from Theorem 3.2 and Definition 3.1. For each pants decomposition P of Σ, the condition that $\rho \in X_k(\mathbb{C})$ lies in some parabolic fiber $X^P_{k,t}$ is evidently nontrivial and algebraic by Theorem 3.2 and the definition of perfect fibers. In particular, the union of all parabolic fibers of the form $X^P_{k,t}$ for fixed P is a proper closed parabolic algebraic subvariety of X_k. Since there are at most finitely many isotopy classes of pants decompositions of Σ up to mapping class group action, the last statement follows. □

We also record the following consequence of Theorem 3.2. By the uniformization theorem, given a marked hyperbolic structure σ on Σ with geodesic boundary curves there is a Fuchsian representation $\rho_\sigma : \pi_1\Sigma \to \mathrm{SL}_2(\mathbb{R})$ such that the quotient $\mathbb{H}^2/\rho_\sigma(\pi_1\Sigma)$ of the upper half plane contains (Σ, σ) as a Nielsen core.

Corollary 3.6. *Every one-parameter polynomial deformation $\rho_t : \pi_1\Sigma \to \mathrm{SL}_2(\mathbb{R})$ of a Fuchsian representation ρ_0 preserving the boundary traces is isotrivial.*

Proof. This is immediate from Theorem 3.2 and the observation that a Fuchsian representation ρ_0 does not lie in an imperfect fiber $X^P_{k,t}$ for any (P, t). □

4. Integral points on curves

4A. ***Nonintegrable curves.*** Let Σ be a surface of type (g, n) with $3g + n - 3 > 0$, and let $X_k = X_k(\Sigma)$ be a relative character variety of Σ. We shall prove Theorem 1.2 using Corollary 2.6. We first repeat our definition of integrable curves on X_k given in Section 1B.

Definition 4.1. An algebraic curve $C \subseteq X_k$ is *integrable* if there is a pants decomposition P of Σ with tr_P constant along C. Otherwise, C is *nonintegrable*.

As in Section 1B, given an algebraic curve $C \subset X_k$ and an arbitrary subset $A \subseteq \mathbb{C}$, let us denote

$$C(A) = \{\rho \in V(\mathbb{C}) : \mathrm{tr}_a(\rho) \in A \text{ for every essential curve } a \subset \Sigma\}.$$

Theorem 1.2. *Let $A \subset \mathbb{C}$ be a closed discrete subset. If $C \subset X_k$ is a nonintegrable geometrically irreducible algebraic curve, then $C(A)$ is finite, with effective bounds on sizes of the coordinates for any given embedding of C into affine space.*

Proof. Let F be the function field of C over $\mathbb{C}$. Let $\pi : C_0 \to C$ be the normalization of C, and let $\overline{C}_0$ be a smooth compactification of C_0. Let

$$\{p_1, \ldots, p_m\} = \overline{C}_0(\mathbb{C}) \setminus C_0(\mathbb{C})$$

be the points at infinity. For each p_i, we have a discrete valuation v_i on F given by order of vanishing at p_i. By Corollary 2.6 and our assumption on C, we deduce that there is an essential curve $a_i \subset \Sigma$ such that $v_i(\pi^*(\mathrm{tr}_{a_i})) \geq 0$, meaning in particular that $\pi^*(\mathrm{tr}_{a_i})$ is bounded on C_0 near the point p_i. By

our hypothesis that C is nonintegrable, we may further assume that each tr_{a_i} is nonconstant on C. In particular, for each $i = 1, \ldots, m$, setting $z_i = \mathrm{tr}_{a_i} \pi(p_i) \in \mathbb{C}$ the set

$$C(A) \cap \{\rho \in C(\mathbb{C}) : \mathrm{tr}\, \rho(a_i) = z_i\}$$

is finite. It therefore remains to consider

$$C(A) \setminus \bigcup_{i=1}^{m} \{\rho \in C(\mathbb{C}) : \mathrm{tr}\, \rho(a_i) = z_i\}.$$

But by the boundedness of each $\pi^* \mathrm{tr}_{a_i}$ near p_i and the discreteness of A, the above set lies in a compact subset of $C(\mathbb{C})$ (under the analytic topology). Again by the discreteness of A, this shows that $C(A)$ is finite, as desired.

To prove the last assertion, note that the desired curves a_i can in principle be found effectively by simply enumerating and going through the list of all isotopy classes of essential curves on Σ, with Corollary 2.6 guaranteeing the termination of this procedure. Once these functions are found, the above proof leads to the desired effective bounds on the sizes of points in $C(A)$. □

Theorem 1.2 yields a broad generalization and strengthening of the following result of [Long and Reid 2003]. Let M be a finite volume hyperbolic three-manifold with a single cusp. By the work of Thurston, its character variety $X(M)$ has an irreducible component $C(M)$, containing the faithful discrete representation of $\pi_1 M$, which is an algebraic curve. As in Section 1B, let O_d denote the ring of integers of the imaginary quadratic field $\mathbb{Q}(\sqrt{-d})$ for each squarefree integer $d > 0$.

Theorem [Long and Reid 2003]. *For M as above, the set $\bigcup_{d>0} C(M)(O_d)$ is finite.*

The inclusion of the cuspidal torus $T \to M$ induces a morphism from $C(M)$ into the so-called Cayley cubic algebraic surface $X(T)$, which is isomorphic to $X_2(\Sigma)$ where Σ is a surface of type $(1, 1)$. The crucial ingredient in the proof of the above [Long and Reid 2003, Lemma 3.4] is that the image of $C(M)$ in $X_2(\Sigma)$ is a nonintegrable curve. Given this, Theorem 1.2 readily recovers the above theorem on the integral points of $C(M)$, bypassing a somewhat involved arithmetical argument in [Long and Reid 2003].

4B. ***Application of Baker's theory.*** In this interlude, we demonstrate a result on certain lattice points lying on algebraic curves in algebraic tori. We shall obtain it as a straightforward consequence of Baker's theory on linear forms in logarithms. Let $\mathbb{G}_m$ denote the multiplicative group, and let $d \geq 1$ be an integer. Let $\overline{\mathbb{Q}} \subset \mathbb{C}$ be the field of algebraic numbers. Fix a sequence of real algebraic numbers

$$(z_1, \ldots, z_d) \in \mathbb{G}_m^d(\overline{\mathbb{Q}} \cap \mathbb{R})$$

with each $z_i \neq \pm 1$, and let $\Gamma \leq \mathbb{G}_m^d(\overline{\mathbb{Q}} \cap \mathbb{R})$ be the subgroup consisting of elements of the form $(z_1^{l_1}, \ldots, z_d^{l_d})$ with $l_i \in \mathbb{Z}$. Let $C \subset \mathbb{G}_m^d$ be an irreducible algebraic curve defined over $\overline{\mathbb{Q}}$, and fix $p \in \mathbb{G}_m^d(\overline{\mathbb{Q}})$. Our goal is to prove the following.

Proposition 4.2. *One of the following holds*:

(1) $C(\overline{\mathbb{Q}}) \cap (\Gamma \cdot p)$ *is finite, or*

(2) C *is a translate of an algebraic subtorus* $T \leq \mathbb{G}_m^d$ *defined over* $\overline{\mathbb{Q}}$ *and invariant under some nontrivial* $z \in \Gamma$.

Moreover, $C(\overline{\mathbb{Q}}) \cap (\Gamma \cdot p)$ *can be effectively determined.*

Proof. We may assume $p = (1, \ldots, 1)$ up to translation. We proceed by induction on $d \geq 1$. The claim is obvious if $d = 1$, so we may assume that $d \geq 2$. Assuming the result in the case $d = 2$, we shall first show below how the general case follows.

Suppose that $C(\overline{\mathbb{Q}}) \cap \Gamma$ is infinite. Up to rearranging the factors of $\mathbb{G}_m^d$, we may assume that the projection morphism $\pi : \mathbb{G}_m^d \to \mathbb{G}_m^{d-1}$ onto the first $d-1$ factors is nonconstant along C, since otherwise our claim is clear. Now, the Zariski closure C' of $\pi(C)$ in $\mathbb{G}_m^{d-1}$ contains infinitely many points lying in the group

$$\Gamma' = \{(z_1^{l_1}, \ldots, z_{d-1}^{l_{d-1}}) : l_i \in \mathbb{Z}\} \leq \mathbb{G}_m^{d-1}(\overline{\mathbb{Q}} \cap \mathbb{R}).$$

By inductive hypothesis, C' is a translate of an algebraic subtorus $T' \leq \mathbb{G}_m^{d-1}$, and C' is preserved by some nontrivial $z' \in \Gamma'$. Up to translation of C within $\mathbb{G}_m^d$, we may assume that $C' = T'$. Let us also denote by $z' \in \mathbb{G}_m(\overline{\mathbb{Q}} \cap \mathbb{R})$ the element corresponding to z' under the identification $T' \simeq \mathbb{G}_m$. Applying the case $d = 2$ of the proposition to the immersion $C \hookrightarrow T' \times \mathbb{G}_m \subset \mathbb{G}_m^d$, we see that C is a translate of an algebraic subtorus in $T' \times \mathbb{G}_m$ which is invariant under some nontrivial element of $\langle z' \rangle \times \langle z_d \rangle \leq \Gamma$. This gives the desired result.

It remains to prove the proposition in the case $d = 2$. We begin by making a number of simplifying remarks. First, it suffices to prove the proposition with Γ replaced by the monoid

$$\{(z_1^{l_1}, z_2^{l_2}) : l_i \in \mathbb{Z}_{\geq 0}\},$$

since Γ is a union of finitely many monoids of the above type (obtained by replacing (z_1, z_2) with $(z_1^{\pm 1}, z_2^{\pm 1})$). Similarly, it suffices to treat the case where $|z_1|, |z_2| > 1$, since we reduce to this case by applying inversions to factors of $\mathbb{G}_m^2$ appropriately.

Let $f \in \overline{\mathbb{Q}}[X, Y]$ be an irreducible polynomial defining the Zariski closure of C in $\mathbb{A}^2$ under the obvious embedding $\mathbb{G}_m^2 \to \mathbb{A}^2$. In what follows, we shall write $(x, y) = (z_1, z_2)$ for convenience, so that in particular $|x|, |y| > 1$ by our assumption above. Assuming that the set

$$S = \{(m, n) \in \mathbb{Z}_{\geq 0}^2 : f(x^m, y^n) = 0\}$$

is infinite, we shall show that C is a translate of algebraic torus and is preserved by some nontrivial $z \in \Gamma$. Let us write

$$f(X, Y) = \sum_{i=1}^{r} a_i X^{d_i} Y^{e_i}$$

with nonzero $a_i \in \overline{\mathbb{Q}}$ and $d_i, e_i \in \mathbb{Z}_{\geq 0}$, such that $(d_i, e_i) \neq (d_j, e_j)$ whenever $i \neq j$. Note that the number r of terms in the above sum is at least 2 since $x, y \neq 0$. Upon relabeling the terms and passing to an infinite

subset of S, we may assume that

$$(r-1)|a_2x^{d_2m}y^{e_2n}| \geq |a_1x^{d_1m}y^{e_1n}| \geq \cdots \geq |a_rx^{d_rm}y^{e_rn}|$$

for all $(m,n) \in S$. Let $q = \gcd(d_1-d_2, e_1-e_2) > 0$, and define $d, e \geq 0$ to be coprime integers such that $|d_1-d_2| = qd$ and $|e_1-e_2| = qe$. The above inequalities imply that there is a constant $K \geq 1$ with

$$K^{-1} \leq \left|\frac{(x^m)^d}{(y^n)^e}\right| \leq K \tag{$*$}$$

for all $(m,n) \in S$. Assigning weights e and d to the variables X and Y respectively, let us write

$$f(X,Y) = g(X,Y) + h(X,Y),$$

where g is the top degree weighted homogeneous part of f, and the remainder h is a polynomial of lower degree. We remark that g must include the term $a_1X^{d_1}Y^{e_1}$. Indeed, for any $i = 1, \ldots, r$ we must have

$$|a_i|K^{-d_i/d}|y^n|^{(d_ie+de_i)/d} \leq |a_i(x^m)^{d_i}(y^n)^{e_i}| \leq |a_1(x^m)^{d_1}(y^n)^{e_1}| \leq |a_1|K^{d_1/d}|y^n|^{(d_1e+de_1)/d}.$$

Thus, if $d_1e + de_1 < d_ie + de_i$, then $|y|^n$ must be bounded for every $(m,n) \in S$, and by a similar argument $|x|^m$ must be bounded for every $(m,n) \in S$, which is a contradiction. We also have

$$\begin{aligned}\deg(a_2X^{d_2}Y^{e_2}) &= d_2e + de_2 = (d_2-d_1)e + d_1e + d(e_2-e_1) + de_1 \\ &= (qde - qde) + (d_1e + de_1) = \deg(a_1X^{d_1}Y^{e_1})\end{aligned}$$

by definition, so g also includes the term $a_2X^{d_2}Y^{e_2}$. In fact, we see that $a_iX^{d_i}Y^{e_i}$ is a term in g if and only if $(d_i, e_i) = (d_1 + q'd, e_1 - q'e)$ for some $q' \in \mathbb{Z}$. This shows that we can factorize $g(X,Y)$ as

$$g(X,Y) = cX^{\lambda}Y^{\mu}\prod_j(\alpha_jX^d - Y^e)^{\nu_j}$$

for some nonzero $c \in \overline{\mathbb{Q}}$ and pairwise distinct nonzero $\alpha_j \in \overline{\mathbb{Q}}$. Up to relabeling and passing to an infinite subset of S, there exists some $\kappa > 0$ with

$$\begin{aligned}\left|\alpha_1\frac{(x^m)^d}{(y^n)^e} - 1\right|^{\nu_1} &= \frac{|g(x^m,y^n)|}{|c(x^m)^{\lambda}(y^n)^{\mu}\prod_{j\neq 1}(\alpha_jx^{md} - y^{ne})^{\nu_j}|}\frac{1}{|y^{ne}|^{\nu_1}} \\ &= \frac{1}{|c(x^m/y^n)^{\lambda}|\prod_{j\neq 1}|\alpha_j(x^{md}/y^{ne}) - 1|^{\nu_j}}\frac{|h(x^m,y^n)|}{|y^{ne}|^{(\lambda+\mu)/e+\sum_j\nu_j}} \\ &\ll \frac{1}{|y^{ne}|^{\kappa}} \ll \frac{1}{\max\{|x^d|^m, |y^e|^n\}^{\kappa}}\end{aligned}$$

for every $(m,n) \in S$. Here, we have used the inequality $(*)$. We remark that we must have $\alpha_1 \in \overline{\mathbb{Q}} \cap \mathbb{R}$.

In particular, there exists $M > 1$ such that

$$|\alpha_1| \frac{|x^d|^m}{|y^e|^n} = 1 + O\left(\frac{1}{M^{\max\{m,n\}}}\right)$$

for every $(m, n) \in S$. Taking logarithms and using the fact that $|\log(1+\xi)| \le 2|\xi|$ for every $|\xi| \le 1/2$, upon passing to a suitable infinite subset of S we have

$$\left|\log|\alpha_1| + m\log|x^d| - n\log|y^e|\right| \ll \frac{1}{M^{\max\{m,n\}}}.$$

By Baker's theorem [1990, Theorem 3.1], if $\log|\alpha_1|$, $\log|x^d|$, and $\log|y^e|$ were $\mathbb{Q}$-linearly independent, then we would have

$$\left|\log|\alpha_1| + m\log|x^d| - n\log|y^e|\right| \gg \max\{m, n\}^{-D}$$

for some effective constant $D > 0$. Therefore, what we have obtained shows that x and y must be multiplicatively dependent. Let $v > 1$ be the unique real algebraic number such that $(|x|, |y|) = (v^a, v^b)$ for positive coprime integers $a, b \in \mathbb{Z}$. We may rewrite $(*)$ as

$$K^{-1} \le v^{adm-ben} \le K.$$

This shows that $adm - ben$ takes at most finitely many values, and so there is some $t \in \mathbb{Z}$ such that, up to passing to an infinite subset of S, we have $adm - ben = t$ for all $(m, n) \in S$. Fix $(m_0, n_0) \in S$. We then have

$$ad(m - m_0) = be(n - n_0)$$

for every $(m, n) \in S$. Since ad and be are coprime, for each $(m, n) \in S$ we have $N = N_{m,n} \in \mathbb{Z}$ with $(m, n) = (m_0, n_0) + N \cdot (be, ad)$. Consider now the morphism

$$\Phi : \mathbb{A}^1 \to \mathbb{A}^2$$

given by $\Phi(T) = (x^{m_0} T^e, y^{n_0} T^d)$. Note that Φ restricts to a morphism $\mathbb{G}_m \to \mathbb{G}_m^2$ so that the image of Φ is a translation of an algebraic torus. Furthermore, given $(m, n) \in S$ and $N = N_{m,n}$ as above, we have

$$\begin{aligned}\Phi(v^{Nab}) &= (x^{m_0} v^{Nabe}, y^{n_0} v^{Nabd}) = (x^{m_0}(v^{abe})^N, y^{m_0}(v^{abd})^N) \\ &= (x^{m_0} x^{beN}, y^{n_0} y^{adN}) = (x^m, y^n).\end{aligned}$$

This shows that the Zariski closure of C has infinitely many points in common with the image of Φ. This implies that $\Phi(\mathbb{G}_m) = C$. This proves Proposition 4.2. □

4C. ***Integrable curves.*** Fix a pants decomposition $P \subset \Sigma$. Let Γ_P be the free abelian subgroup of rank $3g + n - 3$ in the mapping class group $\Gamma(\Sigma)$ generated by Dehn twists along the curves in P. The action of Γ_P on X_k preserves the fiber $X_{k,t}^P$ for each $t \in X(P, \mathbb{C})$. For fixed $t = (t_1, \ldots, t_{3g+n-3})$, let us fix a sequence $(z_1, \ldots, z_{3g+n-3}) \in (\mathbb{C}^\times)^{3g+n-3}$ of complex numbers such that

$$z_i + z_i^{-1} = t_i \quad \text{for all} \quad i = 1, \ldots, 3g + n - 3.$$

Let Γ_z be the subgroup of $\mathbb{G}_m^{3g+n-3}(\mathbb{C})$ generated by translations by z_i (in the multiplicative sense) in the i-th coordinate. Our discussion in Section 2B3 leads us to the following result.

Proposition 4.3. *If $X_{k,t}^P$ is perfect, then there is a morphism*

$$F : \mathbb{G}_m^{3g+n-3} \to X_{k,t}^P$$

of schemes (defined over $\overline{\mathbb{Q}}$ if k and t are algebraic) satisfying the following:

(1) *At the level of complex points, F is surjective with finite fibers.*

(2) *The action of Γ_z on $\mathbb{G}_m^{3g+n-3}$ lifts the Γ_P-action on $X_{k,t}^P$.*

Proof. We shall describe the map induced by F on the complex points. It will be conceptually clear from our construction, even if laborious to show, that the map is induced from a morphism of schemes, and is moreover defined over $\overline{\mathbb{Q}}$ provided that k and t are algebraic.

To construct F, we first fix an $\mathrm{SL}_2(\mathbb{C})$-local system ρ_0 on Σ whose class lies in the fiber $X_{k,t}^P$. Let us write $P = a_1 \sqcup \cdots \sqcup a_{3g+n-3}$ with each a_i a curve, on which we fix a base point $x_i \in a_i$. Let α_i be a choice of a simple loop based at x_i parametrizing a_i. We fix a trivialization of the fiber of ρ above each x_i so that the monodromy along α_i is given by a diagonal matrix of the form

$$\begin{bmatrix} z_i & 0 \\ 0 & z_i^{-1} \end{bmatrix} \in \mathrm{SL}_2(\mathbb{C}). \tag{$*$}$$

This is possible since $t_i \neq \pm 2$ by our hypothesis that $X_{k,t}^P$ is perfect. Setting aside the case $(g, n, k) = (1, 1, 2)$ which is elementary, this hypothesis also implies that the restriction of ρ_0 to each connected component of $\Sigma | P$ is irreducible, and in fact determines the isomorphism type of such restriction for every local system ρ whose class lies in $X_{k,t}^P$.

For each $i = 1, \ldots, 3g+n-3$, let us denote by a_i' and a_i'' the two boundary curves on $\Sigma | P$ corresponding to a_i, and let (x_i', α_i') and (x_i'', α_i'') be the corresponding lifts of (x_i, α_i), respectively. We shall assume that we have chosen the labelings so that the interior of $\Sigma | a$ lies on the left as one travels along α_i'. The above observation shows that any local system with class in $X_{k,t}^P$ is determined by the isomorphisms of local systems $\rho_0|_{a_i'}$ and $\rho_0|_{a_i''}$ compatible with the gluing of a_i' and a_i'' in Σ. For instance, ρ_0 itself is the local system determined by the identity isomorphisms

$$\mathrm{id} : \rho_0|_{a_i'} \simeq \rho_0|_{a_i} \simeq \rho_0|_{a_i''}.$$

More generally, an isomorphism of local systems $\rho_0|_{a_i'} \simeq \rho_0|_{a_i''}$ is specified by an element in the centralizer of the matrix $(*)$ above (namely, the group of diagonal matrices of determinant one). Thus, we have a map

$$F : \mathbb{G}_m^{3g+n-3}(\mathbb{C}) \to X_{k,t}^P(\mathbb{C})$$

sending $(v_1, \ldots, v_{3g+n-3}) \in \mathbb{G}_m^{3g+n-3}(\mathbb{C})$ to the class of the local system determined by the isomorphisms

$$\begin{bmatrix} v_i & 0 \\ 0 & v_i^{-1} \end{bmatrix} : \rho_0|_{a_i'} \simeq \rho_0|_{a_i''}.$$

The fact that this is surjective follows from the above discussion. To see that F has finite fibers, note that if we glue the local systems $\rho_0|_{\Sigma'}$ on the components of Σ' of $\Sigma|P$ along the curves $a_1, \ldots, a_{3g+n-3}$ in a fixed order, at each stage the resulting local system is uniquely determined, except possibly when a_i is a separating curve in which case there is a double ambiguity.

Finally, the compatibility of the action of Γ_z with the action of Γ_P under F follows from the our description of lifts of Dehn twists in Section 2B3. □

Corollary 4.4. *Let $X^P_{k,t}$ be a perfect fiber. Then*

(1) $|\Gamma_P \backslash X^P_{k,t}(A)| < \infty$ *for any closed discrete $A \subset \mathbb{R}$, and*

(2) *if no coordinate of $t \in X(P, \mathbb{C}) \simeq \mathbb{A}^{3g+n-3}$ lies in $[-2, 2]$, then*

$$|\Gamma_P \backslash X^P_{k,t}(A)| < \infty$$

for any closed discrete $A \subset \mathbb{C}$.

Proof. Let us write $t = (t_1, \ldots, t_{3g+n-3})$, and let us first suppose $t_i \notin [-2, 2]$ for each $i = 1, \ldots, 3g+n-3$. This shows in particular that each z_i (defined above so that $z_i + z_i^{-1} = t_i$) has absolute value different from 1. In particular, every point in $\mathbb{G}_m^{3g+n-3}(\mathbb{C})$ is Γ_z-equivalent to a point in a region $K \subset \mathbb{G}_m^{3g+n-3}(\mathbb{C})$ which is compact with respect to the Euclidean topology. Under the map

$$F : \mathbb{G}_m^{3g+n-3}(\mathbb{C}) \to X^P_{k,t}(\mathbb{C})$$

constructed in Proposition 4.3, the image of K in $X^P_{k,t}(\mathbb{C})$ is compact and hence has finite intersection with any closed discrete $A \subset \mathbb{C}$. The equivariance property of F proved in Proposition 4.3 then implies our claim, when $t_i \notin [-2, 2]$ for each $i = 1, \ldots, 3g + n - 3$.

Let us now assume that A is a closed discrete subset of $\mathbb{R}$. Note in particular that the coordinates t_i may be assumed to lie in A and hence are real, since otherwise $X^P_{k,t}(A)$ is empty. Now, let us consider the natural morphism

$$X^P_{k,t}(\Sigma) \to \prod_{i=1}^{3g+n-3} X^{a_i}_{k_i,t_i}(\Sigma_i),$$

where each Σ_i is the surface of type $(0, 4)$ or $(1, 1)$ obtained by gluing together the two boundary curves on $\Sigma|P$ corresponding to a_i, and the boundary traces k_i are appropriately determined from k, P, t, and Σ_i. The fact that the map F constructed in Proposition 4.3 has finite fibers implies that the above morphism also has finite fibers (at the level of complex points). The claim to be proved thus reduces to the case where Σ is of type $(0, 4)$ or $(1, 1)$. But this is obvious by elementary geometric considerations (see also the proof of Theorem 1.4 in [Whang 2017]). □

Proposition 4.5. *Let C be a geometrically irreducible algebraic curve over $\mathbb{Z}$ lying in a perfect fiber $X^P_{k,t}$. Then $C(\mathbb{Z})$ can be effectively determined, and*

(1) *$C(\mathbb{Z})$ is finite, or*

(2) *$C(\mathbb{Z})$ is finitely generated under some nontrivial $\gamma \in \Gamma_P$ preserving C.*

If moreover C is not fixed pointwise by any nontrivial $\gamma \in \Gamma_P$, the same result holds with $C(\mathbb{Z})$ replaced by the set of all imaginary quadratic integral points on C.

Proof. Note that $t \in X(P, \mathbb{Z}) \simeq \mathbb{Z}^{3g+n-3}$. We shall first consider the case where no coordinate of t lies in $[-2, 2]$. By Corollary 4.4, the set $\bigcup_{d>0} X^P_{k,t}(O_d)$ consists of finitely many Γ_P-orbits. Thus, it suffices to consider the intersection of $C(\mathbb{C})$ with the orbit of a single point $\rho \in \bigcup_{d>0} X^P_{k,t}(O_d)$.

Let $F : \mathbb{G}_m^{3g+n-3} \to X^P_{k,t}$ be a Γ_P-equivariant morphism as constructed in the proof of Proposition 4.3. Let $z = (z_1, \ldots, z_{3g+n-3}) \in \mathbb{G}_m^{3g+n-3}$ be as defined earlier in this subsection. By our hypothesis on t, each z_i is a real quadratic integer. Let $C' = F^{-1}(C)$, and choose a point $p \in F^{-1}(\rho)$. Applying Proposition 4.2 to the curve $C' \subset \mathbb{G}_m^{3g+n-3}$ and projecting the result down to $X^P_{k,t}$, we obtain the result.

Let us write $P = a_1 \sqcup \cdots \sqcup a_{3g+n-3}$, and let us denote by $t_i \in \mathbb{Z}$ the component of t corresponding to a_i. Based on the above argument, it remains to consider the case where $t_i \in [-2, 2]$ for some i. Since we must have $t_i \in \{0, \pm 1\}$, we see that the Dehn twist τ_{a_i} acts on the fiber $X^P_{k,t}$ with finite order (as seen from the discussion in Section 2C), so we need only to consider $C(\mathbb{Z})$. Let Σ_i be the surface of type $(0, 4)$ or $(1, 1)$ obtained by gluing together the two boundary curves on $\Sigma|P$ corresponding to a_i, and consider the composition of morphisms

$$C \to X^P_{k,t}(\Sigma) \to X^{a_i}_{k_i,t_i}(\Sigma_i),$$

where $k_i \in X(\partial\Sigma_i, \mathbb{C})$ is appropriately determined from k, t, and Σ_i. Note that the set of real points of $X^{a_i}_{k_i,t_i}(\Sigma_i)$ defines a ellipse in an appropriate coordinate plane, as seen from our discussion in Section 2B3. In particular, $X^{a_i}_{k_i,t_i}(\Sigma_i, \mathbb{Z})$ is finite, and if the above composition is nonconstant then we find that $C(\mathbb{Z})$ is finite, as desired. It thus remains to consider the case where the above composition is constant. This implies that the morphism

$$C \to X^P_{k,t}(\Sigma) \to X^{P'}_{k',t'}(\Sigma|a_i)$$

(where k', P', and t' are appropriately determined from k, P, t, and a_i) must be nonconstant, as seen from the consideration of the morphism $F : \mathbb{G}_m^{3g+n-3} \to X^P_{k,t}$ as constructed in the proof of Proposition 4.3. Thus, we may apply induction and a straightforward modification of the previous paragraph to conclude the result. □

4D. ***Proofs of Theorem 1.3 and Corollary 1.4.*** We obtain Theorem 1.3 by combining Theorem 1.2 for nonintegrable curves with Proposition 4.5 for integrable curves in X_k. Finally, Corollary 1.4 follows easily from Theorem 1.3 (and indeed from Theorem 1.2) and our understanding of the fibers $X^P_{k,t}$ from Section 3A.

Acknowledgements

This work was done as part of the author's Ph.D. thesis at Princeton University. I thank my advisor Peter Sarnak and Phillip Griffiths for their guidance, encouragement, and insightful discussions. I also thank Rafael von Känel for useful historical remarks, and thank the referee for suggesting numerous improvements to the paper.

References

[Baker 1990] A. Baker, *Transcendental number theory*, 2nd ed., Cambridge Univ. Press, 1990. MR Zbl

[Borevich and Shafarevich 1966] A. I. Borevich and I. R. Shafarevich, *Number theory*, Pure and Applied Mathematics **20**, Academic Press, New York, 1966. MR Zbl

[Faltings 1991] G. Faltings, "Diophantine approximation on abelian varieties", *Ann. of Math.* (2) **133**:3 (1991), 549–576. MR Zbl

[Farb and Margalit 2012] B. Farb and D. Margalit, *A primer on mapping class groups*, Princeton Math. Series **49**, Princeton Univ. Press, 2012. MR Zbl

[Ghosh and Sarnak 2017] A. Ghosh and P. Sarnak, "Integral points on Markoff type cubic surfaces", 2017. arXiv

[Goldman 1997] W. M. Goldman, "Ergodic theory on moduli spaces", *Ann. of Math.* (2) **146**:3 (1997), 475–507. MR Zbl

[Goldman 2009] W. M. Goldman, "Trace coordinates on Fricke spaces of some simple hyperbolic surfaces", pp. 611–684 in *Handbook of Teichmüller theory, II*, edited by A. Papadopoulos, IRMA Lect. Math. Theor. Phys. **13**, Eur. Math. Soc., Zürich, 2009. MR Zbl

[Goldman and Xia 2011] W. M. Goldman and E. Z. Xia, "Ergodicity of mapping class group actions on SU(2)-character varieties", pp. 591–608 in *Geometry, rigidity, and group actions*, edited by B. Farb and D. Fisher, Univ. Chicago Press, 2011. MR Zbl

[Horowitz 1972] R. D. Horowitz, "Characters of free groups represented in the two-dimensional special linear group", *Comm. Pure Appl. Math.* **25** (1972), 635–649. MR Zbl

[Lang 2002] S. Lang, *Algebra*, 3rd ed., Graduate Texts in Mathematics **211**, Springer, 2002. MR Zbl

[Long and Reid 2003] D. D. Long and A. W. Reid, "Integral points on character varieties", *Math. Ann.* **325**:2 (2003), 299–321. MR Zbl

[Markoff 1880] A. Markoff, "Sur les formes quadratiques binaires indéfinies", *Math. Ann.* **17**:3 (1880), 379–399. MR Zbl

[Procesi 1976] C. Procesi, "The invariant theory of $n \times n$ matrices", *Advances in Math.* **19**:3 (1976), 306–381. MR Zbl

[Przytycki and Sikora 2000] J. H. Przytycki and A. S. Sikora, "On skein algebras and $\mathrm{Sl}_2(\mathbf{C})$-character varieties", *Topology* **39**:1 (2000), 115–148. MR Zbl

[Saito 1996] K. Saito, "Character variety of representations of a finitely generated group in SL_2", pp. 253–264 in *Topology and Teichmüller spaces*, edited by S. Kojima et al., World Sci. Publ., River Edge, NJ, 1996. MR Zbl

[Simpson 1994] C. T. Simpson, "Moduli of representations of the fundamental group of a smooth projective variety. II", *Inst. Hautes Études Sci. Publ. Math.* 80 (1994), 5–79. MR Zbl

[Sterk 1985] H. Sterk, "Finiteness results for algebraic $K3$ surfaces", *Math. Z.* **189**:4 (1985), 507–513. MR Zbl

[Vogt 1889] H. Vogt, "Sur les invariants fondamentaux des équations différentielles linéaires du second ordre", *Ann. Sci. École Norm. Sup.* (3) **6** (1889), 3–71. MR Zbl

[Vojta 1991] P. Vojta, "Siegel's theorem in the compact case", *Ann. of Math.* (2) **133**:3 (1991), 509–548. MR Zbl

[Vojta 1996] P. Vojta, "Integral points on subvarieties of semiabelian varieties, I", *Invent. Math.* **126**:1 (1996), 133–181. MR Zbl

[Whang 2017] J. P. Whang, "Nonlinear descent on moduli of local systems", 2017. arXiv

[Whang 2020] J. P. Whang, "Global geometry on moduli of local systems for surfaces with boundary", *Compos. Math.* **156**:8 (2020), 1517–1559. MR

[Zannier 2009] U. Zannier, *Lecture notes on Diophantine analysis*, Appunti. Scuola Normale Superiore di Pisa (Nuova Serie) [Lecture Notes. Scuola Normale Superiore di Pisa (New Series)] **8**, Ed. Norm., Pisa, 2009. MR Zbl

Communicated by Joseph H. Silverman
Received 2018-06-29 Revised 2020-07-20 Accepted 2020-08-21

jwhang@mit.edu *Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, United States*

ALGEBRA AND NUMBER THEORY 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2607

Curtis homomorphisms and the integral Bernstein center for GL_n

David Helm

We describe two conjectures, one strictly stronger than the other, that give descriptions of the integral Bernstein center for $\mathrm{GL}_n(F)$ (that is, the center of the category of smooth $W(k)[\mathrm{GL}_n(F)]$-modules, for F a p-adic field and k an algebraically closed field of characteristic ℓ different from p) in terms of Galois theory. Moreover, we show that the weak version of the conjecture (for $m \leq n$), together with the strong version of the conjecture for $m < n$, implies the strong conjecture for GL_n. In a companion paper (*Invent. Math.* **214**:2 (2018), 999–1022) we show that the strong conjecture for $n-1$ implies the weak conjecture for n; thus the two papers together give an inductive proof of both conjectures. The upshot is a description of the Bernstein center in purely Galois theoretic terms; previous work of the author shows that this description implies the conjectural "local Langlands correspondence in families" of (*Ann. Sci. Éc. Norm. Supér.* (4) **47**:4 (2014), 655–722).

1. Introduction

Emerton and the author [Emerton and Helm 2014] described a conjectural "local Langlands correspondence in families" for the group $\mathrm{GL}_n(F)$, where F is a p-adic field. More precisely, we showed that given a suitable coefficient ring A (in particular complete and local with residue characteristic ℓ different from p), and a family of Galois representations $\rho : G_F \to \mathrm{GL}_n(A)$, there is, up to isomorphism, at most one admissible $A[\mathrm{GL}_n(F)]$-module $\pi(\rho)$ that "interpolates the local Langlands correspondence across the family ρ" and satisfies certain technical hypotheses. (We refer the reader to [Emerton and Helm 2014, Theorem 1.1.1] for the precise result.) We further conjecture that such a representation $\pi(\rho)$ exists for any ρ.

The paper [Helm 2016b] gives an approach to the question of actually constructing $\pi(\rho)$ from ρ. The key new idea is the introduction of the integral Bernstein center, which is by definition the center of the category of smooth $W(k)[\mathrm{GL}_n(F)]$-modules. More prosaically, the integral Bernstein center is a ring Z that acts on every smooth $W(k)[\mathrm{GL}_n(F)]$-module, compatibly with every morphism between such modules, and is the universal such ring. The structure of Z encodes deep information about "congruences" between $W(k)[\mathrm{GL}_n(F)]$-modules (for instance, if two irreducible representations of $\mathrm{GL}_n(F)$ in characteristic zero become isomorphic modulo ℓ, the action of Z on these two representations will be via scalars that are congruent modulo ℓ.)

MSC2010: primary 11F33; secondary 11F70, 22E50.
Keywords: Langlands correspondence, modular representation theory, p-adic groups.

Morally, the problem of showing that $\pi(\rho)$ exists for all ρ amounts to showing — for a sufficiently general notion of "congruence" — that whenever there is a congruence between two representations of G_F, there is a corresponding congruence on the other side of the local Langlands correspondence. It is therefore not surprising that one can rephrase the problem of constructing $\pi(\rho)$ in terms of the structure of Z. Indeed, Theorem 7.4 of [Helm 2016b] reduces the question of the existence of $\pi(\rho)$ to a conjectured relationship between the ring Z and the deformation theory of mod ℓ representations of G_F (Conjecture 7.2 of [Helm 2016b]).

The primary goal of this paper, together with its companion paper [Helm and Moss 2018], is to prove a version of this conjecture, and thus establish the local Langlands correspondence in families. More precisely, we introduce a collection of finite type $W(k)$-algebras R_ν that parametrize representations of the Weil group W_F of F with fixed restriction to prime-to-ℓ inertia, and whose completion at a given maximal ideal is a close variant of a universal framed deformation ring. We then conjecture that there is a map $Z \to R_\nu$ that is "compatible with local Langlands" in a certain technical sense (see Conjecture 9.2 below for a precise statement and discussion.) This conjecture, which we will henceforth call the "weak conjecture", becomes Conjecture 7.2 of [Helm 2016b] after one completes R_ν at a maximal ideal, and hence implies both that conjecture and the existence of $\pi(\rho)$.

If a map $Z \to R_\nu$ of the conjectured sort exists it is natural to ask what the image is. The "strong conjecture" (Conjecture 9.3 below) gives a description of this image (and in fact gives a description of the direct factors of Z in purely Galois-theoretic terms.) As the names suggest, the "strong conjecture" implies the "weak conjecture."

The main result of this paper is that if the weak conjecture holds for all $\mathrm{GL}_m(F)$, with m less than or equal to a fixed n, and the strong conjecture holds for $m < n$, then the strong conjecture holds as well for the group $\mathrm{GL}_n(F)$. In the companion paper [Helm and Moss 2018], we show that the strong conjecture for $\mathrm{GL}_{n-1}(F)$ implies the weak conjecture for $\mathrm{GL}_n(F)$. Since the case $n = 1$ is easy (it is a consequence of local class field theory), the two papers together will establish both conjectures for all n, and hence the local Langlands correspondence for GL_n in families.

Our approach relies on three main ingredients. The first is an input from finite group theory, namely the endomorphism ring of the Gelfand–Graev representation $\overline{\Gamma}$ of $\mathrm{GL}_n(\mathbb{F}_q)$. In Section 2 we introduce this ring and describe some of its basic properties, following Bonnafé and Kessar [2008]. A crucial structure on this endomorphism ring is its canonical symmetrizing form, which Bonnafé and Kessar describe in terms of "Curtis homomorphisms" arising from Deligne–Lusztig restriction. In Section 3 we describe the connection between this endomorphism ring and the ring Z.

The second key ingredient is the behavior of the integral Bernstein center Z with respect to parabolic induction; for a Levi M of G there are natural maps $Z \to Z_M$ compatible, in a certain sense, with parabolic induction from M to G. In Section 3 we recall results of [Helm 2016a] (see Theorems 3.9 and 3.12, below) that say that in certain key cases the images of these maps are "large" in a certain sense, and that the failure of these maps to have image that is "as large as possible" is controlled by the endomorphism ring of a Gelfand–Graev representation.

The third key ingredient is the construction of the rings R_ν which occupies Sections 4, 7, and 8. These moduli spaces admit maps between them coming from taking direct sums of representations; these maps serve a purpose analogous to the "parabolic induction" maps from Z to Z_M. The functions on such spaces also admit subalgebras $B_{q,n}$ that play a role analogous to the subalgebras of Z arising from the endomorphism ring $\overline{E}_{q,n}$ of a Gelfand–Graev representation. The strong conjecture leads us to expect that in fact $\overline{E}_{q,n}$ and $B_{q,n}$ are isomorphic, but it seems difficult to show this directly (although it is easy to show if one inverts ℓ). Instead, we make use of the symmetrizing form on $\overline{E}_{q,n}$ to show that *if* there exists a map from $\overline{E}_{q,n}$ to $B_{q,n}$ then it must be an isomorphism (see Sections 5 and 6.)

Once we have established this, our argument goes as follows. First we show that the strong conjecture holds after inverting ℓ; this essentially follows easily from the classical Bernstein–Deligne theory of the Bernstein center over algebraically closed fields. We then assume the strong conjecture for $m < n$, and the weak conjecture for $m \leq n$. This gives us in particular a map $\overline{E}_{q,n} \to B_{q,n}$ that is necessarily an isomorphism. Using this, and considering various parabolic restriction maps from Z to various Levi subgroups, together with the corresponding maps on the rings R_ν of representations of W_F, we show, using our "large image" results for Z, that Z must "fill out" the entire ring of invariant functions in R_ν, thus proving the strong conjecture for GL_n.

In the process of carrying out this inductive argument we prove that $\overline{E}_{q,n}$ is isomorphic to $B_{q,n}$ for all n. This is a statement purely in finite group theory that is of independent interest. We know of no more direct proof of this isomorphism than the one described here.

Throughout this paper we adopt the following conventions: F is a p-adic field with residue field $\mathbb{F}_q$, k is an algebraically closed field of characteristic $\ell \neq p$, $\mathcal{K}$ is the field of fractions of $W(k)$, and $\overline{\mathcal{K}}$ is an algebraic closure of $\mathcal{K}$. Algebraic groups over F with be denoted by uppercase calligraphic letters $\mathcal{T}$, $\mathcal{G}$, etc.; for any such group the corresponding uppercase letters T, G, etc. will denote the groups of F-points of $\mathcal{T}$, $\mathcal{G}$, and so forth. In particular there is an implicit dependence of T on $\mathcal{T}$.

2. Finite groups

Before beginning our study of the Bernstein center we develop some finite group theory that will be essential for our approach. Most of the ideas in this section originally appear in [Bonnafé and Kessar 2008].

Fix distinct primes p and ℓ, and a power q of p. Let $\overline{\mathcal{G}}$ be the group GL_n over $\mathbb{F}_q$, and let $\overline{G} = \overline{\mathcal{G}}(\mathbb{F}_q)$. We will consider the representation theory of $\overline{G}$ over the Witt ring $W(k)$, where k is an algebraic closure of $\mathbb{F}_\ell$. Let $\mathcal{K}$ be the field of fractions of $W(k)$, and fix an algebraic closure $\overline{\mathcal{K}}$ of $\mathcal{K}$.

Our principal object of study in this section will be the Gelfand–Graev representation $\overline{\Gamma}$ of $\overline{G}$, with coefficients in $W(k)$. Fix a Borel $\overline{\mathcal{B}}$ in $\overline{\mathcal{G}}$, with unipotent radical $\overline{\mathcal{U}}$, and let $\overline{B}$, $\overline{U}$ denote the $\mathbb{F}_q$-points of $\overline{\mathcal{B}}$ and $\overline{\mathcal{U}}$ respectively. Also fix a generic character $\Psi : \overline{U} \to W(k)^\times$. Then, by definition, we have $\overline{\Gamma} = \text{c-Ind}_{\overline{U}}^{\overline{G}} \Psi$, where Ψ is considered as a $W(k)[\overline{U}]$-module that is free over $W(k)$ of rank one, with the appropriate action of $\overline{U}$. The module $\overline{\Gamma}$ is then independent of the choice of Ψ, up to isomorphism.

The objective of this first section is to study the endomorphism ring $\mathrm{End}_{W(k)[\overline{G}]}(\overline{\Gamma})$, which we denote by $\overline{E}_{q,n}$. Our main tool for doing so will be the Deligne–Lusztig induction and restriction functors of

[Bonnafé and Rouquier 2003]. Let $\bar{L}$ be the subgroup of $\bar{G}$ consisting of the $\bar{\mathbb{F}}_q$-points of a (not necessarily split) Levi subgroup $\bar{\mathcal{L}}$ of GL_n, and choose a parabolic subgroup $\bar{\mathcal{P}}$ of GL_n whose Levi subgroup is $\bar{\mathcal{L}}$. Let $\mathrm{Rep}_{W(k)}(\bar{G})$ and $\mathrm{Rep}_{W(k)}(\bar{L})$ denote the categories of $W(k)[\bar{G}]$-modules and $W(k)[\bar{L}]$-modules, respectively. Then Deligne–Lusztig induction and restriction are functors:

$$i_{\bar{\mathcal{L}}\subseteq\bar{\mathcal{P}}}^{\bar{G}} : \mathcal{D}^b(\mathrm{Rep}_{W(k)}(\bar{L})) \to \mathcal{D}^b(\mathrm{Rep}_{W(k)}(\bar{G})),$$
$$r_{\bar{G}}^{\bar{\mathcal{L}}\subseteq\bar{\mathcal{P}}} : \mathcal{D}^b(\mathrm{Rep}_{W(k)}(\bar{G})) \to \mathcal{D}^b(\mathrm{Rep}_{W(k)}(\bar{L})).$$

We will be concerned exclusively with the case where $\bar{\mathcal{L}}$ is a maximal torus in $\bar{\mathcal{G}}$. In this case the effect of Deligne–Lusztig restriction on $\bar{\Gamma}$ has been described by Bonnafé and Rouquier when $\bar{\mathcal{L}}$ is a Coxeter torus and by Dudas [2009] in general.

Theorem 2.1 (Bonnafé–Rouquier, Dudas). *When $\bar{\mathcal{L}}$ is the standard maximal torus, there is a natural isomorphism*

$$r_{\bar{G}}^{\bar{\mathcal{L}}\subseteq\bar{\mathcal{P}}}\bar{\Gamma} \cong W(k)[\bar{L}][-\ell(w)]$$

in $\mathcal{D}^b(\mathrm{Rep}_{W(k)}(\bar{L}))$, where w is the element of the Weyl group of $\bar{\mathcal{G}}$ such that $\bar{\mathcal{P}}^w$ is the standard Borel, $\ell(w)$ is its length, and $[-\ell(w)]$ denotes a cohomological shift.

Proof. This is the main theorem of [Dudas 2009]. □

An immediate consequence of this result is that, when $\bar{T}$ is the $\bar{\mathbb{F}}_q$-points of a torus in GL_n, then an endomorphism of $\bar{\Gamma}$ gives rise, by functoriality of Deligne–Lusztig restriction, to an endomorphism of $W(k)[\bar{T}]$ (or, equivalently, an element of $W(k)[\bar{T}]$). We thus obtain homomorphisms

$$\Phi_{\bar{T}} : \bar{E}_{q,n} \to W(k)[\bar{T}]$$

for each torus $\bar{\mathcal{T}}$ in $\bar{\mathcal{G}}$. These are integral versions of the classical "Curtis homomorphisms".

Over $\bar{\mathcal{K}}$, it is not difficult to describe the structure of $\bar{\Gamma}\otimes\bar{\mathcal{K}}$, its endomorphism ring, and the associated Curtis homomorphisms. Recall that an irreducible representation π of $\bar{G}$ is said to be *generic* if π contains the character Ψ, or, equivalently, if there exists a nonzero map from $\bar{\Gamma}$ to π. The irreducible generic representations of $\bar{G}$ over $\bar{\mathcal{K}}$ are indexed by semisimple conjugacy classes s in $\bar{G}'$, where $\bar{G}'$ is the group of $\bar{\mathbb{F}}_q$-points in the group $\bar{\mathcal{G}}'$ that is dual to $\bar{\mathcal{G}}$. More precisely, given such an s, there exists a unique irreducible generic representation $\overline{\mathrm{St}}_s$ in the rational series attached to s.

The association of rational series to semisimple conjugacy classes in $\bar{G}'$ depends on choices which we now recall: let $\mu^{(p)}$ denote the prime-to-p roots of unity in $\bar{\mathcal{K}}$, let $(\mathbb{Q}/\mathbb{Z})^{(p)}$ denote the elements of order prime to p in $(\mathbb{Q}/\mathbb{Z})$, and fix isomorphisms

$$\mu^{(p)} \cong (\mathbb{Q}/\mathbb{Z})^{(p)} \cong \bar{\mathbb{F}}_q^{\times}.$$

Now let t be a semisimple element in $\overline{G}'$, let $\overline{T}'$ be a maximal torus containing s, and let $\overline{T}$ be the dual torus in $\overline{G}$. Let X and X' denote the character groups of $\overline{T}$ and $\overline{T}'$, respectively. We have isomorphisms

$$\overline{T}(\mathbb{F}_q) \cong \mathrm{Hom}(X/(\mathrm{Fr}_q - 1)X, \mathbb{G}_m),$$
$$\overline{T}'(\mathbb{F}_q) \cong \mathrm{Hom}(X'/(\mathrm{Fr}_q - 1)X', \mathbb{G}_m),$$

where Fr_q is the endomorphism induced by the q-power Frobenius. We also have a natural duality $X/(\mathrm{Fr}_q - 1)X \cong \mathrm{Hom}(X'/(\mathrm{Fr}_q - 1)X', (\mathbb{Q}/\mathbb{Z})^{(p)})$. The identifications we fixed above then give rise to isomorphisms

$$\overline{T}'(\mathbb{F}_q) \cong \mathrm{Hom}(X'/(\mathrm{Fr}_q - 1)X', \mathbb{G}_m) \cong X/(\mathrm{Fr}_q - 1)X \cong \mathrm{Hom}(\overline{T}(\mathbb{F}_q), \mu^{(p)}).$$

In this way we associate, to any semisimple element t of $\overline{\mathcal{G}}'(\mathbb{F}_q)$, and any $\overline{T}'$ containing t, a character $\varphi_{\overline{T}',t} : \overline{T}(\mathbb{F}_q) \to \overline{\mathcal{K}}^\times$.

It is immediate (by applying the idempotent of $\overline{\mathcal{K}}[\overline{G}]$ corresponding to the rational series attached to s to Theorem 2.1) that we then have:

Proposition 2.2. *Let $\overline{T}$ be a maximal torus of $\overline{\mathcal{G}}$, and let $\overline{\mathcal{B}}$ be a Borel containing $\overline{T}$. Then, up to a cohomological shift depending only on $\overline{\mathcal{B}}$, we have*

$$r_{\overline{G}}^{\overline{T}\subseteq\overline{\mathcal{B}}}\overline{\mathrm{St}}_s \cong \bigoplus_{t\sim s; t\in\overline{T}'} \varphi_{\overline{T}',t}.$$

Returning to $\overline{\Gamma}$, we have a direct sum decomposition

$$\overline{\Gamma} \otimes \overline{\mathcal{K}} \cong \bigoplus_s \overline{\mathrm{St}}_s$$

It follows immediately that the endomorphism ring of $\overline{\Gamma} \otimes \overline{\mathcal{K}}$ is isomorphic to a product of copies of $\overline{\mathcal{K}}$, indexed by the semisimple conjugacy classes s in $\overline{G}'$. As the endomorphism ring $\mathrm{End}_{W(k)[\overline{G}]}(\overline{\Gamma})$ of $\overline{\Gamma}$ embeds in this product, we see immediately that $\mathrm{End}_{W(k)[\overline{G}]}(\overline{\Gamma})$ is reduced and commutative.

Indeed, it is not difficult to describe the maps $\Phi_{\overline{T}} \otimes \overline{\mathcal{K}}$. The isomorphism

$$\overline{\Gamma} \otimes \overline{\mathcal{K}} \cong \bigoplus_s \overline{\mathrm{St}}_s,$$

where s runs over semisimple conjugacy classes in $\overline{G}'$, gives rise to an isomorphism

$$\overline{E}_{q,n} \otimes \overline{\mathcal{K}} \cong \prod_s \overline{\mathcal{K}}.$$

On the other hand we have a direct sum decomposition

$$\overline{\mathcal{K}}[\overline{T}] \cong \bigoplus_t \varphi_{\overline{T},t}$$

of $\overline{\mathcal{K}}[\overline{T}]$-modules, and hence an algebra isomorphism

$$\overline{\mathcal{K}}[\overline{T}] \cong \prod_t \overline{\mathcal{K}}.$$

It follows immediately from the previous paragraph that $\Phi_{\overline{T}}$ maps the factor of $\overline{\mathcal{K}}$ of $\overline{E}_{q,n} \otimes \overline{\mathcal{K}}$ corresponding to s identically to each factor of $\overline{\mathcal{K}}[\overline{T}]$ that corresponds to a t in the $\overline{G}'$-conjugacy class s, and to zero in the other factors.

Now let $\overline{\mathcal{T}}$ range over all tori in $\overline{G}$, and consider the product map

$$\Phi : \overline{E}_{q,n} \to \prod_{\overline{\mathcal{T}}} W(k)[\overline{T}].$$

For each pair $(\overline{\mathcal{T}}, \varphi)$, where φ is a character $\overline{T} \to \overline{\mathcal{K}}^\times$, we have a map

$$\xi_{\overline{\mathcal{T}},\varphi} : \prod_{\overline{T}} W(k)[\overline{T}] \to \overline{\mathcal{K}}$$

given by composing the projection onto $W(k)[\overline{T}]$ with the map $\varphi : W(k)[\overline{T}] \to \mathcal{K}$.

Define an equivalence relation on such pairs by setting $(\overline{\mathcal{T}}_1, \varphi_1) \sim (\overline{\mathcal{T}}_2, \varphi_2)$ if t_1 and t_2 are conjugate in $\overline{G}'$, where t_1 and t_2 are the elements of the dual tori $\overline{\mathcal{T}}'_1$ and $\overline{\mathcal{T}}'_2$ corresponding to φ_1 and φ_2. Then our description of each $\Phi_{\overline{T}}$ shows that, when $(\overline{\mathcal{T}}_1, \varphi_1) \sim (\overline{\mathcal{T}}_2, \varphi_2)$, one has $\xi_{\overline{\mathcal{T}}_1,\varphi_1} \circ \Phi = \xi_{\overline{\mathcal{T}}_2,\varphi_2} \circ \Phi$. Thus Φ induces a bijection between the $\overline{\mathcal{K}}$-points of $\operatorname{Spec} \overline{E}_{q,n}$ and the equivalence classes of pairs $(\overline{\mathcal{T}}, \varphi)$.

In what follows, it will be necessary for us to consider certain direct factors of $\overline{E}_{q,n}$ arising from idempotents of $W(k)[\overline{G}]$. An ℓ-regular semisimple conjugacy class s in $\overline{G}'$ gives rise, via the choices we have made above, to an idempotent e_s in $W(k)[\overline{G}]$, that acts by the identity on the rational series corresponding to those s' in $\overline{G}$ with ℓ-regular part s, and zero elsewhere. We will denote by $\overline{E}_{q,n,s}$ the direct factor $e_s \overline{E}_{q,n}$ of $\overline{E}_{q,n}$. The $\overline{\mathcal{K}}$-points of $\operatorname{Spec} \overline{E}_{q,n,s}$ are those corresponding to pairs $(\overline{\mathcal{T}}, \varphi)$ such that φ corresponds to an element t of $\overline{\mathcal{T}}'$ whose ℓ-regular part is s.

Now let $s \in \overline{G}'$ be ℓ-regular semisimple and suppose that the characteristic polynomial of s is a power of an irreducible polynomial of degree d. Then the centralizer $\overline{\mathcal{L}}'$ of s in $\overline{\mathcal{G}}'$ is a nonsplit Levi isomorphic to $\operatorname{Res}_{\mathbb{F}_{q^d}/\mathbb{F}_q} \mathrm{GL}_{n/d}$. Let $\overline{\mathcal{L}}$ be the Levi of $\overline{\mathcal{G}}$ dual to $\overline{\mathcal{L}}'$. By [Bonnafé and Rouquier 2003, Théorème 11.8], twisting by the character of $\overline{L}$ associated to s, followed by Deligne–Lusztig induction from $\overline{\mathcal{L}}$ to $\overline{\mathcal{G}}$, is an equivalence of categories from $e_1 \operatorname{Rep}_{W(k)}(\overline{L})$ to $e_s \operatorname{Rep}_{W(k)}(\overline{G})$. Moreover, this equivalence carries $e_1 \overline{\Gamma}_{\overline{\mathcal{L}}}$ to $e_s \overline{\Gamma}$. (This follows from uniqueness of projective envelopes, since the former is the projective envelope of the unique irreducible generic k-representation of $\overline{L}$ in the block corresponding to e_1, and the latter is the projective envelope of the unique irreducible generic k-representation of $\overline{G}$ in the block corresponding to e_s.) We thus have:

Proposition 2.3. *For s an ℓ-regular semisimple element of $\overline{G}'$ whose characteristic polynomial is a power of an irreducible polynomial of degree d over $\mathbb{F}_q$. Then there is a natural isomorphism*

$$\overline{E}_{q,n,s} \cong \overline{E}_{q^d,n/d,1}.$$

The induced map on $\overline{\mathcal{K}}$-points takes the $\overline{\mathcal{K}}$-point of $\operatorname{Spec} \overline{E}_{q^d,n/d,1}$ *corresponding to the ℓ-primary conjugacy class t of $\overline{L}'$ to the $\overline{\mathcal{K}}$-point of* $\operatorname{Spec} \overline{E}_{q,n,s}$ *corresponding to the conjugacy class of st in $\overline{G}'$.*

Proof. The first claim is immediate from the previous paragraph. The second follows from the description of the equivalence of categories on irreducible generic $\overline{\mathcal{K}}$-representations. □

The final structure we will need to consider on $\overline{E}_{q,n}$ is a natural symmetrizing form considered by Bonnafé and Kessar [2008, Section 3.B]. Define a $W(k)$-linear map $\theta : \overline{E}_{q,n} \to W(k)$ by the formula

$$\theta(x) = \frac{1}{n!} \sum_{w \in S_n} \theta_w(\Phi_{\overline{\mathcal{T}}_w}(x)),$$

where $\overline{\mathcal{T}}_w$ is the torus of $\overline{\mathcal{G}}$ associated to the element w of the Weyl group, and $\theta_w : W(k)[\overline{T}_w] \to W(k)$ is the canonical symmetrizing form on $W(k)[\overline{T}_w]$ given by "evaluation at the identity". Note that we can extend θ to a linear map $\overline{E}_{q,n} \otimes \overline{\mathcal{K}} \to \overline{\mathcal{K}}$.

We then have:

Proposition 2.4. *Let t be a semisimple conjugacy class in $\overline{G}'$, and let e_t be the corresponding idempotent of $\overline{E}_{q,n} \otimes \overline{\mathcal{K}}$. Then*

$$\theta(e_t) = \frac{1}{n!} \sum_{w \in S_n} \frac{1}{\#\overline{T}_w} N(w, t),$$

where $N(w, t)$ is the number of elements of $\overline{T}'_w$ in the conjugacy class of t.

Proof. It is easy to see that $\Phi_{\overline{\mathcal{T}}_w}(e_t)$ is equal to the sum, over those $t' \in \overline{\mathcal{T}}'_w$ conjugate to t', of the idempotents $e_{t'}$ of $\overline{\mathcal{K}}[\overline{T}_w]$. The claim is then immediate from the formula for θ. □

3. The integral Bernstein center

We now turn to the first main object of interest in this paper: the integral Bernstein center. Let $G = \mathrm{GL}_n(F)$, and denote by $\mathrm{Rep}_{W(k)}(G)$ and $\mathrm{Rep}_{\overline{\mathcal{K}}}(G)$ the categories of smooth $W(k)[G]$-modules and smooth $\overline{\mathcal{K}}[G]$-modules, respectively.

By the phrase "integral Bernstein center" we mean the center of the category $\mathrm{Rep}_{W(k)}(G)$. We recall what this means:

Definition 3.1. The *center* of an Abelian category $\mathcal{A}$ is the ring of natural transformations $\mathrm{Id}_{\mathcal{A}} \to \mathrm{Id}_{\mathcal{A}}$, where $\mathrm{Id}_{\mathcal{A}}$ denotes the identity functor on $\mathcal{A}$.

By definition, if Z is the center of $\mathcal{A}$, then specifying an element of Z amounts to specifying an endomorphism of every object of $\mathcal{A}$, such that the resulting collection commutes with all arrows in $\mathcal{A}$. The center of $\mathcal{A}$ is thus a commutative ring that acts naturally on every object in $\mathcal{A}$, and this action is compatible with all morphisms in $\mathcal{A}$.

Bernstein and Deligne [1984] gave a complete and explicit description of the center $\widetilde{Z}$ of $\mathrm{Rep}_{\overline{\mathcal{K}}}(G)$. We briefly summarize their results: first, define an equivalence relation on pairs $(M, \tilde{\pi})$, where M is a Levi of G and π is an irreducible supercuspidal representation of M over $\overline{\mathcal{K}}$ by declaring $(M_1, \tilde{\pi}_1)$ to be *inertially equivalent* to $(M_2, \tilde{\pi}_2)$ if $\tilde{\pi}_1$ is G-conjugate to an unramified twist of $\tilde{\pi}_2$. One then has:

Theorem 3.2 [Bernstein and Deligne 1984, Proposition 2.10]. *There is a bijection $(M, \tilde{\pi}) \mapsto e_{(M,\tilde{\pi})}$ between inertial equivalence classes of pairs $(M, \tilde{\pi})$ over $\overline{\mathcal{K}}$ and primitive idempotents of $\widetilde{Z}$, such that for any irreducible smooth representation Π of G over $\overline{\mathcal{K}}$ $e_{(M,\tilde{\pi})}$ acts via the identity on Π if Π has supercuspidal support in the inertial equivalence class of $(M, \tilde{\pi})$, and by zero otherwise.*

The upshot is that $\widetilde{Z}$ decomposes as an infinite product of the rings $e_{(M,\tilde{\pi})}\widetilde{Z}$ as $(M, \tilde{\pi})$ runs over all inertial equivalence classes of pairs. Denote $e_{(M,\tilde{\pi})}\widetilde{Z}$ by $\widetilde{Z}_{(M,\tilde{\pi})}$. Then Bernstein and Deligne gave a complete description of the ring structure of $\widetilde{Z}_{(M,\tilde{\pi})}$ that we now explain.

Let M_0 be the smallest subgroup of M containing every compact open subgroup of M. Then M/M_0 is a free abelian group of finite rank, and $\operatorname{Spec} \overline{\mathcal{K}}[M/M_0]$ is a torus whose $\overline{\mathcal{K}}$-points are in bijection with the characters $M/M_0 \to \overline{\mathcal{K}}^\times$. Let H be the subgroup of these characters consisting of those characters χ such that $\tilde{\pi} \otimes \chi$ is isomorphic to $\tilde{\pi}$. Then H is a finite abelian group that acts on $\overline{\mathcal{K}}[M/M_0]$. The torus $\operatorname{Spec} \overline{\mathcal{K}}[(M/M_0)]^H$ is a quotient of $\operatorname{Spec} \overline{\mathcal{K}}[M/M_0]$; its $\overline{\mathcal{K}}$-points correspond to H-orbits of characters of M/M_0.

Now let W_M be the subgroup of the Weyl group of G consisting of w such that $wMw^{-1} = M$. Let $W_M(\tilde{\pi})$ be the subgroup of W_M consisting of w such that the representation $\tilde{\pi}^w$ of M is an unramified twist of $\tilde{\pi}$. Then we have a natural action of $W_M(\tilde{\pi})$ on $\overline{\mathcal{K}}[(M/M_0)]^H$, characterized by

$$\tilde{\pi} \otimes \chi^w \cong (\tilde{\pi} \otimes \chi)^w$$

for characters χ of M/M_0. We then have:

Theorem 3.3 [Bernstein and Deligne 1984, Théorème 2.13]. *There is a unique natural isomorphism*

$$\widetilde{Z}_{(M,\tilde{\pi})} \cong \left(\overline{\mathcal{K}}[(M/M_0)]^H\right)^{W_M(\tilde{\pi})}$$

such that, for any irreducible representation Π over $\overline{\mathcal{K}}$ whose supercuspidal support has the form $\tilde{\pi} \otimes \chi$, $\widetilde{Z}_{(M,\tilde{\pi})}$ acts on Π via the map

$$\left(\overline{\mathcal{K}}[(M/M_0)]^H\right)^{W_M(\tilde{\pi})} \to \overline{\mathcal{K}}[M/M_0] \to \overline{\mathcal{K}}$$

corresponding to the character $\chi : M/M_0 \to \overline{\mathcal{K}}^\times$. In particular $\widetilde{Z}_{(M,\tilde{\pi})}$ is a reduced, finitely generated, and normal $\overline{\mathcal{K}}$-algebra.

In particular, $\widetilde{Z}$ acts on two irreducible representations Π, Π' of G via the same map $\widetilde{Z} \to \overline{\mathcal{K}}$ if, and only if, Π and Π' have the same supercuspidal support. This defines, for each $(M, \tilde{\pi})$, a bijection between the $\overline{\mathcal{K}}$-points of $\operatorname{Spec} \widetilde{Z}_{(M,\tilde{\pi})}$ and supercuspidal supports in the inertial equivalence class of $(M, \tilde{\pi})$; that is, unramified twists of $\tilde{\pi}$ considered up to $W_M(\tilde{\pi})$-conjugacy.

Now let L be a Levi in GL_n; then L factors as a product of L_i isomorphic to $\mathrm{GL}_{n_i}(F)$. For each i, let M_i be a Levi in L_i, and $\tilde{\pi}_i$ an irreducible supercuspidal $\overline{\mathcal{K}}$-representation of M_i. We then have isomorphisms

$$\widetilde{Z}_{M_i,\tilde{\pi}_i} \cong \left(\overline{\mathcal{K}}[(M_i/(M_i)_0)]^{H_i}\right)^{W_{M_i}(\tilde{\pi}_i)}.$$

Let M be the product of the M_i; we may regard it as a Levi of L and hence as a Levi of $\mathrm{GL}_n(F)$. Let $\tilde{\pi}$ be the tensor product of the $\tilde{\pi}_i$. The quotient M/M_0 factors naturally as a product of $M_i/(M_i)_0$, and this induces a map

$$\big(\overline{\mathcal{K}}[(M/M_0)]^H\big)^{W_M(\tilde{\pi})} \to \bigotimes_i \big(\overline{\mathcal{K}}[(M_i/(M_i)_0)]^{H_i}\big)^{W_{M_i}(\tilde{\pi}_i)}$$

and hence a map

$$\mathrm{Ind}_{\{(M_i,\tilde{\pi}_i)\}} : \widetilde{Z}_{(M,\tilde{\pi})} \to \bigotimes_i \widetilde{Z}_{(M_i,\tilde{\pi}_i)}.$$

On $\overline{\mathcal{K}}$-points this takes the $\mathcal{K}$-point of the tensor product that corresponds to the collection of supercuspidal supports $\{(M_i, \tilde{\pi}_i \otimes \chi_i)\}$ to the point of $\mathrm{Spec}\, \widetilde{Z}_{(M,\tilde{\pi})}$ corresponding to the supercuspidal support $(M, \otimes_i(\tilde{\pi}_i \otimes \chi_i))$.

We now turn to the study of $\mathrm{Rep}_{W(k)}(G)$; let Z denote the center of this category. In this setting there is an analogue of the Bernstein–Deligne characterization of the primitive idempotents of Z. By [Helm 2016a, Theorem 11.8], such idempotents are parametrized by inertial equivalence classes of pairs (L, π), where π is now an irreducible supercuspidal representation of L over k.

If we let $e_{[L,\pi]}$ denote the idempotent of Z corresponding to (L, π), $\mathrm{Rep}_{W(k)}(G)_{[L,\pi]}$ the corresponding block, and $Z_{[L,\pi]}$ the corresponding factor of the Bernstein center, then one has the following basic structure results:

Theorem 3.4 [Helm 2016a, Theorem 12.8]. *The ring $Z_{[L,\pi]}$ is a finitely generated, reduced, flat $W(k)$-algebra.*

It is important to note that, in contrast to the situation over $\overline{\mathcal{K}}$, the ring $Z_{[L,\pi]}$ is in general very far from being normal.

We also have a description of $Z_{[L,\pi]} \otimes \overline{\mathcal{K}}$ in terms of $\widetilde{Z}$. This can be made precise as follows: if $(M, \tilde{\pi})$ is a pair over $\overline{\mathcal{K}}$, and Π is an irreducible integral representation of G over $\overline{\mathcal{K}}$ with supercuspidal support in the inertial equivalence class of $(M, \tilde{\pi})$, then there exists a (possibly proper) Levi subgroup L of M, and an irreducible supercuspidal representation π of L, such that every irreducible subquotient of the mod ℓ reduction of Π has supercuspidal support (L, π). Moreover, the inertial equivalence class of (L, π) depends only on that of $(M, \tilde{\pi})$, and not on the particular choice of π. We say that $(M, \tilde{\pi})$ *reduces modulo ℓ* to (L, π); this defines a finite-to-one map from inertial equivalence classes over $\overline{\mathcal{K}}$ to inertial equivalence classes over k. One then has:

Theorem 3.5 [Helm 2016a, Proposition 12.1]. *The natural map $Z \otimes \overline{\mathcal{K}} \to \widetilde{Z}$ induces an isomorphism*

$$Z_{[L,\pi]} \otimes \overline{\mathcal{K}} \cong \prod_{(M,\tilde{\pi})} \widetilde{Z}_{(M,\tilde{\pi})},$$

where the product is over all pairs $(M, \tilde{\pi})$, up to inertial equivalence, that reduce modulo ℓ to the pair (L, π).

From this and the description of the $\overline{\mathcal{K}}$-points of $\mathrm{Spec}\, \widetilde{Z}_{(M,\tilde{\pi})}$ one immediately deduces:

Corollary 3.6. *The $\overline{\mathcal{K}}$-points of* $\mathrm{Spec}\, Z_{[L,\pi]}$ *are in bijection with the supercuspidal supports of irreducible smooth $\overline{\mathcal{K}}$-representations in* $\mathrm{Rep}_{W(k)}(G)_{[L,\pi]}$.

We now give a more precise description of $Z_{[L,\pi]}$. We first reduce to a more easily studied special case:

Definition 3.7. A pair (L, π) is *simple* if there exist r, m such that $n = rm$, L is isomorphic to $\mathrm{GL}_m(F)^r$, and π, up to unramified twist, is of the form $(\pi')^{\otimes r}$ for an irreducible supercuspidal representation π' of $\mathrm{GL}_m(F)$.

Note that any pair (L, π) factors uniquely as a product of simple pairs (L^i, π^i), with $\pi^i \cong (\pi'_i)^{\otimes r_i}$, such that no π'_i is an unramified twist of any other. One then has:

Theorem 3.8 [Helm 2016a, Theorem 12.4]. *Let $\{(L^i, \pi^i)\}$ be the natural decomposition of (L, π) as a product of simple pairs. Then there is a natural isomorphism*

$$Z_{[L,\pi]} \cong \bigotimes_i Z_{[L^i,\pi^i]}$$

such that, for any sequence $\{(M^i, \tilde{\pi}^i)\}$ reducing modulo ℓ to $\{(L^i, \tilde{\pi}^i)\}$, the diagram

$$\begin{array}{ccc} Z_{[L,\pi]} \otimes \overline{\mathcal{K}} & \longrightarrow & \left[\bigotimes_i Z_{[L^i,\pi^i]}\right] \otimes \overline{\mathcal{K}} \\ \downarrow & & \downarrow \\ \widetilde{Z}_{(M,\tilde{\pi})} & \longrightarrow & \bigotimes_i \widetilde{Z}_{(M^i,\tilde{\pi}^i)} \end{array}$$

commutes, where $(M, \tilde{\pi})$ is the product of the $(M_i, \tilde{\pi}_i)$, and the bottom horizontal map is the map $\mathrm{Ind}_{\{(M^i,\tilde{\pi}^i)\}}$ described above.

We thus focus our attention on the case where (L, π) is simple. Fix an integer n_1 and an irreducible supercuspidal representation π' of $\mathrm{GL}_{n_1}(F)$ over k. For each $m > 0$, let L_m be a Levi of $\mathrm{GL}_{n_1 m}(F)$ isomorphic to $\mathrm{GL}_{n_1}(F)^m$, and let π_m be the representation $(\pi')^{\otimes m}$ of L_m. We can then consider the family of rings $Z_m := Z_{[L_m,\pi_m]}$ as n varies.

Section 13 of [Helm 2016a] contains detailed information about the structure of the family Z_m. In particular this structure theory is closely related to the endomorphism rings of certain projective objects $\mathcal{P}_{K_m,\tau_m}$ for particular m. More precisely, consider the group of unramified characters χ of $\mathrm{GL}_{n_1}(F)$ such that $\pi' \otimes \chi$ is isomorphic to π'. This is a finite group; denote its order by f'. Then attached to the system of pairs (L_m, π_m) we have a system of projective objects $\mathcal{P}_{K_m,\tau_m}$, where m lies in the set $\{1, e_{q^{f'}}, \ell e_{q^{f'}}, \ell^2 e_{q^{f'}}, \dots\}$. (We refer the reader to Sections 7 and 9 of [Helm 2016a] for a construction and structure theory of these objects.) For brevity, denote the representation $\mathcal{P}_{K_m,\tau_m}$ by $\mathcal{P}_m$.

For such m, let E_m denote the endomorphism ring of $\mathcal{P}_m$. Then, by Corollary 9.2 of [Helm 2016a], E_m is a reduced, finite type, ℓ-torsion free $W(k)$-algebra. Moreover, we have a map $Z_m \to E_m$ that gives the action of Z_m on the object $\mathcal{P}_m$ of $\mathrm{Rep}_{W(k)}(\mathrm{GL}_{n_1 m}(F))_{[L_m,\pi_m]}$.

If m is arbitrary, the relationship between the rings Z_m and E_m is more complicated. For a partition ν of m, we will say that ν is q-relevant if each ν_i belongs to the set $\{1, e_q, \ell e_q, \ell^2 e_q, \dots\}$, where e_q is the multiplicative order of q modulo ℓ (relevant partitions were called admissible in [Helm 2016a]). Let ν be the maximal $q^{f'}$-relevant partition of m. Let M_ν and P_ν be the standard Levi and (upper triangular) parabolic subgroups of $\mathrm{GL}_{n_1 m}$ attached to $n_1 \nu$, so that M_ν is a product of $\mathrm{GL}_{n_1 \nu_i}(F)$, and consider the representation $\bigotimes_i \mathcal{P}_{\nu_i}$ of M_ν. Then Z_m acts on the parabolic induction $i_{P_\nu}^{\mathrm{GL}_{n_1 m}(F)} \bigotimes_i \mathcal{P}_{\nu_i}$, and we have:

Theorem 3.9 [Helm 2016a, Theorem 13.7]. *The action of Z_m on $i_{P_\nu}^{\mathrm{GL}_{n_1 m}(F)} \bigotimes_i \mathcal{P}_{\nu_i}$ factors through the action of $\bigotimes_i E_{\nu_i}$ on $\bigotimes_i \mathcal{P}_{\nu_i}$. Moreover, the resulting map*

$$Z_m \to \bigotimes_i E_{\nu_i}$$

is injective with saturated image, and is an isomorphism if m lies in $\{1, e_{q^{f'}}, \ell e_{q^{f'}}, \dots\}$. (Note that in this case ν is the one-element partition $\{m\}$ of m.)

For m in $\{1, e_{q^{f'}}, \ell e_{q^{f'}}, \dots\}$ we thus have a natural identification of Z_m with E_m. For arbitrary m, we can regard the map $Z_m \to \bigotimes_i E_{\nu_i}$ as a map $Z_m \to \bigotimes_i Z_{\nu_i}$. Denote this map by Ind_ν. It is injective with saturated image.

For m in $\{1, e_{q^{f'}}, \ell e_{q^{f'}}, \dots\}$, the results of Sections 7 and 9 of [Helm 2016a] give very precise information about E_m, and hence Z_m. In particular there is an integer f dividing f', and a cuspidal k-representation σ_m of $\mathrm{GL}_{mf'/f}(\mathbb{F}_{q^f})$ (attached to an ℓ-regular conjugacy class $(s_1')^m$ with s_1' irreducible of degree f' over $\mathbb{F}_{q^f}$), such that the projective $\mathcal{P}_m$ is a compact induction $\text{c-Ind}_{K_m}^{\mathrm{GL}_{n_1 m}(F)} \tilde{\kappa}_m \otimes \mathcal{P}_{\sigma_m}$, where $\tilde{\kappa}_m$ comes from type theory and $\mathcal{P}_{\sigma_m}$ is the projective envelope of σ_m, inflated to a representation of K_m via a natural map $K_m \to \mathrm{GL}_{mf'/f}(\mathbb{F}_{q^f})$.

Section 5 of [Helm 2016a] shows that $\mathcal{P}_{\sigma_m}$ is the projection of the Gelfand–Graev representation of $\mathrm{GL}_{mf'/f}(\mathbb{F}_{q^f})$ to the block containing σ_m. In particular, the results of Section 2 identify the endomorphisms of $\mathcal{P}_{\sigma_m}$ with $\bar{E}_{q^f, md, s}$, where we have written $s = (s_1)^m$ and $d = \frac{f'}{f}$. By Proposition 2.3 we may identify $\bar{E}_{q^f, md, s}$ with $\bar{E}_{q^{f'}, m, 1}$.

We thus obtain an embedding of $\bar{E}_{q^{f'}, m, 1}$ in E_m for such m. Furthermore, Section 9 of [Helm 2016a] constructs an invertible element $\Theta_{m,m}$ of E_m. We thus obtain a map

$$\bar{E}_{q^{f'}, m, 1}[T, T^{-1}] \to E_m$$

taking T to $\Theta_{m,m}$. It follows easily from the description of E_m as a Hecke algebra in Section 9 of [Helm 2016a] that the image of this map consists of the elements of E_m supported on double cosets of the form $K_m z_{m,m}^r K_m$ for various r. (In particular, this image is saturated in E_m.)

The image of $\bar{E}_{q^{f'}, m, 1}$ in Z_m is easy to describe. Indeed, we have:

Proposition 3.10. *Let m lie in $\{1, e_{q^{f'}}, \ell e_{q^{f'}}, \dots\}$, and let x be an element of $\bar{E}_{q^{f'}, m, 1}$, where the latter is considered as a subalgebra of Z_m. Then for any irreducible $\overline{\mathcal{K}}$-representations Π, Π' of $\mathrm{GL}_{n_1 m}(F)$ in the same block of $\mathrm{Rep}_{\overline{\mathcal{K}}}(\mathrm{GL}_{n_1 m}(F))$, the action of x on Π and Π' is via the same scalar. Conversely, any element of Z_m with this property lies in $\bar{E}_{q^{f'}, m, 1}$.*

Proof. The ring Z_m annihilates both Π and Π' unless Π and Π' belong to a block of the form $\mathrm{Rep}_{\overline{\mathcal{K}}}(\mathrm{GL}_{n_1 m}(F))_{(M_s, \pi_s)}$ for a suitable s, in the notation of [Helm 2016a, Section 9]. In this case the action of Z_m on Π and Π' factors through the action of Z_m on the summand $\text{c-Ind}_{K_m}^{\mathrm{GL}_{n_1 m}(F)} \tilde{\kappa}_m \otimes \mathrm{St}_s$ of $\text{c-Ind}_{K_m}^{\mathrm{GL}_{n_1 m}(F)} \tilde{\kappa}_m \otimes \mathcal{P}_{\sigma_m} \otimes \overline{\mathcal{K}}$. In particular the action of x on Π and Π' factors through the action of x on St_s, which is by a scalar.

Since $\overline{E}_{q^{f'},m,1}$ is saturated in Z_m, it suffices to prove the converse over $\overline{\mathcal{K}}$. But it follows easily from our factorization of Z_m in characteristic zero that every idempotent of $Z_m \otimes \overline{\mathcal{K}}$ is contained in $\overline{E}_{q^{f'},m,1}$; since these idempotents correspond to the blocks of $\mathrm{Rep}_{\overline{\mathcal{K}}}(\mathrm{GL}_{n_1 m}(F))_{M,\pi}$ the claim follows. □

We also make the following observation about the action of $\Theta_{m,m} \in Z_m$:

Proposition 3.11. *Let P be a parabolic subgroup of $\mathrm{GL}_{n_1 m}(F)$, with Levi subgroup M, and let π be an irreducible cuspidal $\overline{\mathcal{K}}$-representation of M such that $i_P^G \pi$ lies in the block corresponding to L_m, π_m. Suppose that M decomposes as a product of groups $M_i = \mathrm{GL}_{n_1 m_i}(F)$, and let χ be an unramified character of M, of the form $\otimes_i (\chi_i \circ \det)$, where we regard $(\chi_i \circ \det)$ as a character of M_i.*

Let $x \in \overline{\mathcal{K}}^\times$ be the scalar by which $\Theta_{m,m}$ acts on $i_P^G \pi$. Then $\Theta_{m,m}$ acts on $i_P^G \pi \otimes \chi$ via $x \prod_i \chi_i^{f'}(\varpi_F)$.

Proof. For some s, the pair (M, π) is conjugate to an unramified twist of one of the pairs (M_s, π_s) described in Section 9 of [Helm 2016a]. Thus, by Theorem 9.4 of [Helm 2016a], the action of $\Theta_{m,m}$ on π is via the element $\theta_{m,s}$ of Z_{M_s,π_s} defined in Section 9 of [Helm 2016a], and the claim is immediate from the definition of $\theta_{m,s}$ in that section. □

Finally, let m' and m be two consecutive elements of $\{1, e_{q^{f'}}, \ell e_{q^{f'}}, \dots\}$, and set $j = \frac{m}{m'}$. Theorem 13.5 of [Helm 2016a] then provides a map

$$\mathrm{Ind}_{m',m} : Z_m \to Z_{m'}^{\otimes j}$$

that is compatible with parabolic induction, in the sense that the action of x in Z_m on $i_P^{\mathrm{GL}_{n_1 m}(F)} \pi$ (where $P = MU$ is a parabolic such that M is isomorphic to $\mathrm{GL}_{n_1 m'}(F)^j$) is induced by the action of $\mathrm{Ind}_{m',m}(x)$ on π. The image of this map is not saturated but we have:

Theorem 3.12 [Helm 2016a, Theorem 13.6]. *Let y be an element of $Z_{m'}^{\otimes j}$ such that, for some a, $\ell^a y$ lies in the image of $\mathrm{Ind}_{m',m}$. Then there exists an element $\tilde{y}$ of Z_m, an element x of $\overline{E}_{q^{f'},m,1}[T^{\pm 1}]$, and an integer $b > 0$ such that $\mathrm{Ind}_{m',m}(x) = \ell^b (y - \mathrm{Ind}_{m',m}(\tilde{y}))$.*

The map $\mathrm{Ind}_{m',m}$ is not injective, but its kernel has a rather simple structure:

Proposition 3.13. *There exists an ideal $I_{m',m}$ of $\overline{E}_{q^{f'},m,1}$ such that the kernel of $\mathrm{Ind}_{m',m}$ is equal to $I_{m',m}[\Theta_{m,m}^{\pm 1}]$.*

Proof. Since $\overline{E}_{q^{f'},m,1}[\Theta_{m,m}^{\pm 1}]$ is saturated in Z_m we can prove this after tensoring with $\overline{\mathcal{K}}$. We have a decomposition

$$Z_m \otimes \overline{\mathcal{K}} \cong \prod_i \widetilde{Z}_{(M_i, \tilde{\pi}_i)},$$

where $(M_i, \tilde{\pi}_i)$ run over the $\overline{\mathcal{K}}$-inertial equivalence classes in the block corresponding to $[L_m, \pi_m]$. In particular the partitions corresponding to the M_i are all $q^{f'}$-relevant. Fix a factor in this product corresponding to a pair $(M_i, \tilde{\pi}_i)$. On this factor, we can describe the map $\mathrm{Ind}_{m',m}$ in the following way: let $(M_{ij}, \tilde{\pi}_{ij})$ run over the set of M_ν-inertial equivalence classes of pairs that are $\mathrm{GL}_{n_1 m}$-inertially equivalent to $(M_i, \tilde{\pi}_i)$, where ν is the partition $(m', \dots, m')$ of m and M_ν is the corresponding Levi of $\mathrm{GL}_{n_1 m}$. Since

M_{ij} is a Levi contained in M_ν the pair $(M_{ij}, \tilde{\pi}_{ij})$ breaks up as a product of $\frac{m}{m'}$ pairs $(M_{ijk}, \tilde{\pi}_{ijk})$ in $\mathrm{GL}_{n_1 m'}$. On the factor $\widetilde{Z}_{(M_i,\tilde{\pi}_i)}$ of $Z_m \otimes \overline{\mathcal{K}}$, $\mathrm{Ind}_{m,n}$ is the sum of the maps

$$\mathrm{Ind}_{(M_{ij},\tilde{\pi}_{ij})} : \widetilde{Z}_{(M_i,\tilde{\pi}_i)} \to \bigotimes_k \widetilde{Z}_{(M_{ijk},\tilde{\pi}_{ijk})}.$$

In particular $\mathrm{Ind}_{m',m}$ is injective on the factor $\widetilde{Z}_{(M_i,\tilde{\pi}_i)}$ if M_i is a proper Levi subgroup and zero otherwise. When M_i is not a proper Levi then the pair $(M_i, \tilde{\pi}_i)$ gives a cuspidal inertial equivalence class, so $\widetilde{Z}_{(M_i,\tilde{\pi}_i)}$ is isomorphic to $\overline{\mathcal{K}}[\Theta_{m,m}^{\pm 1}]$. Thus the kernel of $\mathrm{Ind}_{m',m} \otimes \overline{\mathcal{K}}$ is equal to $\tilde{I}_{m',m}[\Theta_{m,m}^{\pm 1}]$, where $\tilde{I}_{m,m}$ is the ideal of $\bar{E}_{q^{f'},m,1} \otimes \overline{\mathcal{K}}$ generated by the idempotents of the latter that correspond to cuspidal inertial equivalence classes $(M_i, \tilde{\pi}_i)$. □

4. The ring $R_{q,n}$

We now turn to the second principal object of study of this paper, which is a moduli space of representations of W_F. We begin by studying spaces of tame representations. Let $X_{q,n}$ be the affine $W(k)$-scheme parametrizing pairs of invertible n by n matrices (Fr, σ) such that $\mathrm{Fr}\,\sigma\,\mathrm{Fr}^{-1} = \sigma^q$, and let $X_{q,n}^0$ be the connected component of $X_{q,n}$ containing the k-point $\mathrm{Fr} = \sigma = \mathrm{Id}_n$. Let $S_{q,n}$ and $R_{q,n}$ be the rings of functions on $X_{q,n}$ and $X_{q,n}^0$, respectively, so that $X_{q,n} = \mathrm{Spec}\, S_{q,n}$ and $X_{q,n}^0 = \mathrm{Spec}\, R_{q,n}$.

Lemma 4.1. *Let L be an algebraically closed field that is a $W(k)$-algebra and x be an L-point of $X_{q,n}$ corresponding to a pair $(\mathrm{Fr}_x, \sigma_x)$ of elements of $\mathrm{GL}_n(L)$. Then x lies in $X_{q,n}^0$ if, and only if, the eigenvalues of σ_x are ℓ-power roots of unity.*

Proof. Consider the map $X_{q,n} \to \mathbb{A}^n_{W(k)}$ that takes a point x to the coefficients of the characteristic polynomial of σ_x. Let Y be the image of this map. For all L and x, σ_x is an element of $\mathrm{GL}_n(L)$ conjugate to its q-th power, so its image in $Y(L)$ is a polynomial of degree n whose roots, counted with multiplicities, are stable under the q-th power map. That is, every point of $Y(L)$ corresponds to the characteristic polynomial of a diagonal matrix that is conjugate to its q-th power. Conversely, given such a matrix σ it is easy to construct an L-point x of $X_{q,n}$ with $\sigma_x = \sigma$.

Let $\widetilde{Y} \subset \mathbb{A}^n_{W(k)}$ be the space of diagonal matrices that are conjugate to their q-th powers; we then have a map $\widetilde{Y} \to \mathbb{A}^n_{W(k)}$ that sends such a matrix to the coefficients of its characteristic polynomial. The argument of the previous paragraph shows that the (set-theoretic) image of $\widetilde{Y}$ is equal to Y. On the other hand, $\widetilde{Y}(\overline{\mathcal{K}})$ is a finite collection of points; indeed, the entries of any diagonal matrix that is conjugate to its q-th power are roots of unity of order bounded in terms of q and n. Thus the "coordinates" of each $\overline{\mathcal{K}}$-point of $\widetilde{Y}$ are integral over $W(k)$, and every point of $\widetilde{Y}(k)$ is in the closure of some point of $\widetilde{Y}(\overline{\mathcal{K}})$. It follows that the same is true for Y; in particular Y is the closure of a finite set of $\overline{\mathcal{K}}$-points, and the closure of any $\overline{\mathcal{K}}$-point of Y meets the special fiber of Y. Therefore, the connected component Y^0 of Y containing the image of $X_{q,n}^0$ is the closure of the set of $\overline{\mathcal{K}}$-points of Y that "specialize" mod ℓ to the characteristic polynomial $(X-1)^n$ of the identity matrix. The only k-point of this component arises from the characteristic polynomial of the identity matrix, and the $\overline{\mathcal{K}}$-points of this component correspond to

characteristic polynomials of elements of $\widetilde{Y}(\overline{\mathcal{K}})$ whose roots reduce to 1 modulo ℓ. The roots of such a polynomial are ℓ-power roots of unity. Therefore, for x in $X^0_{q,n}(L)$ the roots of the characteristic polynomial of σ_x are ℓ-power roots of unity, as required.

Conversely, let x be an L-point of $X_{q,n}$, and suppose that the eigenvalues of σ_x are ℓ-power roots of unity. Note that $\mathrm{GL}_n(L)$ acts on $X_{q,n}(L)$, by conjugation on both F and σ, and this action preserves the connected components. We may thus assume σ_x is in Jordan normal form; in particular its entries lie in k or an integral extension $\mathcal{O}$ of $W(k)$. Moreover, for a fixed σ_x, the set of Fr_x such that $\mathrm{Fr}_x\,\sigma_x = \sigma_x^q\,\mathrm{Fr}_x$ is a linear space; there is thus an invertible Fr'_x whose entries lie in k or $W(k)$, such that $\mathrm{Fr}'_x\,\sigma_x = \sigma_x^q\,\mathrm{Fr}'_x$ and $(\mathrm{Fr}'_x, \sigma_x)$ lies on the same connected component as x.

If L has characteristic ℓ, the above construction yields a k-point of $X_{q,n}$ in the same connected component as x. If L has characteristic zero, the closure of the point (Fr', σ) constructed above contains a k-point (Fr'', σ') of $X_{q,n}$ in the same connected component as x. Moreover, σ' is unipotent and in Jordan normal form. Thus in the closure of orbit of (Fr'', σ') under conjugation by diagonal matrices there is a point where σ is the identity. It is clear that such a point lies in the connected component of the k-point x where $\mathrm{Fr}_x = \sigma_x = \mathrm{Id}_n$. □

The ring $R_{q,n}$ is rather well-behaved from an algebraic standpoint. In particular, one has:

Proposition 4.2. *The ring $R_{q,n}$ is reduced and locally a complete intersection. Moreover, $R_{q,n}$ is flat as a $W(k)$-algebra.*

Proof. This argument is a slight elaboration of an argument due to Choi [2009]. We give a sketch here.

First note that $X_{q,n}$ is given by n^2 relations in a space of dimension $2n^2+1$. Consider the map $X_{q,n} \to \mathbb{A}^{n^2}_{W(k)}$ that sends a point x to the matrix σ_x. Let L be an algebraically closed field that is a $W(k)$-algebra, and let x be an L-point of $X_{q,n}$.

The group $\mathrm{GL}_n(L)$ acts on the set of L-points of $X_{q,n}$ by conjugation. Consider the locally closed subset U_{σ_x} of $\mathrm{Spec}\,\mathbb{A}^{n^2}_L$ consisting of those σ' conjugate to σ_x. For any L-point σ' of U_{σ_x}, the fiber of $X_{q,n} \times_{W(k)} L$ over σ' consists of pairs $(\mathrm{Fr}'\,h, \sigma')$, where Fr' is a fixed element of GL_n such that $\mathrm{Fr}'\,\sigma'(\mathrm{Fr}')^{-1} = (\sigma')^q$ and h commutes with σ'

In particular, the dimension of the preimage of U_σ in $X_{q,n} \times_{W(k)} L$ is equal to the dimension of U_σ plus the dimension of the stabilizer of σ under conjugation; this is clearly n^2. As σ varies over a finite list of conjugacy classes, the preimages of the U_σ cover $X_{q,n} \times_{W(k)} L$; thus $X_{q,n} \times_{W(k)} L$ is equidimensional of dimension n^2. On the other hand the dimension of $X_{q,n}$ is at least n^2+1. It follows that the Zariski closures of the preimages of sets U_σ are irreducible components of $X_{q,n}$, and that no irreducible component of $X_{q,n}$ is contained in the special fiber (as it would then be a component of $X_{q,n} \times_{W(k)} k$ of dimension at most n^2). It also follows that every irreducible component of $X_{q,n}$ has dimension n^2+1, because if we had a component of larger dimension then its base change to $\overline{\mathcal{K}}$ would have dimension greater than n^2. In particular $X_{q,n}$ is a complete intersection. It follows that $R_{q,n}$ is a local complete intersection.

An argument of Choi [2009, Theorem 3.0.13] shows that $(\mathrm{Spec}\,R_{q,n})_m\left[\frac{1}{\ell}\right]$ is generically smooth for any maximal ideal m of $R_{q,n}$; in particular $X^0_{q,n}$ is generically reduced. By the unmixedness theorem the

local complete intersection $X^0_{q,n}$ has no embedded points, so $R_{q,n}$ is reduced. As the generic points of Spec $R_{q,n}$ all have characteristic zero, we may conclude that $R_{q,n}$ is flat over $W(k)$. □

We have a universal pair of matrices (Fr, σ) in $\mathrm{GL}_n(R_{q,n})$. The above result immediately implies:

Corollary 4.3. *There exists a power ℓ^a of ℓ such that σ^{ℓ^a} is unipotent in $\mathrm{GL}_n(R_{q,n})$.*

Proof. Since $R_{q,n}$ is reduced and flat over $W(k)$, it suffices to check that σ^{ℓ^a} is unipotent for some a at each of the generic points of Spec $R_{q,n}$, all of which lie in characteristic zero. This is an immediate consequence of Lemma 4.1. □

Let L be a finite extension of $\mathcal{K}$. We call an L-point of $X^0_{q,n}$ *integral* if the corresponding map $R_{q,n} \to L$ factors through the ring of integers $\mathcal{O}_L$.

Lemma 4.4. *Let x be an L-point of $X^0_{q,n}$, and suppose that the eigenvalues of Fr_x lie in $\mathcal{O}^\times_{L'}$ for some finite extension L' of L. Then there is an integral point of $X^0_{q,n}$ in the GL_n-orbit of x.*

Proof. Extending L if necessary, we may assume that the eigenvalues of σ_x are in L, and hence $\mathcal{O}_L$. Then (for instance, by putting σ_x in Jordan normal form) we can find an $\mathcal{O}_L$-sublattice M of L^n preserved by σ_x. Using $\mathrm{Fr}_x\, \sigma_x\, \mathrm{Fr}_x^{-1} = \sigma_x^q$, we find that $\mathrm{Fr}_x\, M$, $\mathrm{Fr}_x^2\, M$, etc. are also preserved by σ_x. Consider the lattice M' given by $M + \mathrm{Fr}_x\, M + \cdots + \mathrm{Fr}_x^{n-1}\, M$; it is clearly preserved by σ_x. On the other hand, since Fr_x is annihilated by a polynomial with integral coefficients, $\mathrm{Fr}_x^n\, M$ is contained in M', and hence $\mathrm{Fr}_x\, M'$ is contained in M'. Since Fr_x has unit determinant we must have $\mathrm{Fr}_x\, M' = M'$. Thus M' is stable under both Fr_x and σ_x. Choosing a basis for M', we find an integral point of $X^0_{q,n}$ in the same GL_n-orbit as x. □

Lemma 4.5. *For any positive integer m, and any element λ of $\mathcal{O}^\times_L$, there is an element $g_{m,\lambda}$ of $\mathrm{GL}_m(L)$, with unit eigenvalues, such that $g_{m,\lambda} J_{m,\lambda^q} g_{m,\lambda}^{-1} = J^q_{m,\lambda}$, where $J_{m,\lambda}$ is the unipotent Jordan block of size m.*

Proof. The matrices J_{m,λ^q} and $J^q_{m,\lambda}$ are regular with the same eigenvalues, hence conjugate by some $g' \in \mathrm{GL}_m(L)$. Since J_{m,λ^q} is contained in a unique Borel subgroup of GL_m (namely, the standard one), the same is true of J_{m,λ^q}. Thus g' normalizes the standard Borel, so g' is upper triangular. The eigenvalues of g' are thus given by its diagonal entries $g'_1, \ldots, g'_m$. Comparing the $(i, i+1)$ entries of $J^q_{m,\lambda}$ and J_{m,λ^q} we find that $g'_{i+1}/g'_i = \lambda^{q-1} q$. In particular, multiplying g' by a suitable scalar we may assume g' has integral eigenvalues, as desired. □

Proposition 4.6. *The images of the integral points of $X^0_{q,n}$ are dense in $X^0_{q,n}$.*

Proof. Fix a point $(\mathrm{Fr}_x, \sigma_x)$ of $X^0_{q,n}$. After conjugating σ_x appropriately we may assume that σ_x is in Jordan normal form (and thus in particular has integral entries, since we have shown that the eigenvalues of σ_x are roots of unity). Moreover, since σ_x is conjugate to its q-th power, for any eigenvalue λ of σ there is a size-preserving bijection between the Jordan blocks of σ_x of eigenvalue λ and those of eigenvalue λ^q. Let (m_i, λ_i) denote the size and eigenvalue of the i-th Jordan block of σ_x. Then we can find a permutation matrix w such that $w \sigma_x w^{-1}$ is also in Jordan normal form, but where the i-th Jordan block is of size m_i with eigenvalue λ_i^q. Let g be the block diagonal matrix whose i-th block is the matrix g_{m_i,λ_i} from the above lemma. Then $gw\sigma_x(gw)^{-1} = \sigma_x^q$. Moreover gw has unit eigenvalues, as some power of gw is

block diagonal with blocks given by powers of the matrices g_{m_i,λ_i}. Thus by Lemma 4.4 we can find an integral point $(\mathrm{Fr}'_x, \sigma'_x)$ of $X^0_{q,n}$ in the GL_n-orbit of the point (gw, σ_x).

Now consider the condition $g'\sigma'_x = \sigma'_x g'$, for arbitrary matrices g'. This is a linear condition on g' with coefficients in $\mathcal{O}_L$. The scheme parametrizing such g' is not quite a vector space scheme over $\mathcal{O}_L$ (it need not be flat over $\mathcal{O}_L$), but the closure of its general fiber is such a scheme. Let U be the open subscheme of this closure consisting of invertible g'. Then U contains the identity in particular, so its special fiber is nonempty. However, in an open subset of a vector space scheme over $\mathcal{O}_L$ whose special fiber is nonempty, the $\mathcal{O}_L$-points form a dense subset. Thus integral points are dense in U.

On the other hand, the points $(\mathrm{Fr}'_x\, u, \sigma'_x)$, as u runs over the integral points of U, are all integral points of $X^0_{q,n}$, and (since integral points of U are dense in U) their closure is the set of all points $(\mathrm{Fr}'_y, \sigma'_x)$ in $X^0_{q,n}$. Conjugating by integral points of GL_n, which are clearly dense in GL_n, we find that the closure of the integral points contains the entire locus of points $(\mathrm{Fr}''_x, \sigma''_x)$ with σ''_x conjugate to σ_x. Since σ_x was chosen arbitrarily the result follows. □

Corollary 4.7. *The ring $R_{q,n}$ is ℓ-adically separated; that is, the intersection of the ideals $\ell^i R_{q,n}$ is zero.*

Proof. Let f be an element of $R_{q,n}$ that is divisible by ℓ^i for all i. Then, for any integral point $x : R_{q,n} \to \mathcal{O}_L$, the image $x(f)$ is divisible by ℓ^i for all i and is therefore zero. In other words, f vanishes on a dense subset of $X^0_{q,n}$. Since $X^0_{q,n}$ is reduced, f is zero. □

Now fix a Frobenius element $\widetilde{\mathrm{Fr}}$ in W_F, and a topological generator $\tilde{\sigma}$ of the quotient $I_F/I_F^{(\ell)}$. Let t_ℓ be the isomorphism of $I_F/I_F^{(\ell)}$ with the additive group of $\mathbb{Z}_\ell$ that takes $\tilde{\sigma}$ to 1. By Corollary 4.3, for some positive integer a the matrix σ^{ℓ^a} in $\mathrm{GL}_n(R_{q,n})$ is unipotent; that is, its characteristic polynomial is $(X-1)^n$. The following lemma allows us to make sense of $(\sigma^{\ell^a})^b$ for any $b \in \mathbb{Z}_\ell$:

Lemma 4.8. *Let R be a flat, ℓ-adically separated $\mathbb{Z}_\ell$-algebra, and $M \in \mathrm{GL}_n(R)$ such that $(M-1)^n = 0$. Then there exists a unique ℓ-adically continuous homomorphism $\phi_M : \mathbb{Z}_\ell \to \mathrm{GL}_n(R)$ such that for all $b \in \mathbb{Z}$, $\phi_M(b) = M^b$.*

Proof. Consider the power series $\exp t \log(1+X)$ in $\mathbb{Q}[t][[X]]$, and let $p_i(t)$ be the coefficient of X^i in this power series. For any i, and any integer b, Let N_i be the $(i+1)$ by $(i+1)$ Jordan block with eigenvalue zero; then $p_i(b)$ is the upper right entry of $(1+N_i)^b$, and is thus an integer. In particular each p_i is a $\mathbb{Z}_\ell$-valued function on $\mathbb{Z}_\ell$. Given M as above, and $t \in \mathbb{Z}_\ell$, we may thus define ϕ_M by

$$\phi_M(t) = 1 + p_1(t)(M-1) + \cdots + p_{n-1}(t)(M-1)^{n-1},$$

and it is clear that this has the claimed properties. □

(Recall that for an ℓ-adically separated ring A, and a locally profinite group H, a representation $\rho : H \to \mathrm{GL}_n(A)$ is ℓ-adically continuous if, for all positive integers i, the preimage of the subgroup $\mathrm{Id} + \ell^i M_n(A)$ of $\mathrm{GL}_n(A)$ is open in H.)

We will henceforth write $(\sigma^{\ell^a})^b$ for $\phi_{(\sigma^{\ell^a})}(b)$, given $b \in \mathbb{Z}_\ell$.

We thus have an ℓ-adically continuous representation $\rho_{F,n} : W_F \to \mathrm{GL}_n(R_{q,n})$ defined by

$$\rho_{F,n}(\widetilde{\mathrm{Fr}}^i w) = \mathrm{Fr}^i \sigma^j (\sigma^{\ell^a})^b$$

for any $w \in I_F$ and any $j \in \mathbb{Z}$, $b \in \mathbb{Z}_\ell$ such that $j + \ell^a b = t_\ell(w)$. Note that, by the above lemma, this is the *unique* ℓ-adically continuous representation that takes $\widetilde{\mathrm{Fr}}$ to Fr and $\tilde{\sigma}$ to σ.

The pair $(R_{q,n}, \rho_{F,n})$ has the following universal property, which is easily seen to characterize the pair up to isomorphism:

Proposition 4.9. *For any finitely generated, ℓ-adically separated $W(k)$-algebra A, and any framed, ℓ-adically continuous representation $\rho : W_F/I_F^{(\ell)} \to \mathrm{GL}_n(A)$, there is a unique map: $R_{q,n} \to A$ such that ρ is the base change of $\rho_{F,n}$.*

Proof. Given ρ, we have a pair of matrices $(\rho(\widetilde{\mathrm{Fr}}), \rho(\tilde{\sigma}))$ in $\mathrm{GL}_n(A)$, satisfying

$$\rho(\widetilde{\mathrm{Fr}})\rho(\tilde{\sigma})\rho(\widetilde{\mathrm{Fr}})^{-1} = \rho(\tilde{\sigma})^q,$$

and hence a map $S_{q,n} \to A$. Moreover, since the restriction of ρ to I_F factors through $I_F/I_F^{(\ell)}$ and is ℓ-adically continuous, the eigenvalues of $\rho(\tilde{\sigma})$ are ℓ-power roots of unity. Thus the map $S_{q,n} \to A$ factors through $R_{q,n}$ and the result follows. □

If we regard the $\overline{\mathcal{K}}$-points of $X^0_{q,n}$ as framed representations of $W_F/I_F^{(\ell)}$, then one can show:

Proposition 4.10. *Let x be a $\overline{\mathcal{K}}$-point of $X^0_{q,n}$. Then there is a point y in the closure of the GL_n-orbit of x such that the representation ρ_y is semisimple.*

Proof. Replacing x with a point in the same GL_n-orbit, we may assume that the framing on ρ_x is such that ρ_x is block upper triangular, with block sizes $n_1, \ldots n_r$, and that for $1 \le i \le r$, the restriction ρ_i of ρ_x to the i-th diagonal block is irreducible. Let M be the block diagonal matrix whose i-th block is given by t^i times the n_i by n_i identity matrix, for some parameter t. Then the limit, as t approaches zero, of $M\rho_x M^{-1}$ exists and is semisimple. □

We will later need the following observation about the representation $\rho_{F,n}$.

Proposition 4.11. *As x varies over the $\overline{\mathcal{K}}$-points of $X^0_{q,n}$, the restriction of ρ_x^{ss} to I_F is constant on connected components of $X^0_{q,n} \times_{W(k)} \overline{\mathcal{K}}$.*

Proof. The restriction of ρ_x^{ss} to I_F is determined by the characteristic polyomial of σ_x; since the eigenvalues of σ_x have bounded ℓ-power order there are only finitely possible characteristic polynomials of σ_x. □

5. The inertial subalgebra of $S_{q,n}$

Our next goal is to study the finite rank $W(k)$-subalgebra of $S_{q,n}$ generated by the coefficients of the characteristic polynomial of σ. Consider the map

$$W(k)[r_1, \ldots, r_n, r_n^{-1}] \to S_{q,n}$$

that takes r_i to the coefficient of X^{n-i} in this characteristic polynomial.

By the theory of symmetric functions, for $1 \leq i \leq n$ there are unique polynomials $P_{i,q}$ in the variables $r_1, \ldots, r_n$ with the following property: for all $t_1, \ldots, t_n \in \overline{\mathcal{K}}$, define $r_1, \ldots, r_n \in \overline{\mathcal{K}}$ by the identity

$$(X - t_1) \cdots (X - t_n) = X^n + r_1 X^{n-1} + r_2 X^{n-2} + \cdots + r_n.$$

Then the $P_{i,q}$ are the unique polynomials satisfying

$$(X - t_1^q) \cdots (X - t_n^q) = X^n + P_{1,q}(r_1, \ldots, r_n) X^{n-1} + \cdots + P_{n,q}(r_1, \ldots, r_n).$$

Since σ is conjugate to its q-th power, for $1 \leq i \leq n$ the element $P_{i,q}(r_1, \ldots, r_n) - r_i$ lies in the kernel of the map $W(k)[r_1, \ldots, r_n, r_n^{-1}] \to S_{q,n}$. Let $I_{q,n}$ denote the ideal of $W(k)[r_1, \ldots, r_n, r_n^{-1}]$ generated by the $P_{i,q}(r_1, \ldots, r_n) - r_i$, and let $B_{q,n}$ denote the quotient $W(k)[r_1, \ldots, r_n, r_n^{-1}]/I_{q,n}$. We will show that in fact the map $B_{q,n} \to S_{q,n}$ is injective, and that moreover its image in $S_{q,n}$ is saturated.

We will now realize $B_{q,n}$ as a quotient of $S_{q,n}$ in a natural way. We are grateful to Jack Shotton for making us aware of the following construction, which is adapted from Proposition 7.10 in [Shotton 2018]. (Shotton uses a slightly different form for the matrix σ, that is less convenient for our purposes, but the arguments are otherwise exactly analogous.)

Let $Y \subseteq \operatorname{Spec} S_{q,n}$ denote the locus on which σ has the form

$$\begin{pmatrix} 0 & 0 & 0 & \ldots & 0 & -r_n \\ 1 & 0 & 0 & \ldots & 0 & -r_{n-1} \\ 0 & 1 & 0 & \ldots & 0 & -r_{n-2} \\ & \vdots & & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -r_1 \end{pmatrix}.$$

(that is, on which σ is the "companion matrix" of the polynomial $X^n + r_1 X^{n-1} + \cdots + r_n$.) We may embed Y as an open subscheme of the scheme Y' parametrizing pairs of matrices (Fr, σ) such that σ is invertible of the above form, the characteristic polynomial of σ is equal to that of σ^q, and $\mathrm{Fr}\,\sigma = \sigma^q\,\mathrm{Fr}$. Then Y is simply the open subscheme of Y' on which Fr is invertible. The scheme Y' then maps to $\operatorname{Spec} B_{q,n}$ via the map that takes (Fr, σ) to the tuple $(r_1, \ldots, r_n)$.

We have a map $Y' \to \operatorname{Spec} B_{q,n} \times_{W(k)} \mathbb{A}^n_{W(k)}$ that takes (Fr, σ) to the point $(r_1, \ldots, r_n, \mathrm{Fr}(e_1))$, where $e_1, \ldots, e_n$ is the standard basis for $W(k)^n$. In fact, one then has:

Proposition 5.1. *The map $Y' \to \operatorname{Spec} B_{q,n} \times_{W(k)} \mathbb{A}^n_{W(k)}$ is an isomorphism.*

Proof. We describe an inverse map. Given $(r_1, \ldots, r_n, v)$ in $\operatorname{Spec} B_{q,n} \times_{W(k)} \mathbb{A}^n_{W(k)}$ we associate the pair (Fr, σ), where σ has the above form with $-r_n, \ldots, -r_1$ in the right column, and Fr is defined by $\mathrm{Fr}\, e_i = \sigma^{(i-1)q} v$ for $1 \leq i \leq n$. One verifies easily that for $1 \leq i \leq n-1$, we have $\mathrm{Fr}\,\sigma(e_i) = \sigma^q\,\mathrm{Fr}(e_i)$. On the other hand, we have

$$\sigma^q\,\mathrm{Fr}\, e_n - \mathrm{Fr}\,\sigma e_n = ((\sigma^q)^n + r_1(\sigma^q)^{n-1} + \cdots + r_n)v = P_\sigma(\sigma^q)v,$$

where P_σ is the characteristic polynomial of σ. The relations on $\tilde{B}_{q,n}$ guarantee that $P_\sigma = P_{\sigma^q}$, so $P_\sigma(\sigma^q)v = 0$ by Cayley–Hamilton.

We thus have a well-defined map that is clearly a right inverse to the map constructed above. To see that it is also a left inverse, note that if $\mathrm{Fr}\,\sigma = \sigma^q\,\mathrm{Fr}$, and $\mathrm{Fr}(e_1) = v$, then we must have

$$\mathrm{Fr}\, e_i = \mathrm{Fr}\,\sigma(e_{i-1}) = \sigma^q\,\mathrm{Fr}\, e_{i-1}$$

so by induction Fr is determined by $\mathrm{Fr}(e_1)$. □

Lemma 5.2. *Let B be a finite rank $W(k)$-algebra, and V an open subset of* $\mathrm{Spec}\, B \times_{W(k)} \mathbb{A}^n_{W(k)}$ *such that the projection $V \to \mathrm{Spec}\, B$ is surjective. Then the map from B to $\mathcal{O}_V$ induced by the projection of V onto* $\mathrm{Spec}\, B$ *is injective. If moreover, B is flat over $W(k)$, then the image of B in $\mathcal{O}_V$ is saturated.*

Proof. For each closed point x of $\mathrm{Spec}\, B$, there exists an element $\bar{a}_x$ of k^n such that $(x, \bar{a}_x)$ lies in V. Lift $\bar{a}_x$ to a $W(k)$-point a_x of $\mathbb{A}^n_{W(k)}$, and let $V_x = V \cap (\mathrm{Spec}\, B \times_{W(k)} a_x)$. Then the projection of V to $\mathrm{Spec}\, V$ identifies V_x with an open subset of $\mathrm{Spec}\, B$, and as x varies, the V_x cover $\mathrm{Spec}\, B$. If b is an element of B that maps to zero in $\mathcal{O}_V$, then it vanishes in particular on each V_x and hence on $\mathrm{Spec}\, B$, so injectivity is clear.

Now consider an element b of $B/\ell B$, and suppose B maps to zero in $\mathcal{O}_V/\ell\mathcal{O}_V$. Then b maps to zero in $\mathcal{O}_{V_x}/\ell\mathcal{O}_{V_x}$ for all x, but since the V_x are an open cover of $\mathrm{Spec}\, B$ this means b is zero in $B/\ell B$. □

We can now show:

Proposition 5.3. *The map $B_{q,n} \to S_{q,n}$ is injective with saturated image.*

Proof. We first show that the projection map from Y to $\mathrm{Spec}\, B$ is surjective. Indeed, for any algebraically closed field L that is a $W(k)$-algebra, and any L-point $(r_1, \ldots, r_n)$ of $\mathrm{Spec}\, B$, the corresponding σ is a regular element of L whose characteristic polynomial is equal to that of σ^q. In particular the eigenvalues of σ are roots of unity of order prime to q. It is then clear, by considering the Jordan normal form of σ, that σ^q is also regular. Over L any two regular matrices with the same characteristic polynomial are conjugate, so there exists an element Fr of $\mathrm{GL}_n(L)$ that conjugates σ to σ^q. Then (Fr, σ) is an L-point of T mapping to $(r_1, \ldots, r_n)$.

The lemma now shows that the map from $B_{q,n}$ to $\mathcal{O}_Y$ is injective; since this map factors through $S_{q,n}$ we see that $B_{q,n}$ embeds in $S_{q,n}$. Thus $B_{q,n}$ is flat over $W(k)$, and the lemma then shows that its image in $\mathcal{O}_Y$ is saturated. Once again using that the map from $B_{q,n}$ to $\mathcal{O}_Y$ factors through $S_{q,n}$ we see that $B_{q,n}$ is also saturated in $S_{q,n}$. □

The map $B_{q,n} \to S_{q,n}$ induces a map $B_{q,n,1} \to R_{q,n}$, where $B_{q,n,1}$ is the direct factor of $B_{q,n}$ whose $\overline{\mathcal{K}}$-points correspond to conjugacy classes whose reduction modulo ℓ is the identity. Proposition 4.11, together with Proposition 5.3, shows that $B_{q,n,1}$ is precisely the subalgebra of $R_{q,n}$ consisting of elements whose value at a $\overline{\mathcal{K}}$-point x of $\mathrm{Spec}\, R_{q,n}$ depends only on the semisimplification of the restriction of ρ_x to I_F.

6. The symmetrizing form on $B_{q,n}$

We now relate $B_{q,n}$ with the endomorphism ring $\bar{E}_{q,n}$ of the Gelfand–Graev representation. We first work over $\bar{\mathcal{K}}$; since both $B_{q,n}$ and $\bar{E}_{q,n}$ are reduced, constructing an isomorphism of $B_{q,n} \otimes \bar{\mathcal{K}}$ with $\bar{E}_{q,n} \otimes \bar{\mathcal{K}}$ amounts to constructing a bijection on their $\bar{\mathcal{K}}$-points.

Recall that the $\bar{\mathcal{K}}$-points of $\operatorname{Spec} \bar{E}_{q,n}$ are in bijection with the isomorphism classes of irreducible generic representations of $\bar{G}$ and therefore (via Deligne–Lusztig restriction) with the equivalence classes of pairs (w, φ) where w is an element of the Weyl group of $\bar{G}$ and $\varphi : \bar{T}_w \to \bar{\mathcal{K}}^\times$ is a character. On the other hand, a $\bar{\mathcal{K}}$-point of $\operatorname{Spec} B_{q,n}$ is represented by an invertible diagonal matrix, with entries in $\bar{\mathcal{K}}$, that is conjugate to its q-th power; that is, it is an invertible diagonal matrix t such that there exists a permutation matrix w with $t^w = t^q$.

In order to construct a natural bijection between these two sets we must fix some choices. First, we identify $\mathrm{GL}_n(\bar{\mathcal{K}})$ with the Langlands dual group $\widehat{G}$ of G, with (diagonal) maximal torus $\widehat{T}$. Second, we choose a topological generator $\tilde{\sigma}$ of the tame inertia group I_F/P_F of F. Local class field theory gives an isomorphism

$$I_F/P_F \cong \varprojlim \mathbb{F}_{q^n}^\times \cong \varprojlim \operatorname{Hom}\left(\frac{1}{q^n-1}\mathbb{Z}/\mathbb{Z}, \bar{\mathbb{F}}_q^\times\right),$$

where the first limit is over the norm maps, and the transition maps in the second limit, for m dividing n, are given by "multiplication by $(q^n-1)/(q^m-1)$".

On the other hand we have a chain of natural isomorphisms,

$$\operatorname{Hom}((\mathbb{Q}/\mathbb{Z})^{(p)}, \bar{\mathbb{F}}_q^\times) = \operatorname{Hom}\left(\varinjlim \left(\frac{1}{q^n-1}\mathbb{Z}/\mathbb{Z}\right), \bar{\mathbb{F}}_q^\times\right) \cong I_F/P_F,$$

so our choice of $\tilde{\sigma}$ gives us a natural map $(\mathbb{Q}/\mathbb{Z})^{(p)} \to \bar{\mathbb{F}}_q^\times$ that is easily seen to be an isomorphism.

Now fix a w in the Weyl group $W(\bar{G})$; we identify $W(\bar{G})$ with the group of permutation matrices in $\mathrm{GL}_n(\bar{\mathcal{K}})$. Let X be the character group of the torus $\overline{T}_w$ of $\bar{\mathcal{G}}$; then X is dual to the character group X' of the group of diagonal matrices in $\mathrm{GL}_n(\bar{\mathcal{K}})$. We have an isomorphism $\overline{T}_w(\bar{\mathbb{F}}_q) \cong \operatorname{Hom}(X/(\mathrm{Fr}_q - 1)X, \bar{\mathbb{F}}_q^\times)$, where Fr_q is the q-power Frobenius. If we denote by $\mu^{(p)}$ the prime-to-p roots of unity in $\bar{\mathcal{K}}^\times$, then we have an isomorphism

$$\operatorname{Hom}(\overline{T}_w, \mu^{(p)}) \cong X/(\mathrm{Fr}_q - 1)X \otimes \operatorname{Hom}(\bar{\mathbb{F}}_q^\times, \mu^{(p)}).$$

Noting that Fr_q acts on X by qw, and applying the duality isomorphism

$$X/(qw-1)X \cong \operatorname{Hom}(X'/(qw-1)X', (\mathbb{Q}/\mathbb{Z})^{(p)})$$

as well as our isomorphism of $(\mathbb{Q}/\mathbb{Z})^{(p)}$ with $\bar{\mathbb{F}}_q^\times$ arising from our choice of s, we see that $\operatorname{Hom}(\overline{T}_w, \mu^{(p)})$ is naturally isomorphic to $\operatorname{Hom}(X'/(qw-1)X', \mu^{(p)})$. An element of the latter is precisely a diagonal matrix t, with entries in $\bar{\mathcal{K}}$, such that $(t^w)^q = t$. We let $T_q^{w^{-1}}$ denote the set of such matrices.

This construction associates to every w, and every character $\varphi : \overline{T}_w \to \bar{\mathcal{K}}^\times$, an element of $T_q^{w^{-1}}$. One easily verifies that it sends equivalent pairs $(\overline{T}_w, \varphi)$ and $(\overline{T}_{w'}, \varphi')$ to conjugate diagonal matrices, and

further induces a bijection between $\overline{\mathcal{K}}$-points of Spec $\overline{E}_{q,n}$ and those of Spec $B_{q,n}$. We thus obtain an isomorphism of $\overline{E}_{q,n} \otimes \overline{\mathcal{K}}$ with $B_{q,n} \otimes \overline{\mathcal{K}}$. This isomorphism is $\mathrm{Gal}(\overline{\mathcal{K}}/\mathcal{K})$-equivariant and thus descends to an isomorphism of $\overline{E}_{q,n}\big[\frac{1}{\ell}\big]$ with $B_{q,n}\big[\frac{1}{\ell}\big]$.

Remark 6.1. The choices made in defining the bijection above means that this bijection is compatible with local Langlands in the following sense: let π be an irreducible depth zero generic representation of G over $\overline{\mathcal{K}}$, and let ρ be its Langlands parameter. If K_1 denotes the kernel of the map $\mathcal{G}(\mathcal{O}_G) \to \bar{\mathcal{G}}(\mathbb{F}_q)$, then π^{K_1} is an irreducible generic $\overline{\mathcal{K}}$-representation of $\bar{G}$, and hence gives rise to a $\overline{\mathcal{K}}$-point of Spec $\overline{E}_{q,n}$. On the other hand, the conjugacy class of the semisimplification of $\rho(\sigma)$ gives a $\overline{\mathcal{K}}$-point of Spec $B_{q,n}$. The bijection constructed above identifies these two points for every choice of π and ρ.

Since $B_{q,n}$ and $\overline{E}_{q,n}$ are ℓ-torsion free, we may regard them as $W(k)$-lattices in $B_{q,n}\big[\frac{1}{\ell}\big] \cong \overline{E}_{q,n}\big[\frac{1}{\ell}\big]$. A priori it is not clear that either lattice is contained in the other. We will show later that in fact these lattices coincide, but this is quite difficult — it will emerge from the same inductive argument that proves both the weak and strong conjecture in Section 10. For the moment, it will suffice to prove something much weaker.

Recall that one has a symmetrizing form $\theta : \overline{E}_{q,n} \to W(k)$; the inclusion $B_{q,n} \to \overline{E}_{q,n}\big[\frac{1}{\ell}\big]$ allows us to regard θ as a map from $B_{q,n}$ to $\overline{\mathcal{K}}$. The goal of the remainder of this section is to prove:

Theorem 6.2. *The map $\theta : B_{q,n} \to \overline{\mathcal{K}}$ takes values in $W(k)$.*

As a corollary, we immediately deduce

Corollary 6.3. *Suppose that the isomorphism $B_{q,n}\big[\frac{1}{\ell}\big] \cong \overline{E}_{q,n}\big[\frac{1}{\ell}\big]$ identifies $\overline{E}_{q,n}$ with a subring of $B_{q,n}$. Then this isomorphism identifies $\overline{E}_{q,n}$ with $B_{q,n}$.*

Proof. (see Lemma 3.8 of [Bonnafé and Kessar 2008]) If $\overline{E}_{q,n}$ is contained in $B_{q,n}$, then $\theta(be)$ lies in $W(k)$ for all $b \in B_{q,n}$, $e \in \overline{E}_{q,n}$; thus $B_{q,n}$ is contained in the dual lattice to $\overline{E}_{q,n}$ with respect to θ. But since θ is a symmetrizing form on $\overline{E}_{q,n}$, this dual lattice is $\overline{E}_{q,n}$. Thus $B_{q,n}$ and $\overline{E}_{q,n}$ must coincide inside $\overline{E}_{q,n}\big[\frac{1}{\ell}\big]$. □

In order to prove Theorem 6.2 we compute the values of θ on a $W(k)$-spanning set for $B_{q,n}$. By definition we have a surjection

$$W(k)[X']^{S_n} = W(k)[r_1, \ldots, r_n, r_n^{-1}]^{S_n} \to B_{q,n}$$

with kernel $I_{q,n}$. For each character $\lambda \in X'$, let N_λ denote the subgroup of S_n normalizing λ. Then the elements

$$r_\lambda = \frac{1}{\#N_\lambda} \sum_{w \in S_n} \lambda^w$$

form a $W(k)$-basis of $W(k)[X']^{S_n}$, as λ runs over the elements of X', so their images in $B_{q,n}$ (which we also, slightly abusively, denote by r_λ) span $B_{q,n}$ over $W(k)$.

Lemma 6.4. *For $\lambda \in X'$, let M_λ denote the number of $w \in S_n$ such that the restriction of λ to the subgroup T_q^w of $\mathrm{Hom}(X', \overline{\mathcal{K}}^\times)$ is trivial. Then we have*

$$\theta(r_\lambda) = \frac{M_\lambda}{\#N_\lambda}.$$

Proof. Let x be a $\overline{\mathcal{K}}$-point of Spec $B_{q,n}$, and let e_x denote the element of $B_{q,n} \otimes \overline{\mathcal{K}}$ that takes the value 1 at x and zero at all other $\overline{\mathcal{K}}$-points of Spec $B_{q,n}$. Our construction of the isomorphism $\overline{E}_{q,n} \otimes \overline{\mathcal{K}} \cong B_{q,n} \otimes \overline{\mathcal{K}}$, together with Proposition 2.4, shows that

$$\theta(e_x) = \frac{1}{n!} \sum_{w \in S_n} \frac{N(w^{-1}, x)}{\#\overline{T}_w(\mathbb{F}_q)} = \frac{1}{n!} \sum_{w \in S_n} \frac{N'(w, x)}{\#T_q^{w^{-1}}},$$

where $N'(w, x)$ denotes the number of elements of T_q^w in the equivalence class corresponding to x. It follows that we have

$$\theta(r_\lambda) = \frac{1}{n!} \sum_x r_\lambda(x) \sum_{w \in S_n} \frac{N'(w, x)}{\#T_q^{w^{-1}}}.$$

Since $r_\lambda(t)$ depends only on the equivalence class of $t \in T_q^w$, we can rewrite this as

$$\theta(r_\lambda) = \frac{1}{n!} \sum_{w \in S_n} \frac{1}{\#T_q^{w^{-1}}} \sum_{t \in T_q^{w^{-1}}} \frac{1}{\#N_\lambda} \sum_{v \in S_n} \lambda^v(t).$$

Changing the order of the summation, we obtain

$$\theta(r_\lambda) = \frac{1}{n! \#N_\lambda} \sum_{v \in S_n} \sum_{w \in S_n} \frac{1}{\#T_q^{w^{-1}}} \sum_{t \in T_q^{w^{-1}}} \lambda^v(t),$$

and the innermost sum is equal to 0 if λ^v is nontrivial on $T_q^{w^{-1}}$ and equal to $\#T_q^{w^{-1}}$ otherwise. Thus the sum over w is equal to M_{λ^v} which is equal to M_λ. We thus have $\theta(r_\lambda) = M_\lambda / \#N_\lambda$ as claimed. □

In light of this result, the proof of Theorem 6.2 is reduced to the following result:

Lemma 6.5. *For any $\lambda \in X'$, the order of N_λ divides M_λ.*

It is clear that the set of w such that λ is trivial on T_w^q is stable under conjugation by elements of N_λ, but of course this action is not faithful, so the divisibility is not immediate.

We begin by observing that N_λ is the Weyl group of the Levi subgroup of GL_n centralizing λ. This Levi corresponds to a partition of the $\{1, 2, \ldots, n\}$ into subsets, and N_λ is then the subgroup of S_n that preserves this partition. In particular if w lies in N_λ, then any cycle occurring in the cycle decomposition of w also lies in N_λ.

Now let $N_{\lambda,w}$ denote the centralizer of w in N_λ. Let $O(w)$ be the partition of the set $\{1, \ldots, n\}$ into orbits under the action of w; then conjugation by $N_{\lambda,w}$ permutes the orbits of w, yielding a map $N_{\lambda,w} \to \mathrm{Aut}(O(w))$, where $\mathrm{Aut}(O(w))$ is the group of permutations of $O(w)$.

Definition 6.6. We will say that w is N_λ-minimal if the map $N_{\lambda,w} \to \mathrm{Aut}(O(w))$ is injective.

Note that the property of being N_λ-minimal is stable under N_λ-conjugacy. Given an arbitrary N_λ-conjugacy class in S_n, we will associate an N_λ-minimal conjugacy class in a natural way. On the level of specific permutations w this construction will depend not just on w but on a particular choice of cycle representation for w. Here by a "cycle representation" of w we mean an unordered collection of expressions of the form $(x_1 \dots x_r)$, with $x_1, \dots, x_r$ distinct elements of $\{1, \dots, n\}$, that correspond to a disjoint set of cycles whose product is w. To give a cycle representation of w is equivalent to specifying, for each orbit x of w on $\{1, \dots, n\}$, a distinguished element x_1 of the orbit x.

Now fix $w \in S_n$, along with a cycle representation of w, and let K be the kernel of the map from $N_{\lambda,w}$ to $\mathrm{Aut}(O(w))$. Then K acts on each orbit $O(w)$; such an orbit x comes from a cycle $(x_1 \dots x_r)$ in our chosen cycle representation of w. Since K centralizes w, it must "cyclically permute" the elements of this orbit; that is, the action of K factors through a map $K \to \mathbb{Z}/r\mathbb{Z}$, where $s \in \mathbb{Z}/r\mathbb{Z}$ acts by sending each x_i to x_{i+s}, and the indices are considered modulo r. Let m be the order of the image of the map $K \to \mathbb{Z}/r\mathbb{Z}$, and set $s = \frac{r}{m}$. Let $x^{\min}$ be the permutation given by the product of the m disjoint cycles $(x_1 \dots x_s)(x_{s+1} \dots x_{2s}) \dots (x_{r-s+1} \dots x_r)$. We then define $w^{\min}$ to be the product, over all cycles $x \in O(w)$, of the permutations $x^{\min}$. It is clear from the construction that if w is N_λ-minimal then $w^{\min} = w$.

This construction depends on our choice of cycle representation of w; in particular if we represented the cycle $x = (x_1 \dots x_r)$ as $(x_{t+1} \dots x_{t+r})$ instead then we would obtain the product of cycles

$$(x_{t+1} \dots x_{t+s})(x_{t+s+1} \dots x_{t+2s}) \dots (x_{r+t-s+1} \dots x_{r+t})$$

instead of the product

$$(x_1 \dots x_s)(x_{s+1} \dots x_{2s}) \dots (x_{r-s+1} \dots x_r).$$

Note that the former is N_λ-conjugate to the latter, via the permutation that, for each $0 \le a < \frac{r}{s}$ and each $1 \le b \le s$, takes x_{as+b} to x_{t+as+c}, where c is the unique integer between 1 and s such that $as + b$ is congruent to $as + t + c$ modulo s. However, the two permutations are of course not equal. Thus changing the cycle representation of w conjugates $w^{\min}$ by an element of N_λ. In particular the N_λ-conjugacy class $[w^{\min}]$ depends only on w and not its cycle representation.

On the other hand, if we fix a $v \in N_\lambda$, and a cycle representation of w, then conjugating this cycle representation by v gives a cycle representation of vwv^{-1}. Then if we compute $w^{\min}$ and $(vwv^{-1})^{\min}$ using these cycle representations it is easy to see that $(vwv^{-1})^{\min} = vw^{\min}v^{-1}$. In particular $[w^{\min}]$ depends only on the N_λ-conjugacy class of w.

Lemma 6.7. *For any $w \in S_n$, $w^{\min}$ is N_λ-minimal.*

Proof. Suppose for a contradiction that $w^{\min}$ is not N_λ-minimal, and let K be the kernel of the map $N_{\lambda,w^{\min}} \to \mathrm{Aut}(O(w^{\min}))$. Choose an element k of K other than the identity. By definition k preserves every orbit in $O(w)$ and acts nontrivially on at least one such orbit $x = (x_1 \dots x_r)$; we have an s such that $kx_i = x_{i+s}$ for all i. Let k' denote the permutation that sends x_i to x_{i+s} for all i and fixes all other elements. Then k' lies in N_λ, since k does and k' is a product of cycles of k. Moreover it is clear that k' commutes with $w^{\min}$.

Our construction of $w^{\min}$ from w implies that the $w^{\min}$-cycle x is contained in a w-cycle x' of the form $(x_1 \dots x_{r'})$ for some multiple r' of r, and that the cycles $(x_{r+1} \dots x_{2r})$, etc. are cycles of $w^{\min}$. Let k'' be the permutation that takes x_i to x_{i+s} for all $1 \leq i \leq r'$; then it is clear that k'' centralizes w. We will show that in fact k'' lies in N_λ; this gives a contradiction as then we have an element of $N_{\lambda,w}$ that acts by a shift of length s on the cycle x', meaning that in passing from w to $w^{\min}$ the cycle x' should decompose into cycles of length dividing s, and not cycles of length r as we have supposed.

To show that k'' lies in N_λ it suffices to show that for all i, x_{i+s} and x_i lie in the same N_λ-orbit. For $1 \leq i \leq r-s$ this is clear since k' lies in N_λ. On the other hand, since x' decomposes into cycles of length r in the cycle decomposition of $w^{\min}$, there is an element of N_λ that carries x_i to x_{i+r} for all i. The claim follows. □

The association $w \mapsto [w^{\min}]$ defines an equivalence relation $\sim$ on S_n, such that $w \sim v$ if, and only if, $[w^{\min}] = [v^{\min}]$. It is clear that each equivalence class for $\sim$ is a union of N_λ-orbits. We will show that in fact each equivalence class has cardinality equal to $\#N_\lambda$. We begin by fixing an N_λ-minimal w. Then we have an injection $N_{\lambda,w} \to \mathrm{Aut}(O(w))$. We will say two orbits x, x' in $O(w)$ are $N_{\lambda,w}$-equivalent if there is an element of $N_{\lambda,w}$ that takes x to x'. We then have:

Lemma 6.8. *Suppose w is N_λ-minimal, and let v be a permutation of $O(w)$ such that for all $x \in O(w)$, vx is $N_{\lambda,w}$-equivalent to x. Then there is a unique element $\tilde{v}$ of $N_{\lambda,w}$ whose image in $\mathrm{Aut}(O(w))$ is v. In particular, $N_{\lambda,w}$ is a product of symmetric groups.*

Proof. Uniqueness is clear from the definition of N_λ-minimality. For existence, fix an orbit $x \in O(w)$. Then there is an element v'_x of $N_{\lambda,w}$ that takes x to vx. We can then define $\tilde{v}$ to be the bijection on $\{1, 2, \dots, n\}$ that agrees with v'_x on x for all orbits x. Note that for all $1 \leq i \leq n$, we have $\tilde{v}(i) = v'_x(i)$ for x the w-orbit containing i; since v'_x is in N_λ we have $\lambda_i = \lambda_{v'_x(i)} = \lambda_{\tilde{v}(i)}$, so $\tilde{v}$ lies in N_λ. □

We now fix a particular N_λ-minimal w, and a particular cycle representation of w. Since w is N_λ-minimal we may (and do) choose this cycle representation so that it is preserved by the action of $N_{\lambda,w}$. Then given any $v \in N_{\lambda,w}$, define $\widetilde{w}(v)$ to be the permutation constructed as follows: for orbit of v on $O(w)$, choose an x representing that orbit. The orbit x then corresponds to a term $(x_1 \dots x_r)$ in our chosen cycle representation of w. Let $\widetilde{w}(v)_x$ be the permutation $(x_1 \dots x_r\ vx_1 \dots vx_r \dots v^{d-1}x_1 \dots v^{d-1}x_r)$, where d is the order of the v-orbit of x. Let $\widetilde{w}(v)$ be the product, over a set of representatives x for the orbits of v on $O(w)$, of $\widetilde{w}(v)_x$. Note that as a permutation, $\widetilde{w}(v)$ is independent of our choices of representatives x but does depend on our choice of cycle representation of w. On the other hand, our initial choice of cycle representation of w, together with the choices of representatives x, gives rise to a cycle representation of $\widetilde{w}(v)$.

Lemma 6.9. *Let u be an element of N_λ. Then u conjugates $\widetilde{w}(v)$ to $\widetilde{w}(v')$ if, and only if, u normalizes w and conjugates v to v'. Moreover, we have $\widetilde{w}(v)^{\min} = w$.*

Proof. First assume that u normalizes w. Then u actually fixes our chosen cycle representation of w, since w is N_λ-minimal. It is then easy to see from the construction that $\widetilde{w}(uvu^{-1}) = u\widetilde{w}(v)u^{-1}$.

Conversely, assume u conjugates $\widetilde{w}(v)$ to $\widetilde{w}(v')$. Let $x = (x_1 \dots x_r)$ be a cycle in our chosen representation of w, such that the induced cycle of $\widetilde{w}(v)$ is $(x_1 \dots x_r\ vx_1 \dots vx_r \dots v^{d-1}x_1 \dots v^{d-1}x_r)$. Since u

conjugates $\widetilde{w}(v)$ to $\widetilde{w}(v')$the cycle $(ux_1 \dots ux_r\, uvx_1 \dots uvx_r \dots uv^{d-1}x_1 \dots uv^{d-1}x_r)$ is a cycle of $\widetilde{w}(v')$. This cycle contains a cycle $(y_1 \dots y_{r'})$ of our chosen representation of w. Thus, by construction of $\widetilde{w}(v')$, there is an $s \in \mathbb{Z}/dr\mathbb{Z}$ such that the sequence

$$ux_1, \dots ux_r, uvx_1, \dots uvx_r, \dots uv^{d-1}x_1, \dots uv^{d-1}x_r$$

coincides with the cyclic shift by s of the sequence

$$y_1, \dots y_{r'}, vy_1, \dots vy_{r'}, \dots v^{d'-1}y_1, \dots v^{d'-1}y_{r'},$$

where $dr = d'r'$.

Since u and v both lie in N_λ, it follows that for all $1 \le i \le r$, and all integers j, x_i lies in the same N_λ-orbit as $y_{i+s+jr'}$, where the indices are taken modulo r. Let $a = (r, r')$. Then for all i, x_i lies in the same N_λ-orbit as x_{i+a}. Thus the permutation that takes x_i to x_{i+a} for all i and fixes all other elements lies in N_λ. This permutation clearly normalizes w and fixes all orbits of w, so must be the identity since w is N_λ-minimal. Thus $a = r$, so r divides r'. Similar reasoning shows that r' divides r, so in fact r equals r'.

Now for all $1 \le i \le r$, x_i is in the same N_λ-orbit as y_{i+s}; there is thus an element of $N_{\lambda,w}$ that carries the cycle $(x_1 \dots x_r)$ of w to the cycle $(y_1 \dots y_r)$. Since we chose our cycle representation of w to be $N_{\lambda,w}$-stable, there is also an element of $N_{\lambda,w}$ that takes x_i to y_i for all i. There is thus an element of $N_{\lambda,w}$ that takes x_i to x_{i+s} for all i, and fixes all other elements of $\{1, \dots, n\}$. Since w is minimal, this is impossible unless r divides s.

We have thus established that u takes the cycle $x = (x_1 \dots x_r)$ of w to the cycle $(v^e y_1, \dots, v^e y_r)$ for some e, which is also a cycle of w. Since x was arbitrary, u preserves the cycles of w and thus normalizes w. But now we have $\widetilde{w}(uvu^{-1}) = u\widetilde{w}(v)u^{-1} = \widetilde{w}(v')$, and it is easy to see that this implies that $uvu^{-1} = v'$.

For the final claim, let $x = (x_1 \dots x_r)$ be a cycle in our chosen representation of w, contained in the cycle $(x_1 \dots x_r vx_1 \dots vx_r \dots v^{d-1}x_1 \dots v^{d-1}x_r)$ of $\widetilde{w}(v)$. The subgroup of $N_{\lambda,w}$ preserving the latter cycle acts on it by cyclic shifts, and minimality of w implies that r divides the length of any of these shifts. On the other hand it is clear that the permutation that agrees with v on the set $\{x_1, \dots, x_r, vx_1, \dots, vx_r, \dots, v^{d-1}x_1, \dots v^{d-1}x_r\}$ and is the identity elsewhere induces a shift of length r on this cycle. Our construction of $\widetilde{w}(v)^{\min}$ thus demands that we break this cycle of $\widetilde{w}(v)$ into cycles of length r. Doing this for all cycles of $\widetilde{w}(v)$ recovers w. □

We now show:

Lemma 6.10. *Suppose w is N_λ-minimal and $w' \sim w$. Then there exists $v \in N_\lambda(w)$ such that w' is N_λ-conjugate to $\widetilde{w}(v)$.*

Proof. We first construct a cycle representation of w' such that the induced cycle representation of $(w')^{\min}$ is $N_{\lambda,(w')^{\min}}$-invariant. To do this, first fix any orbit of w' and choose a representation of the corresponding cycle; we then obtain representations of one or more cycles in $(w')^{\min}$, all of which are N_λ-conjugate. We then proceed inductively: for each orbit x of w', choose a cycle representation arbitrarily and consider

the resulting cycles of $(w')^{\min}$. If these cycles are not N_λ-conjugate to other cycles of $(w')^{\min}$ that have already been constructed, there is nothing further to do and we may proceed to the next orbit of w'. If they are conjugate to cycles we have already constructed, it need not be the case that the corresponding cycle *representations* are N_λ-conjugate to those already extant (they may differ by a cyclic shift). However, adjusting our choice of cycle representation of x by a suitable shift we may arrange that this holds. Proceeding inductively we arrive at a $(w')^{\min}$ and an $N_{\lambda,(w')^{\min}}$-invariant cycle representation of it.

Now for each cycle x of w', our chosen decompositions give $x = (x_1 \dots x_{rs})$ in w', for some integers r, s such that the corresponding cycles of $(w')^{\min}$ are $(x_1 \dots x_r)$, $(x_{r+1} \dots x_{2r})$, etc. Let v'_x be the permutation that takes x_i to x_{i+r} for all i (indices modulo rs); then v'_x lies in N_λ. Taking v' to be the product over the orbits x of the v'_x we obtain an element of $N_{\lambda,(w')^{\min}}$ such that

$$w' = \widetilde{(w')^{\min}}(v').$$

Now if $w' \sim w$ then there exists a $u \in N_\lambda$ such that $u(w')^{\min}u^{-1} = w$; taking $v = uv'u^{-1}$ we find that $uw'u^{-1} = \widetilde{w}(v)$. □

Corollary 6.11. *Suppose w is N_λ-minimal. The number of w' such that $w' \sim w$ is equal to the order of N_λ.*

Proof. The previous lemmas show that the set of such w' is the union of the N_λ-conjugacy classes of $\widetilde{w}(v)$, as v runs over a set of representatives for the conjugacy classes in $N_{\lambda,w}$. For each such v the size of its N_λ-conjugacy class is equal to $\#N_\lambda/\#N_{\lambda,v}$. For each v, the index of $N_{\lambda,w}$ in $N_{\lambda,v}$ is equal to the size of the $N_{\lambda,w}$-conjugacy class C_v of v. Thus the total number of such w' is the sum

$$\#N_\lambda \sum_v \frac{\#C_v}{\#N_{\lambda,w}}$$

which is clearly equal to $\#N_\lambda$. □

We now relate the equivalence $\sim$ to M_λ. Specifically, we observe:

Proposition 6.12. *Suppose that $w \sim w'$. Then λ is trivial on T_q^w if, and only if, λ is trivial on $T_q^{w'}$.*

Proof. It suffices to show this in the case where $w' = w^{\min}$ (for some chosen cycle representation of w), as we can deduce any other case from this one and N_λ-conjugacy.

Let S_λ be the set of N_λ-orbits on $\{1, \dots, n\}$, and $f : \{1, \dots, n\} \to S_\lambda$ the map that sends an element to its N_λ-orbit. There exists a map $g : S_\lambda \to \mathbb{Z}$ such that on the diagonal matrix t with entries $t_1, \dots, t_n$, we have $\lambda(t) = \prod_i t_i^{g(f(i))}$.

An element of T_q^w is a diagonal matrix whose entries t_i satisfy $t_{w(i)} = t_i^q$ for all i. In particular, for each i, t_i is a $(q^{d_i}-1)$-st root of unity, where d_i is the size of the w-orbit of i. In particular, λ is trivial on T_q^w if, and only if, for all i the sum

$$\Sigma_i = \sum_{j=0}^{d_i-1} q^j g(f(w^j(i)))$$

is divisible by $q^{d_i} - 1$.

In $w^{\min}$ the w-orbit of i breaks up as a union of N_λ-conjugate orbits, each of size r. In particular for each j, the elements $w^j(i)$ and $w^{j+r}(i)$ lie in the same N_λ-orbit, so $g(w^j(i)) = g(w^{j+r}(i))$. This means that the sum Σ_i can be rewritten as

$$\Sigma_i = (1 + q^r + \cdots + q^{d_i - r}) \sum_{j=0}^{r-1} q^j g(f(w^j(i))).$$

In particular Σ_i is divisible by $q^{d_i} - 1$ if, and only if, the sum

$$\sum_{j=0}^{r-1} q^j g(f(w^j(i)))$$

is divisible by $q^r - 1$. But this is precisely the condition for λ to be trivial on $w^{\min}$. □

From this it follows that the quotient $M_\lambda/\#N_\lambda$ counts the number of N_λ-minimal orbits of w in S_n such that λ is trivial on T_q^w. In particular this quotient is an integer. This completes the proof of Lemma 6.5 and hence of Theorem 6.2.

7. Deformation theory

In this section we examine the local deformation theory of a representation $\bar\rho : G_F \to \mathrm{GL}_n(k)$. As in previous sections, let $I_F^{(\ell)}$ denote the prime to ℓ part of the inertia group of F, and fix a topological generator $\tilde\sigma$ of $I_F/I_F^{(\ell)}$ and a Frobenius element $\widetilde{\mathrm{Fr}}$ in $W_F/I_F^{(\ell)}$.

We first recall some results of Clozel, Harris and Taylor:

Proposition 7.1 [Clozel et al. 2008, Lemmas 2.4.11–2.4.13]. *Let $\bar\tau$ be an irreducible representation of $I_F^{(\ell)}$ over k, and let $G_{\bar\tau}$ be the subgroup of G_F that preserves $\bar\tau$ under conjugation. Then*

(1) *$\bar\tau$ lifts uniquely to a representation τ of $I_F^{(\ell)}$ over $W(k)$,*

(2) *τ extends uniquely to a representation of $I_F \cap G_{\bar\tau}$ of determinant prime to ℓ,*

(3) *τ extends (nonuniquely) to a representation of $G_{\bar\tau}$.*

If we fix a representation τ of $G_{\bar\tau}$ as in part (3), *we obtain an action of $G_{\bar\tau}/I_F^{(\ell)}$ on $\mathrm{Hom}_{I_F^{(\ell)}}(\tau, \rho)$ for any G_F-module ρ. Moreover, we have a direct sum decomposition of G_F-modules,*

$$\rho \cong \bigoplus_{[\bar\tau]} \mathrm{Ind}_{G_{\bar\tau}}^{G_F}[\mathrm{Hom}_{I_F^{(\ell)}}(\tau, \rho) \otimes \tau],$$

where $\bar\tau$ runs over G_F-conjugacy classes of irreducible representations of $I_F^{(\ell)}$ over k.

Fix, for each G_F-conjugacy class of $\bar\tau$, a τ as in the proposition. Suppose we are given a representation $\rho_A : G_F \to \mathrm{GL}_n(A)$. We then obtain a direct sum decomposition

$$\rho_A = \bigoplus_{[\bar\tau]} \mathrm{Ind}_{G_{\bar\tau}}^{G_F}[\mathrm{Hom}_{I_F^{(\ell)}}(\tau, \rho_A) \otimes \tau].$$

It is clear that $\mathrm{Hom}_{I_F^{(\ell)}}(\tau, \rho_A)$ is a free A-module for all τ, and that the collection of $G_{\bar\tau}$-representations $\mathrm{Hom}_{I_F^{(\ell)}}(\tau, \rho)_A$ determines the representation ρ_A up to isomorphism.

Definition 7.2. A *pseudoframing* of a continuous representation $\rho_A : G_F \to \mathrm{GL}_n(A)$ is a choice, for each $\bar\tau$, of basis for each $\mathrm{Hom}_{I_F^{(\ell)}}(\tau, \rho_A)$. A *pseudoframed deformation* of a continuous representation $\bar\rho : G_F \to \mathrm{GL}_n(k)$ (together with a chosen pseudoframing) is a lift $\rho_A : G_F \to \mathrm{GL}_n(A)$ of $\bar\rho$, together with a pseudoframing of ρ_A that lifts the chosen pseudoframing of ρ.

Fix a $\bar\rho$ and a pseudoframing of $\bar\rho$, and, for each $\bar\tau$, let $\bar\rho_{\bar\tau}$ be the $G_{\bar\tau}$-representation $\mathrm{Hom}_{I_F^{(\ell)}}(\tau, \bar\rho)$. Let $R^\diamond_{\bar\rho}$ be the completed tensor product

$$\widehat{\bigotimes}_{[\bar\tau]} R^\square_{\bar\rho_{\bar\tau}}$$

of the universal framed deformation rings of the $\bar\rho_{\bar\tau}$. Over each such ring we have the universal framed deformation $\rho^\square_{\bar\tau}$ of $\bar\rho_{\bar\tau}$.

Using these, we construct a representation

$$\rho^\diamond := \bigoplus_{[\bar\tau]} \mathrm{Ind}_{G_{\bar\tau}}^{G_F}[\rho^\square_{\bar\tau} \otimes \tau]$$

that has a natural pseudoframing induced by the universal framings of the representations $\rho^\square_{\bar\tau}$. One easily verifies that the pair $R^\diamond_\rho$, $\rho^\diamond$ is a universal object for pseudoframed deformations of ρ.

For each $\bar\tau$, the formal group $\mathcal{G}^\square_{\bar\rho_{\bar\tau}}$ acts on $\mathrm{Spf}\, R^\square_{\rho_{\bar\tau}}$ by "change of frame". Let $\mathcal{G}^\diamond_{\bar\rho}$ be the product of the $\mathcal{G}^\square_{\bar\rho_{\bar\tau}}$. Then $\mathcal{G}^\diamond_{\bar\rho}$ acts on $\mathrm{Spf}\, R^\diamond_{\bar\rho}$ by "change of pseudoframing".

For computational purposes it is often easier to work with $R^\diamond_{\bar\rho}$ rather than $R^\square_{\bar\rho}$, as $R^\diamond_{\bar\rho}$ can be made quite explicit. The two rings are related in a natural way: one has a ring $R^{\square,\diamond}_{\bar\rho}$ that is universal for triples consisting of a deformation ρ of $\bar\rho$, a framing of ρ lifting that of $\bar\rho$, and a pseudoframing of ρ lifting that of $\bar\rho$. Then $\mathrm{Spf}\, R^{\square,\diamond}_{\bar\rho}$ is a (split) $\mathcal{G}^\diamond_{\bar\rho}$-torsor over $\mathrm{Spf}\, R^\square_{\bar\rho}$ and a (split) $\mathcal{G}^\square_{\bar\rho}$-torsor over $\mathrm{Spf}\, R^\diamond_{\bar\rho}$.

We immediately deduce:

Corollary 7.3. *The ring $R^\square_{\bar\rho}$ is a reduced, ℓ-torsion free local complete intersection.*

Proof. The construction above shows that it suffices to prove the same claim with $R^\square_{\bar\rho}$ replaced by $R^\diamond_{\bar\rho}$. But the latter is a completed tensor product of rings of the form $R^\square_{\bar\rho_{\bar\tau}}$, and each of these is isomorphic to the completion of a ring of the form $R_{q,n}$ (with q and n depending on $\bar\tau$) at a maximal ideal. The result thus follows from the results of Section 4. □

Moreover, we may canonically identify both the $\mathcal{G}^\square_{\bar\rho}$-invariant elements of $R^\square_{\bar\rho}$ and the $\mathcal{G}^\diamond_{\bar\rho}$-invariant elements of $R^\diamond_{\bar\rho}$ with the $\mathcal{G}^\square_{\bar\rho} \times \mathcal{G}^\diamond_{\bar\rho}$-invariant elements of $R^{\square,\diamond}_{\bar\rho}$. In particular these spaces of invariants are naturally isomorphic.

Given a choice of framing of $\rho^\diamond$, we get a map $R^\square_{\bar\rho} \to R^\diamond_{\bar\rho}$. When restricted to $\mathcal{G}^\square_{\bar\rho}$-invariants this map is the isomorphism of $(R^\square_{\bar\rho})^{\mathcal{G}^\square_{\bar\rho}}$ with $(R^\diamond_{\bar\rho})^{\mathcal{G}^\diamond_{\bar\rho}}$ constructed above. Summarizing, we have:

Lemma 7.4. *For any choice of framing of $\rho^\diamond$, the induced map: $R^\square_{\bar\rho} \to R^\diamond_{\bar\rho}$ identifies the $\mathcal{G}^\square_{\bar\rho}$-invariant elements of $R^\square_{\bar\rho}$ with the $\mathcal{G}^\diamond_{\bar\rho}$-invariant elements of $R^\diamond_{\bar\rho}$. (In particular the image of this set of invariant elements is saturated in $R^\diamond_{\bar\rho}$.)*

8. The rings R_ν

Let $\bar\rho : W_F/I_F^{(\ell)} \to \mathrm{GL}_n(k)$ be a representation. Then we have a corresponding map $x : R_{q,n} \to k$, with kernel $\mathfrak{m}$. It follows easily from the universal property of the pair $(R_{q,n}, \rho_{F,n})$ that the completion $(R_{q,n})_\mathfrak{m}$ is isomorphic to $R^\diamond_{\bar\rho}$, and that this isomorphism is induced by the base change of $\rho_{F,n}$ to $(R_{q,n})_\mathfrak{m}$. In other words, $R_{q,n}$ is a global object that interpolates the formal deformation rings $R^\diamond_{\bar\rho}$ for $\bar\rho$ trivial on $I_F^{(\ell)}$.

We would like to construct similar objects for $\bar\rho$ whose restriction to $I_F^{(\ell)}$ is nontrivial. Let us define:

Definition 8.1. An ℓ-inertial type is a representation ν of $I_F^{(\ell)}$ over k that extends to a representation of W_F.

Note that (as $I_F^{(\ell)}$ is a profinite group of pro-order prime to ℓ), such a representation lifts uniquely to a representation of $I_F^{(\ell)}$ over $W(k)$, and this lift also extends to a representation of W_F. We will thus consider an ℓ-inertial type ν as a representation over $W(k)$ rather than over k whenever it is convenient to do so.

Now fix an ℓ-inertial type ν, and for each irreducible representation $\bar\tau$ of $I_F^{(\ell)}$ over k, let $n_{\bar\tau}$ be the multiplicity of $\bar\tau$ in ν (note that $n_{\bar\tau}$ depends only on the W_F-conjugacy class of $\bar\tau$.) Let $W_{\bar\tau}$ be the subgroup of W_F that fixes $\bar\tau$ under conjugation, let $F_{\bar\tau}$ be the fixed field of $W_{\bar\tau}$, and let $q_{\bar\tau}$ denote the cardinality of the residue field of $F_{\bar\tau}$.

We define R_ν to be the tensor product,

$$R_\nu := \bigotimes_{\bar\tau} R_{q_{\bar\tau}, n_{\bar\tau}},$$

where $\bar\tau$ runs over a set of representatives for the W_F-conjugacy classes of irreducible representations appearing in ν. For each $\bar\tau$ we have a representation $\rho_{F_{\bar\tau}, n_{\bar\tau}}$ over $R_{q_{\bar\tau}, n_{\bar\tau}}$, which we regard as a representation over R_ν in the obvious way.

Define the representation $\rho_\nu : W_F \to \mathrm{GL}_n(R_\nu)$ as follows:

$$\rho_\nu := \bigoplus_{\bar\tau} \mathrm{Ind}_{W_{\bar\tau}}^{W_F} \rho_{F_{\bar\tau}, n_{\bar\tau}} \otimes \tau,$$

where $\bar\tau$ runs over a set of representative for the W_F-conjugacy classes of irreducible representations appearing in ν, and for each such $\bar\tau$, we have chosen an extension τ of $\bar\tau$ to a representation $W_F \to \mathrm{GL}_n(W(k))$ as in Proposition 7.1. Note that ρ_ν inherits a pseudoframing from the natural framings of the $\rho_{F_{\bar\tau}, n_{\bar\tau}}$, and that the restriction of ρ_ν to $I_F^{(\ell)}$ is given by ν.

For a map $x : R_\nu \to k$, the specialization $(\rho_\nu)_x$ is a pseudoframed representation $W_F \to \mathrm{GL}_n(k)$, whose restriction to $I_F^{(\ell)}$ is given by ν. This defines a bijection between k-points of $\mathrm{Spec}\, R_\nu$ and such pseudoframed representations. Moreover, it follows directly from the constructions of R_ν and $R^\diamond_{(\rho_\nu)_x}$ that the completion of R_ν at the maximal ideal corresponding to x is naturally isomorphic to $R^\diamond_{(\rho_\nu)_x}$, in a manner compatible with the universal family on the latter.

Moreover, the universal property for each $R_{q_{\bar\tau},n_{\bar\tau}}$ immediately yields:

Proposition 8.2. *For any finitely generated, ℓ-adically separated $W(k)$-algebra A, and any pseudoframed, ℓ-adically continuous representation $\rho : W_F \to \mathrm{GL}_n(A)$ whose restriction to $I_F^{(\ell)}$ is isomorphic to ν, there is a unique map: $R_\nu \to A$ such that ρ is the base change of ρ_ν.*

For each $\bar\tau$, the group $\mathrm{GL}_{n_{\bar\tau}}$ acts on $R_{q_{\bar\tau},n_{\bar\tau}}$. Let $\mathcal{G}_\nu$ be the product of the $\mathrm{GL}_{n_{\bar\tau}}$; then $\mathcal{G}_\nu$ acts on $\mathrm{Spec}\, R_\nu$ by "changing the pseudoframe".

9. Maps from $Z_{[L,\pi]}$ to R_ν

Now fix a pair (L,π), where L is a Levi subgroup of $\mathrm{GL}_n(F)$ and π is an irreducible supercuspidal k-representation of L. The mod ℓ semisimple local Langlands correspondence of Vignéras [2001] attaches to π a semisimple k-representation ρ of W_F. Let $\bar\nu$ be the restriction of ρ to $I_F^{(\ell)}$. Then $\bar\nu$ lifts uniquely to a $W(k)$-representation ν of $I_F^{(\ell)}$, and we have:

Proposition 9.1. *The irreducible $\overline{\mathcal{K}}$-representations of $\mathrm{GL}_n(F)$ that are objects of $\mathrm{Rep}_{W(k)}(\mathrm{GL}_n(F))_{[L,\pi]}$ correspond, via local Langlands, to the $\overline{\mathcal{K}}$-representations of W_F whose restriction to $I_F^{(\ell)}$ is isomorphic to ν.*

Proof. This is an easy consequence of the compatibility of Vigneras' mod ℓ correspondence with reduction mod ℓ. □

This proposition shows that for any $\overline{\mathcal{K}}$-point x of $\mathrm{Spec}\, R_\nu$, the representation ρ_x corresponds, via local Langlands (and Frobenius semisimplification if necessary) to an irreducible $\overline{\mathcal{K}}$-representation Π_x in $\mathrm{Rep}_{W(k)}(\mathrm{GL}_n(F))_{[L,\pi]}$, and hence to a $\overline{\mathcal{K}}$-point of $\mathrm{Spec}\, Z_{[L,\pi]}$. It is a natural question to ask whether this map is induced by a map $Z_{[L,\pi]} \to R_\nu$. Indeed, we conjecture:

Conjecture 9.2 (weak local Langlands in families). *There is a map $Z_{[L,\pi]} \to R_\nu$ such that the induced map on $\overline{\mathcal{K}}$-points takes a point x of $\mathrm{Spec}\, R_\nu$ to the $\overline{\mathcal{K}}$-point of $Z_{[L,\pi]}$ that gives the action of $Z_{[L,\pi]}$ on the representation Π_x corresponding to ρ_x by local Langlands.* (*We will say such a map is* ***compatible with local Langlands***.)

Since R_ν is reduced and ℓ-torsion free, such a map is unique if it exists. Note also that the image of any element of $Z_{[L,\pi]}$ under such a map is invariant under the action of $\mathcal{G}_\nu$, and so any such map must factor through the subalgebra R_ν^{inv} of $\mathcal{G}_\nu$-invariant elements of R_ν. We further conjecture:

Conjecture 9.3 (strong local Langlands in families). *There is an isomorphism $Z_{[L,\pi]} \cong R_\nu^{\mathrm{inv}}$ such that the composition*

$$Z_{[L,\pi]} \to R_\nu^{\mathrm{inv}} \to R_\nu$$

is compatible with local Langlands.

If one completes at a maximal ideal of R_ν, corresponding to a representation $\bar\rho$ of W_F over k, and uses Lemma 7.4 to relate the invariant elements of $R_{\bar\rho}^{\square}$ and $R_{\bar\rho}^{\diamond}$, one recovers Conjectures 7.5 and 7.6 of [Helm 2016b]. In particular (see Theorem 7.9 of [Helm 2016b]), Conjecture 9.2 above implies the "local Langlands in families" conjecture [Emerton and Helm 2014, Conjecture 1.1.3].

These conjectures should be viewed as relating "congruences" between admissible representations (which are in some sense encoded in the structure of $Z_{[L,\pi]}$) with "congruences" between representations of W_F (encoded in R_ν). Since inverting ℓ destroys information about such congruences, one expects such conjectures to be relatively straightforward with ℓ inverted. We will show that this is indeed the case.

First, note that any map

$$Z_{[L,\pi]} \otimes \overline{\mathcal{K}} \to R_\nu \otimes \overline{\mathcal{K}}$$

that is compatible with local Langlands is Galois equivariant, and hence descends to a map

$$Z_{[L,\pi]}\left[\frac{1}{\ell}\right] \to R_\nu\left[\frac{1}{\ell}\right]$$

compatible with local Langlands. It thus suffices to show:

Theorem 9.4. *There is a map $Z_{[L,\pi]} \otimes \overline{\mathcal{K}} \to R_\nu \otimes \overline{\mathcal{K}}$ compatible with local Langlands* (*and therefore a corresponding map over $\mathcal{K}$*). *Moreover, the image of this map is $R_\nu^{\mathrm{inv}} \otimes \overline{\mathcal{K}}$.*

To prove this, we first work on the level of connected components. We have an isomorphism

$$Z_{[L,\pi]} \otimes \overline{\mathcal{K}} \cong \prod_{M,\tilde{\pi}} \widetilde{Z}_{(M,\tilde{\pi})},$$

by Theorem 3.5, where $(M, \tilde{\pi})$ varies over the inertial equivalence classes of pairs that reduce modulo ℓ to (L, π). Thus the connected components of $\operatorname{Spec} Z_{[L,\pi]} \otimes \overline{\mathcal{K}}$ are in bijection with the pairs $(M, \tilde{\pi})$. Via local Langlands, these correspond to representations of I_F. More precisely, let Π be an admissible representation of G, let $\rho : W_F \to \mathrm{GL}_n(\overline{\mathcal{K}})$ correspond to Π via local Langlands, and let $\tilde{\rho} : W_F \to \mathrm{GL}_n(\overline{\mathcal{K}})$ be the representation of W_F corresponding to $\tilde{\pi}$ via local Langlands. Then Π belongs to the block corresponding to $(M, \tilde{\pi})$ if and only if the restriction of ρ^{ss} to I_F coincides with the restriction of $\tilde{\rho}$ to I_F.

On the other hand, it is an easy consequence of Proposition 4.11 that as x varies over $\overline{\mathcal{K}}$-points of $\operatorname{Spec} R_\nu$, the restriction of $\rho_{\nu,x}^{\mathrm{ss}}$ to I_F is constant on connected components of $\operatorname{Spec} R_\nu \otimes \overline{\mathcal{K}}$. We can thus let $R_\nu^{\tilde{\rho}}$ be the direct factor of $R_\nu \otimes \overline{\mathcal{K}}$ corresponding to the union of the connected components of $\operatorname{Spec} R_\nu \otimes \overline{\mathcal{K}}$ on which the restriction of $\rho_{\nu,x}^{\mathrm{ss}}$ to I_F is isomorphic to the restriction of $\tilde{\rho}$ to I_F. We will see later that $\operatorname{Spec} R_\nu^{\tilde{\rho}}$ is in fact connected.

It then suffices to construct, for each $(M, \tilde{\pi})$, an isomorphism

$$\widetilde{Z}_{(M,\tilde{\pi})} \to (R_\nu^{\tilde{\rho}})^{\mathrm{inv}}$$

compatible with local Langlands. Since $(M, \tilde{\pi})$ is only well-defined up to inertial equivalence, we may assume that $\tilde{\pi}$ has the form

$$\tilde{\pi} \cong \bigotimes_i \tilde{\pi}_i^{\otimes r_i},$$

where the $\tilde{\pi}_i$ are pairwise inertially inequivalent representations of $\mathrm{GL}_{n_i}(F)$. Unwinding the Bernstein–Deligne description of $\widetilde{Z}_{(M,\tilde{\pi})}$, we obtain an isomorphism

$$\widetilde{Z}_{(M,\tilde{\pi})} \cong \bigotimes_i \overline{\mathcal{K}}[X_{i,1}^{\pm 1}, \ldots, X_{i,r_i}^{\pm 1}]^{S_{r_i}},$$

where the symmetric group S_{r_i} acts by permuting the elements $X_{i,1}, \ldots, X_{i,r_i}$.

For each i, and any $\alpha \in \overline{\mathcal{K}}$, let $\chi_{i,\alpha}$ denote the unramified character of $\mathrm{GL}_{n_i}(F)$ that takes the value α on any element of $\mathrm{GL}_{n_i}(F)$ with determinant ϖ_F. An irreducible Π in $\mathrm{Rep}_{\overline{\mathcal{K}}}(M, \tilde{\pi})$ has supercuspidal support $(M, \tilde{\pi}')$ for some $\tilde{\pi}'$ of the form

$$\tilde{\pi}' \cong \bigotimes_i \bigotimes_{j=1}^{r_i} \tilde{\pi}_i \otimes \chi_{i,\alpha_{i,j}}$$

for suitable $\alpha_{i,j}$. Then the d-th elementary symmetric function in $X_{i,1}, \ldots, X_{i,r_i}$, considered as an element of $\widetilde{Z}_{(M,\tilde{\pi})}$, acts on Π via the d-th elementary symmetric function in the $\alpha_{i,1}^{f_i'}, \ldots, \alpha_{i,r_i}^{f_i'}$, where f_i' is the order of the group of unramified characters χ such that $\tilde{\pi}_i \otimes \chi$ is isomorphic to $\tilde{\pi}_i$.

For each i, the irreducible representation $\tilde{\rho}_i$ of W_F corresponding to $\tilde{\pi}_i$ via local Langlands decomposes, when restricted to I_F, as a direct sum of distinct irreducible representations of I_F, all of which are W_F-conjugate. Fix an irreducible representation $\tilde{\tau}_i$ of I_F contained in $\tilde{\rho}_i$, and let W_i be the normalizer of $\tilde{\tau}_i$ in W_F. Then there is a unique way of extending $\tilde{\tau}_i$ to a representation of W_i such that the induction of the resulting extension to W_F is isomorphic to $\tilde{\rho}_i$. (Note that this implies that W_i has index f_i' in W_F.)

This choice of extension of $\tilde{\tau}_i$ to W_i gives rise to an action of W_i on the space $\mathrm{Hom}_{I_F}(\tilde{\tau}_i, \rho_v)$. The quotient of this space that lives over $R_v^{\tilde{\rho}}$ is a free $R_v^{\tilde{\rho}}$-module of rank r_i, with an unramified action of W_i.

Let $\widetilde{\mathrm{Fr}}_i$ be a Frobenius element of W_i, and let $P_i(x) = \sum_{j=0}^{r_i} a_{i,j} X^j$ be the characteristic polynomial of $\widetilde{\mathrm{Fr}}_i$ on $\mathrm{Hom}_{I_F}(\tilde{\tau}_i, \rho_v)$ (over $R_v^{\tilde{\rho}}$). Consider the map $\widetilde{Z}_{(M,\tilde{\pi})} \to R_v^{\tilde{\rho}}$ that sends the d-th elementary symmetric function in $X_{i,1}, \ldots, X_{i,r_i}$ to the element $(-1)^d a_{i,r_i-d}$ of $R_v^{\tilde{\rho}}$. One verifies easily that this map is compatible with local Langlands.

It remains to show that $(R_v^{\tilde{\rho}})^{\mathrm{inv}}$ is generated by the images of these elements. Given a polynomial P_i of degree r_i, with coefficients in a ring R, we can associate to it the unramified R-representation $M_i(P_i)$ of W_i on which $\widetilde{\mathrm{Fr}}_i$ acts via the companion matrix of P_i. The representation $\rho(\{P_i\})$ given by

$$\rho(\{P_i\}) = \bigoplus_i \mathrm{Ind}_{W_i}^{W_F} M_i(P_i) \otimes \tilde{\tau}_i$$

is then an R-point of $\mathrm{Spec}\, R_v^{\tilde{\rho}}$. In this way we obtain a natural map

$$R_v^{\tilde{\rho}} \to \bigotimes_i \overline{\mathcal{K}}[Y_{i,1}, \ldots, Y_{i,r_i}]$$

that in particular takes the element $(-1)^d a_{i,r_i-d}$ of $R_v^{\tilde{\rho}}$ to $Y_{i,d}$. On the other hand, it is easy to see that for every y in $(\mathrm{Spec}\, R_v^{\tilde{\rho}})(\overline{\mathcal{K}})$, there is a point x in $(\mathrm{Spec}\, R_v^{\tilde{\rho}})(\overline{\mathcal{K}})$ arising from a collection of polynomials $\{P_i(x)\}$ such that y is in the closure of the G_v-orbit of x. It follows that the map

$$R_v^{\tilde{\rho}} \to \bigotimes_i \overline{\mathcal{K}}[Y_{i,1}, \ldots, Y_{i,r_i}]$$

is injective on $(R_v^{\tilde{\rho}})^{\mathrm{inv}}$. Therefore $((R_v)^{\tilde{\rho}})^{\mathrm{inv}}$ is generated by the elements a_{i,r_i-d}, completing the proof.

It is not hard to go slightly further, and show:

Theorem 9.5. *The image of $Z_{[L,\pi]}$ in $R_\nu\left[\frac{1}{\ell}\right]$ under the map of Theorem 9.4 lies in the normalization of R_ν.*

Proof. Fix an element x of $Z_{[L,\pi]}$, and let y be its image in $R_\nu\left[\frac{1}{\ell}\right]$. Let A be a discrete valuation ring that is a $W(k)$-algebra, with field of fractions K of characteristic zero, and fix a map $R_\nu \to A$. This corresponds to a pseudoframed representation ρ_A of W_F. Let Π_K denote the admissible K-representation corresponding to $\rho_A \otimes_A K$ via local Langlands. Since $\rho_A \otimes_A K$ admits an A-lattice, so does Π_K. In particular the action of x on Π_K is via an element of A, so y maps to an element of A under the map $R_\nu\left[\frac{1}{\ell}\right] \to K$. Since this is true for every A and every map $R_\nu \to A$, y lives in the normalization of R_ν as claimed. □

10. Main results

The main objective of this section (and, indeed, the paper) is to show the following:

Theorem 10.1. *Suppose that Conjecture 9.2 holds for all $\mathrm{GL}_m(F)$, $m \le n$, and Conjecture 9.3 holds for $m < n$. Then*

(1) *the map $\bar{E}_{q,n}\left[\frac{1}{\ell}\right] \to B_{q,n}\left[\frac{1}{\ell}\right]$ of Section 6 induces an isomorphism of $\bar{E}_{q,n}$ with $B_{q,n}$, and*

(2) *Conjecture 9.3 holds for $\mathrm{GL}_n(F)$.*

We begin by proving the first claim, using the weak conjecture for GL_n in depth zero. Let Z_n^0 be the product of the depth zero blocks of $\mathrm{Rep}_{W(K)}(G)$. The weak conjecture then gives rise to a map $Z_n^0 \to S_{q,n}$ compatible with the local Langlands correspondence. The subalgebra of Z_n^0 consisting of elements that are constant on inertial equivalence classes is isomorphic to $\bar{E}_{q,n}$, by Proposition 3.10. By compatibility with local Langlands together with Propositions 4.11 and 5.3 the image of $\bar{E}_{q,n}$ in $S_{q,n}$ is contained in $B_{q,n}$, and the induced map $\bar{E}_{q,n}\left[\frac{1}{\ell}\right] \to B_{q,n}\left[\frac{1}{\ell}\right]$ is the map considered in Section 6. It thus follows from Corollary 6.3 that the map $\bar{E}_{q,n} \to B_{q,n}$ is an isomorphism.

We now turn to the second claim. Fix a mod ℓ supercuspidal inertial equivalence class $[L, \pi]$, corresponding to an ℓ-inertial type ν, and note that we have tensor factorizations

$$Z_{[L,\pi]} \cong \bigotimes_i Z_{[L_i,\pi_i]}, \qquad R_\nu \cong \bigotimes_{\bar{\tau}} R_{q_{\bar{\tau}},n_{\bar{\tau}}},$$

where the $[L_i, \pi_i]$ are simple blocks. The former factorization is compatible with parabolic induction and the latter arises from the direct sum decomposition

$$\rho_\nu = \bigoplus_{\bar{\tau}} \mathrm{Ind}_{W_{\bar{\tau}}}^{W_F} \rho_{F_{\bar{\tau}},n_{\bar{\tau}}} \otimes \tau.$$

Since simple blocks correspond to types ν with only one $n_{\bar{\tau}}$ nonzero, these factorizations are compatible, in the sense that if we have maps $Z_{[L_i,\pi_i]} \to R_{\nu_i}$ for each i that are compatible with local Langlands, then their tensor product gives a map $Z_{[L,\pi]} \to R_\nu$ compatible with local Langlands. Thus both Conjecture 9.2 and Conjecture 9.3 reduce to the corresponding conjectures on simple blocks. We thus henceforth assume that $[L, \pi]$ is of the form $[L_n, \pi_n]$ with $\pi_n \cong \pi_1^{\otimes n}$ for a supercuspidal representation π_1. Following Section 3 we set $Z_n = Z_{[L_n,\pi_n]}$. The corresponding R_{ν_n} is then isomorphic to $R_{q_{\bar{\tau}},n}$ for some fixed $\bar{\tau}$.

We first consider the case in which n is not $q_{\bar{\tau}}$-relevant. Let ν be the maximal $q_{\bar{\tau}}$-relevant partition of n. We have a commutative diagram

$$\begin{array}{ccc} Z_n & \longrightarrow & R^{\mathrm{inv}}_{q_{\bar{\tau}},n} \\ \downarrow & & \downarrow \\ \bigotimes_i Z_{\nu_i} & \longrightarrow & \bigotimes_i R^{\mathrm{inv}}_{q_{\bar{\tau}},\nu_i} \end{array}$$

in which the horizontal maps are those arising from the weak conjecture, the left-hand vertical map is Ind_ν, and the right-hand vertical map is induced by the map $\mathrm{Spec} \otimes_i R_{q_{\bar{\tau}},\nu_i} \to R_{q_{\bar{\tau}},n}$ that takes a collection $(\mathrm{Fr}_i, \sigma_i)$ of matrices with $\mathrm{Fr}_i\, \sigma_i\, \mathrm{Fr}_i^{-1} = \sigma_i^{q_{\bar{\tau}}}$ to the pair $\left(\bigoplus_i \mathrm{Fr}_i, \bigoplus_i \sigma_i\right)$.

The horizontal maps are isomorphisms after inverting ℓ, and our hypotheses imply that the lower horizontal map is an isomorphism integrally. Moreover the left-hand vertical map is injective with saturated image by Theorem 3.9 and the discussion in the paragraph following it. It follows immediately that the top horizontal map must also be an isomorphism.

We now assume that n is $q_{\bar{\tau}}$-relevant (that is, it lies in $\{1, e_{q_{\bar{\tau}}}, \ell e_{q_{\bar{\tau}}}, \dots\}$). Let m be the largest element of this set that is strictly less than n. Set $j = \frac{n}{m}$.

We have a subalgebra $\bar{E}_{q^{f'},n,1}$ of Z_m and compatibility with local Langlands shows that $q^{f'} = q_{\bar{\tau}}$. Thus the map $Z_n \to R_{q_{\bar{\tau}},n}$ induces a map $\bar{E}_{q_{\bar{\tau}},n,1} \to R_{q_{\bar{\tau}},n}$. Reasoning as in the depth zero setting we see that the image of this map is contained in $B_{q_{\bar{\tau}},n,1}$. It seems likely that the resulting map $\bar{E}_{q_{\bar{\tau}},n,1} \to B_{q_{\bar{\tau}},n,1}$ is the one considered in Section 6, but we do not prove this here. Instead we use the fact that we have shown these two rings to be abstractly isomorphic, together with the following lemma:

Lemma 10.2. *Let $\bar{E}$ be a finite rank, reduced, ℓ-torsion free $W(k)$-algebra, and let $f : \bar{E} \to \bar{E}$ be an injection. Then f is an isomorphism.*

Proof. Clearly f is an isomorphism after inverting ℓ. On the other hand, the hypotheses guarantee that $\bar{E}\left[\frac{1}{\ell}\right]$ is a product of finite extensions of $\mathcal{K}$, and f is a $\mathcal{K}$-linear automorphism of this product. In particular there is some power of f that is the identity. □

We thus conclude that the map $Z_n \to R_{q_{\bar{\tau}},n}$ coming from the weak conjecture induces an isomorphism of $\bar{E}_{q_{\bar{\tau}},n,1}$ with $B_{q_{\bar{\tau}},n,1}$.

Now consider the commutative diagram

$$\begin{array}{ccc} K & \longrightarrow & K' \\ \downarrow & & \downarrow \\ Z_n & \longrightarrow & R^{\mathrm{inv}}_{q_{\bar{\tau}},n} \\ \downarrow & & \downarrow \\ Z_m^{\otimes j} & \longrightarrow & (R^{\mathrm{inv}}_{q_{\bar{\tau}},m})^{\otimes j} \end{array}$$

in which the horizontal maps are induced by the weak conjecture, the lower left vertical map is $\mathrm{Ind}_{m,n}$, the lower right vertical map is the one taking a collection of pairs $(\mathrm{Fr}_i, \sigma_i)$ to their direct sum, and K

and K' are the kernels of the lower left and lower right vertical maps, respectively. As in the previous case, all horizontal maps become isomorphisms after inverting ℓ and the bottom horizontal map is an isomorphism integrally.

By Proposition 3.13 K is contained in the subalgebra $\bar{E}_{q_{\bar{\tau}},n,1}[\Theta_{n,n}^{\pm 1}]$ of Z_n, and the image of this subalgebra in $R_{q_{\bar{\tau}},n}$ is saturated. It follows that the map from K to K' is an isomorphism: if x is an element of K', then for some a, the product $\ell^a x$ is in the image of K. But then $\ell^a x$ is in the image of $\bar{E}_{q_{\bar{\tau}},n,1}[\Theta_{n,n}^{\pm 1}]$, so x is as well. On the other hand, the image of K in $\bar{E}_{q_{\bar{\tau}},n,1}$ is saturated (as K is the kernel of a map of rings that have no ℓ-torsion), so x must lie in the image of K.

Let r be an element of $R^{\mathrm{inv}}_{q_{\bar{\tau}},n}$, and let r' be its image in $(R^{\mathrm{inv}}_{q_{\bar{\tau}},m})^{\otimes j}$. There is then an element y of $Z_m^{\otimes j}$ whose image under the bottom horizontal map is r'. Since the map $Z_n \to R^{\mathrm{inv}}_{q_{\bar{\tau}},n}$ is an isomorphism after inverting ℓ, there exists a such that $\ell^a y$ is in the image of $\mathrm{Ind}_{m,n}$.

By Theorem 3.12, there exist $\tilde{y}$ in Z_n and x in $\bar{E}_{q_{\bar{\tau}},n,1}[\Theta_{n,n}^{\pm 1}]$ such that $\mathrm{Ind}_{m,n}(x) = \ell^b(\mathrm{Ind}_{m,n}(\tilde{y}) - y)$. Let s be the image of $\tilde{y}$ in $R^{\mathrm{inv}}_{q_{\bar{\tau}},n}$. The image of $\ell^b(s-r)$ in $(R^{\mathrm{inv}}_{q_{\bar{\tau}},m})^{\otimes j}$ coincides with the image of $\mathrm{Ind}_{m,n}(x)$. Thus $\ell^b(s-r)$ lies in the image of $\bar{E}_{q_{\bar{\tau}},n,1}[\Theta_{n,n}^{\pm 1}]$. Since this image is saturated, the element $s-r$ also lives in this image. Thus the map $Z_n \to R^{\mathrm{inv}}_{q_{\bar{\tau}},n}$ is surjective, so it is an isomorphism.

We have thus completed the proof of Theorem 10.1. In [Helm and Moss 2018] we show that the strong conjecture for GL_{n-1} implies the weak conjecture for GL_n. Together with Theorem 10.1 and the fact that the strong conjecture for GL_1 is an easy consequence of local class field theory, we obtain an unconditional proof both of the strong conjecture, and of the existence of an isomorphism $\bar{E}_{q,n} \cong B_{q,n}$. We refer the reader to the final section of [Helm and Moss 2018] for the details.

Remark 10.3. The isomorphism of $\bar{E}_{q,n}$ with $B_{q,n}$ is an interesting result in finite group theory in its own right. We are aware of no proof other than the one presented here; it is an interesting question to find a purely group-theoretic proof of this result.

11. Affine Curtis homomorphisms

Having established both Conjectures 9.2 and 9.3 we now turn to an interesting consequence of Conjecture 9.2. Fix a w in S_n (which we identify with the Weyl group of $\mathcal{G}$). The conjugacy class of w gives rise to a conjugacy class of nonsplit, unramified tori in $\mathcal{G}$; we let $\mathcal{T}_w$ denote a representative of this conjugacy class. In particular we have $\mathcal{T}_w \cong \prod_{w_i} \mathrm{Res}_{F_i/F} \mathbb{G}_m$, where the product is over the cycles w_i of w and F_i/F is unramified of degree equal to the length of w_i. Let d be the order of w in S_n.

Let X be the character group of $\mathcal{T}_w$, and let $\mathcal{T}_w^L$ denote the algebraic group $\mathrm{Hom}(X', \mathbb{G}_m) \rtimes \mathbb{Z}/d\mathbb{Z}$ (regarded as an algebraic group over $W(k)$), where the action of $1 \in \mathbb{Z}/d\mathbb{Z}$ on X' is via w^{-1}. Then $\mathcal{T}_w^L$ is the L-group of $\mathcal{T}_w$. Moreover, if we identify GL_n (over $W(k)$) with the L-group of $\mathcal{G}$ in such a way that X' becomes identified with the character group of the diagonal torus in GL_n, then we have a natural L-homomorphism from $\mathcal{T}_w^L$ to GL_n that takes $\mathrm{Hom}(X', \mathbb{G}_m)$ to the diagonal torus and takes $1 \in \mathbb{Z}/n\mathbb{Z}$ to w^{-1}. This allows us to transfer a Langlands parameter $\rho_w : W_F \to \mathcal{T}_w^L(\bar{\mathcal{K}})$ for $\mathcal{T}_w$ to a Langlands parameter $\rho : W_F \to \mathrm{GL}_n(\bar{\mathcal{K}})$ for $\mathcal{G}$.

It will be useful to understand the interaction between this transfer and the block decompositions for $\mathrm{Rep}_{W(k)}(T_w)$ and $\mathrm{Rep}_{W(k)}(G)$. Note that

$$T_w = \mathcal{T}_w(F) = \mathrm{Hom}(X, (F^{\mathrm{ur}})^\times)^{\widetilde{\mathrm{Fr}}},$$

where $\widetilde{\mathrm{Fr}}$ is a fixed Frobenius element of W_F, and its action on X is via w. Let $T_w^{(\ell)}$ denote the subgroup $\mathrm{Hom}(X, (\mathcal{O}_{F^{\mathrm{ur}}}^\times)^{(\ell)})^{\widetilde{\mathrm{Fr}}}$ of T_w, where $(\mathcal{O}_{F^{\mathrm{ur}}}^\times)^{(\ell)}$ denotes the elements of pro-order prime to ℓ in $\mathcal{O}_{F^{\mathrm{ur}}}^\times$. Then $T_w^{(\ell)}$ is profinite, of pro-order prime to ℓ, and the quotient $T_w/T_w^{(\ell)}$ is a discrete group. Indeed, explicitly, one has

$$T_w/T_w^{(\ell)} \cong \prod_{w_i} F_i^\times/(\mathcal{O}_{F_i}^\times)^{(\ell)} \cong \prod_{w_i}(\mathbb{Z}\cdot\varpi \times \mathbb{F}_{q^i}^\times),$$

where ϖ is a uniformizer of F (hence also of F_i.)

The blocks of $\mathrm{Rep}_{W(k)}(T_w)$ are thus given by characters $\chi^{(\ell)} : T_w^{(\ell)} \to W(k)^\times$. Choose an extension χ of $\chi^{(\ell)}$ to a character $T_w \to W(k)^\times$. Then "twisting by χ" induces an equivalence of categories between the block of $\mathrm{Rep}_{W(k)}(T_w)$ corresponding to the trivial character of $T_w^{(\ell)}$ and the block corresponding to $\chi^{(\ell)}$. Denote the centers of these blocks by $Z_{w,1}$ and $Z_{w,\chi^{(\ell)}}$, respectively; our choice of χ then gives an isomorphism of $Z_{w,1}$ with $Z_{w,\chi^{(\ell)}}$.

On the other side of the Langlands correspondence, the local Langlands correspondence for tori associates to χ a Langlands parameter $\tilde{\nu}_w : W_F \to \mathcal{T}_w^L(\overline{\mathcal{K}})$; the restriction ν_w of $\tilde{\nu}_w$ to $I_F^{(\ell)}$ depends only on $\chi^{(\ell)}$. Consider the functor that associates to a $W(k)$-algebra R the set of parameters $W_F \to \mathcal{T}_w^L(R)$ whose restriction to $I_F^{(\ell)}$ is equal to ν_w. This functor is easily seen to be representable by a finite type affine scheme $\mathrm{Spec}\, R_\nu^w$, and there is a universal Langlands parameter $\rho_{w,\nu} : W_F \to \mathcal{T}_w^L(R_\nu^w)$. Note that the torus $\mathrm{Hom}(X', \mathbb{G}_m) \subseteq \mathcal{T}_w^L$ acts on $\mathrm{Spec}\, R_\nu^w$ by conjugation; let $(R_\nu^w)^{\mathrm{inv}}$ be the subring of R_ν^w invariant under this action.

We then have the following proposition, which can be seen as an analogue of Conjecture 9.3 for the nonsplit torus T_w:

Proposition 11.1. *There is a unique isomorphism*

$$\mathbb{L}_w : Z_{w,\chi^{(\ell)}} \to (R_\nu^w)^{\mathrm{inv}}$$

which is compatible with the local Langlands correspondence for tori, in the sense that for any Langlands parameter $\rho : W_F \to \mathcal{T}_w^L(\overline{\mathcal{K}})$, corresponding to a character χ_ρ of T_w, and any $z \in Z_{w,\chi^{(\ell)}}$, the value of χ_ρ at z is equal to the value of $\mathbb{L}_w$ at the point of $\mathrm{Spec}(R_\nu^w)$ corresponding to ρ.

Proof. Any parameter $W_F \to \mathcal{T}_w^L(R)$ of type ν_w differs from $\tilde{\nu}_w$ by a parameter $W_F \to \mathcal{T}_w^L(R)$ that is trivial on $I_F^{(\ell)}$. Thus "twisting by ν_w" induces an isomorphism of $\mathrm{Spec}\, R_\nu^w$ with $\mathrm{Spec}\, R_1^w$, where 1 is the trivial character of $I_F^{(\ell)}$. On $\overline{\mathcal{K}}$-points, this isomorphism is compatible with the local Langlands correspondence for tori and the "twisting by χ" isomorphism of $Z_{w,1}$ with $Z_{w,\chi^{(\ell)}}$. We can thus reduce to the case where $\chi^{(\ell)}$ and ν_w are the trivial character.

In this case we can be very explicit: on the one hand, we have isomorphisms

$$Z_{w,1} = W(k)[T_w/T_w^{(\ell)}] \cong W(k)[\mathrm{Hom}(X, (F^{\mathrm{ur}})^\times/(\mathcal{O}_{F^{\mathrm{ur}}}^\times)^{(\ell)})^{\widetilde{\mathrm{Fr}}}],$$

where $\widetilde{\mathrm{Fr}}$ acts on X via w and on $\mathcal{O}_{F^{\mathrm{ur}}}^\times$ in the usual way. Let ϖ be a uniformizer of F corresponding to our Frobenius element $\widetilde{\mathrm{Fr}}$. We then have a canonical isomorphism

$$(F^{\mathrm{ur}})^\times/(\mathcal{O}_{F^{\mathrm{ur}}}^\times)^{(\ell)} \cong \mathbb{Z}\cdot\varpi \times \bar{\mathbb{F}}_q^\times,$$

where $\widetilde{\mathrm{Fr}}$ acts trivially on the first factor and by q-th powers on the second. We thus obtain an isomorphism

$$Z_{w,1} \cong W(k)[\mathrm{Hom}(X, \mathbb{Z})^w] \otimes W(k)[\mathrm{Hom}(X/(qw-1)X, \bar{\mathbb{F}}_q^\times)].$$

On the other side of the Langlands correspondence, fix a generator $\tilde{\sigma}$ of I_F/P_F. Then a Langlands parameter $W_F \to \mathcal{T}_w^L$ trivial on $I_F^{(\ell)}$ is determined by the images of $\widetilde{\mathrm{Fr}}$ and $\tilde{\sigma}$; these form a pair of diagonal matrices $\mathcal{F}$ and σ such that $\mathcal{F}w^{-1}\sigma(\mathcal{F}w^{-1})^{-1} = \sigma^q$. Since $\mathcal{F}$ and σ commute, this condition is equivalent to the condition $\sigma^{w^{-1}} = \sigma^q$. Thus $\mathrm{Spec}\, R_1^w$ decomposes as a product,

$$\mathrm{Spec}\, R_1^w \cong \mathrm{Spec}\, W(k)[X'] \times \mathrm{Spec}\, W(k)[X'/(q-w)X'],$$

where the first factor parametrizes $\mathcal{F}$ and the second parametrizes σ. The conjugation action of $t \in \mathrm{Hom}(X', \mathbb{G}_m)$ on this product fixes the second factor and acts by multiplication by $t^{w^{-1}-1}$ on the first. We thus obtain a product decomposition

$$\mathrm{Spec}(R_1^w)^{\mathrm{inv}} \cong \mathrm{Spec}\, W(k)[X'/(1-w)X'] \times \mathrm{Spec}\, W(k)[X'/(q-w)X'].$$

On the first factor, the isomorphism of $Z_{w,1}$ with $(R_1^w)^{\mathrm{inv}}$ is induced by the isomorphism $\mathrm{Hom}(X, \mathbb{Z})^w \cong X'/(w-1)X'$. On the second factor we have to work a bit harder. Note that $qw-1$ divides q^r-1, where r is a multiple of the order of w. Thus $\mathrm{Hom}(X/(qw-1)X, \bar{\mathbb{F}}_q^\times)$ is isomorphic to $\mathrm{Hom}(X/(qw-1)X, \mathbb{F}_{q^r}^\times)$. Our choice of s gives rise to a system of generators for $\mathbb{F}_{q^r}^\times$ for all r, compatible with respect to norm maps; we can thus identify $\mathrm{Hom}(X/(qw-1)X, \mathbb{F}_{q^r}^\times)$ with the kernel of $qw^{-1}-1$ on $X'/(q^r-1)X'$, via the isomorphism

$$X'/(q^r-1)X' \cong \mathrm{Hom}(X/(q^r-1)X, \mathbb{Z}/(q^r-1)\mathbb{Z}).$$

Finally, multiplication by $1+qw^{-1}+\cdots+q^{r-1}w^{1-r}$ identifies this kernel with $X'/(qw^{-1}-1)X'$. The resulting isomorphism of $\mathrm{Hom}(X/(qw-1)X, \bar{\mathbb{F}}_q^\times)$ with $X'/(qw^{-1}-1)X'$ is independent of r, and gives the desired map from the second factor of $Z_{w,1}$ to the second factor of $(R_1^w)^{\mathrm{inv}}$. One checks easily that the resulting isomorphism is compatible with local Langlands. □

The L-homomorphism of $\mathcal{T}_w^L$ into GL_n takes Langlands parameters for T_w to Langlands parameters for G. If the former has type ν_w, then so does the latter (where we regard ν_w as an ℓ-inertial type by embedding it in $\mathrm{GL}_n(W(k))$ by identifying $\mathrm{Hom}(X', \mathbb{G}_m)$ with the diagonal matrices.) Thus

this L-homomorphism induces a map $R^{\nu} \to R^{w}_{\nu}$ that takes R^{inv}_{ν} to $(R^{w}_{\nu})^{\mathrm{inv}}$. Combining this with Proposition 11.1 and Conjecture 9.2, we obtain a map

$$eZ_n \to Z_{w,\chi^{(\ell)}},$$

where e is the idempotent of Z_n corresponding to the ℓ-inertial type ν. On $\overline{\mathcal{K}}$-points this map takes a point of $\mathrm{Spec}\, Z_{w,\chi^{(\ell)}}$ corresponding to a character with Langlands parameter ρ to the point of $\mathrm{Spec}\, eZ_n$ corresponding to the Langlands parameter obtained by composing ρ with the L-homomorphism of $\mathcal{T}^{L}_{w}$ into GL_n.

On the other hand, if we fix a generic character Ψ of the unipotent radical U of G, and let Γ be the module $\text{c-Ind}^{G}_{U}\,\Psi$, then it follows from results in [Helm 2016b] that the natural map $eZ_n \to \mathrm{End}_{W(k)[G]}(\Gamma)$ is an isomorphism. We can thus view the map $eZ_n \to Z_{w,\chi^{(\ell)}}$ as the affine group analogue of a Curtis homomorphism. Since the Curtis homomorphisms have such a nice interpretation via Deligne–Lusztig theory, it is natural to ask if a similar phenomenon is at play here:

Question 11.2. Does there exist an adjoint pair of functors

$$i_w : \mathcal{D}^b(\mathrm{Rep}_{W(k)}(T_w)) \to \mathcal{D}^b(\mathrm{Rep}_{W(k)}(G)),$$
$$r_w : \mathcal{D}^b(\mathrm{Rep}_{W(k)}(G)) \to \mathcal{D}^b(\mathrm{Rep}_{W(k)}(F))$$

such that $r_w(\Gamma)$ is a shift of the induction $\text{c-Ind}^{T_w}_{e}\,1$, and the induced homomorphism

$$Z_n \to Z_w$$

is the product over suitable idempotents of the "affine Curtis homomorphisms" constructed above? Moreover, is there a natural geometric construction of such an adjoint pair?

Acknowledgements

We are grateful to Jean-Francois Dat, Robert Kurinczuk, Vincent Sécherre, David Ben-Zvi, and Richard Taylor for helpful conversations and suggestions, and to Gil Moss for his comments on an earlier draft of this paper. We are also deeply indebted to Jack Shotton for noticing a serious error in an earlier version of this paper, and for bringing to our attention the argument of Proposition 7.10 of [Shotton 2018], which proved crucial to correcting this error. This research was partially supported by NSF grant DMS-1161582 and EPSRC grant EP/M029719/1.

References

[Bernstein and Deligne 1984] J. N. Bernstein and P. Deligne, "Le "centre" de Bernstein", pp. 1–32 in *Représentations des groupes réductifs sur un corps local*, edited by P. Deligne, Hermann, Paris, 1984. MR Zbl

[Bonnafé and Kessar 2008] C. Bonnafé and R. Kessar, "On the endomorphism algebras of modular Gelfand–Graev representations", *J. Algebra* **320**:7 (2008), 2847–2870. MR Zbl

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

[Choi 2009] S. H. Choi, *Local deformation lifting spaces of mod l Galois representations*, Ph.D. thesis, Harvard University, 2009, available at https://www.proquest.com/docview/304892280. MR

[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

[Dudas 2009] O. Dudas, "Deligne-Lusztig restriction of a Gelfand–Graev module", *Ann. Sci. Éc. Norm. Supér.* (4) **42**:4 (2009), 653–674. MR Zbl

[Emerton and Helm 2014] M. Emerton and D. Helm, "The local Langlands correspondence for GL_n in families", *Ann. Sci. Éc. Norm. Supér.* (4) **47**:4 (2014), 655–722. MR Zbl

[Helm 2016a] D. Helm, "The Bernstein center of the category of smooth $W(k)[\mathrm{GL}_n(F)]$-modules", *Forum Math. Sigma* **4** (2016), art. id. e11. MR Zbl

[Helm 2016b] D. Helm, "Whittaker models and the integral Bernstein center for GL_n", *Duke Math. J.* **165**:9 (2016), 1597–1628. MR Zbl

[Helm and Moss 2018] D. Helm and G. Moss, "Converse theorems and the local Langlands correspondence in families", *Invent. Math.* **214**:2 (2018), 999–1022. MR Zbl

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

[Vignéras 2001] M.-F. Vignéras, "Correspondance de Langlands semi-simple pour $\mathrm{GL}(n, F)$ modulo $\ell \neq p$", *Invent. Math.* **144**:1 (2001), 177–223. MR Zbl

Communicated by Wee Teck Gan
Received 2018-11-13 Revised 2020-03-11 Accepted 2020-06-30

dhelm@imperial.ac.uk *Department of Mathematics, Imperial College London, United Kingdom*

ALGEBRA AND NUMBER THEORY 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2647

Moduli spaces of symmetric cubic fourfolds and locally symmetric varieties

Chenglong Yu and Zhiwei Zheng

We realize the moduli spaces of cubic fourfolds with specified group actions as arithmetic quotients of complex hyperbolic balls or type IV symmetric domains, and study their compactifications. We prove the geometric (GIT) compactifications are naturally isomorphic to the Hodge theoretic (Looijenga, in many cases Baily–Borel) compactifications. The key ingredients of the proof are the global Torelli theorem by Voisin, the characterization of the image of the period map given by Looijenga and Laza independently, and the functoriality of Looijenga compactifications proved in the Appendix.

A list of symbols can be found on page 2680.

1. Introduction

Cubic fourfolds are intensively studied objects in algebraic geometry. There are many interesting relations and analogues between cubic fourfolds and $K3$ surfaces. The Hodge structure on the primitive middle cohomology $H^4_0(X)$ of a smooth cubic fourfold X is of $K3$ type. On the other hand, the Fano scheme of lines on a smooth cubic fourfold is a hyper-Kähler fourfold of $K3^{[2]}$ type; see [Beauville and Donagi 1985]. Similar to $K3$ surfaces, people have a good understanding of the period map for cubic fourfolds. The period map $\mathscr{P}$ gives an algebraic map from the moduli of smooth cubic fourfolds to an arithmetic quotient of a 20-dimensional type IV domain. This period map is an open embedding due to the global Torelli theorem by Voisin [1986]. The image of the period map is the complement of certain hypersurface arrangement. This was proved by Looijenga [2009] and Laza [2010] independently.

Zarhin [1983] classified the Mumford–Tate groups of $K3$-type Hodge structures. The corresponding Mumford–Tate domains are either complex hyperbolic balls or type IV domains. Examples of those Mumford–Tate groups can arise when the Hodge structures admit extra symmetries. This leads us to study moduli spaces of cubic fourfolds with specified group actions. For cubic fourfolds, any automorphisms are induced from linear automorphisms of $\mathbb{P}^5$. This is a general fact for almost all hypersurfaces in projective spaces with degree at least 3; see [Matsumura and Monsky 1963]. Moreover, in [Zheng 2019], the second author checked that any automorphism of the polarized Hodge structure on the middle cohomology of a smooth cubic fourfold is induced by a unique automorphism of the cubic fourfold. Therefore, the symmetries of polarized Hodge structures for cubic fourfolds can be detected geometrically by linear

MSC2010: primary 14D23; secondary 14D07.
Keywords: cubic fourfold, locally symmetric space, Looijenga compactification.

symmetries. These facts give rise to identifications between moduli spaces constructed by GIT and arithmetic quotients of complex hyperbolic balls or type IV domains. We next review two such examples.

One example regarding the complex hyperbolic ball is given by Looijenga and Swierstra [2007] and Allcock, Carlson and Toledo [2011] independently on the moduli space of cubic threefolds. They attach to cubic threefolds the Hodge structures of cubic fourfolds with specified automorphism with order 3. Explicitly, suppose the cubic threefold is given by a polynomial $F(x_1, \ldots, x_5)$, then the corresponding symmetric cubic fourfold we looking at is $x_0^3 + F(x_1, \ldots, x_5) = 0$. Via this construction, the moduli space of cubic threefolds with at worst ADE singularities is identified with the complement of an irreducible totally geodesic hypersurface in an arithmetic quotient of a complex hyperbolic ball of dimension 10. The phenomena that the image of a period map is the complement of some totally geodesic hypersurfaces in a locally symmetric variety appear in many examples besides cubic fourfolds and cubic threefolds. In fact, type IV domains and complex hyperbolic balls are the only irreducible Hermitian symmetric domains admitting totally geodesic hypersurfaces. Coming back to cubic threefolds, on the geometric side, we have the natural GIT compactification of the moduli space of cubic threefolds. On the Hodge theoretic side, there is a natural compactification of the complement of the hypersurface arrangements, building upon Baily–Borel compactification. The construction of this Hodge-theoretic compactification was carried out by Looijenga [2003a], inspired by work of Shah, consisting of two steps. The first step is a partial blowup of the boundary components of Baily–Borel compactification, sitting between toroidal compactification and Baily–Borel compactification. The second step is a successive blowup of the intersection strata of hyperplane arrangements and blowdown in the opposite direction; see the Appendix for a discussion of Looijenga compactification. The fascinating result proved in [Looijenga and Swierstra 2007; Allcock et al. 2011] is the existence of a natural isomorphism between the GIT compactification and Looijenga compactification, which are from totally different origins.

Another example regarding type IV domain was given by Laza, Pearlstein and Zhang [Laza et al. 2018] recently. They considered the moduli space of pairs consisting of a cubic threefold $F(x_1, \ldots, x_5) = 0$ and a hyperplane section $H(x_1, \ldots, x_5) = 0$, or equivalently the moduli of cubic fourfolds

$$x_0^2 H(x_1, \ldots, x_5) + F(x_1, \ldots, x_5) = 0$$

which have natural involutions $x_0 \mapsto -x_0$. The period map gives rise to an identification between the moduli space of the pairs and an arrangement complement in an arithmetic quotient of a type IV domain of dimension 14. Moreover, Laza, Pearlstein and Zhang showed that with a careful choice of linearization (which is indeed natural as we will discuss in Proposition 6.8) in the GIT construction, the pairs which give rise to symmetric cubic fourfolds with at worst ADE singularities are stable, and their moduli can be identified with the whole arithmetic quotient. Finally, they showed that the GIT compactification is isomorphic to the Baily–Borel compactification of the arithmetic quotient.

The first key observation of this work is that the phenomena in the above two examples should also appear in a much more general situation, namely, for cubic fourfolds with any given symmetry. Along this direction, we are able to unify many examples studied before (including the two above), and produce

many new identifications between GIT compactifications and Hodge-theoretic compactifications. Before giving the main theorems, we introduce some notation.

For a smooth cubic fourfold X, we have the Hodge decomposition

$$H^4(X,\mathbb{C}) = H^{3,1}(X) \oplus H^{2,2}(X) \oplus H^{1,3}(X),$$

where $\dim(H^{3,1}) = \dim(H^{1,3}) = 1$, and $\dim H^{2,2} = 21$. We denote by $\varphi_X : H^4(X,\mathbb{C}) \times H^4(X,\mathbb{C}) \to \mathbb{C}$ the topological intersection pairing, whose restriction to $H^4(X,\mathbb{Z}) \times H^4(X,\mathbb{Z})$ is an integral unimodular bilinear form of signature $(21, 2)$.

Let X be a smooth cubic fourfold with an action of a finite group A. All deformations of the pair (X, A) form a quasiprojective variety $\mathcal{F}$, which is called the moduli space of smooth cubic fourfolds with this given group action. See Section 2 for a GIT construction of $\mathcal{F}$. There is a natural morphism (via forgetting the action of A) from $\mathcal{F}$ to the moduli space $\mathcal{M}$ of smooth cubic fourfolds, which is a finite morphism (see Proposition 2.8). Moreover, when the action of A realizes all the automorphisms of X, the morphism $j : \mathcal{F} \to \mathcal{M}$ is a normalization of its image.

On the other hand, we look at the induced action of A on the Hodge structure of the cubic fourfold X. This induces a character $\zeta : A \to \mathrm{GL}(H^{3,1}(X)) \cong \mathbb{C}^\times$ of A. Denote by $H^4(X)_\zeta$ the ζ-eigenspace of the action of A. There is a Hermitian form $h : H^4(X)_\zeta \times H^4(X)_\zeta \to \mathbb{C}$ defined by $h(x, y) = \varphi_X(x, \bar{y})$ for any $x, y \in H^4(X)_\zeta$. If $\zeta = \bar{\zeta}$, then h has signature $(n', 2)$ and there is an type IV domain $\mathbb{D}$ associated with $(H^4(X)_\zeta, h)$. If $\zeta \neq \bar{\zeta}$, the form h has signature $(n', 1)$ and there is an associated complex hyperbolic ball, which we still denote by $\mathbb{D}$ for the moment, with $(H^4(X)_\zeta, h)$; see Proposition 4.1. The discussion above applies to any cubic fourfolds X' in $\mathcal{F}$ and the Hodge structures on $H^4(X')_\zeta$ give rise to a period map from $\mathcal{F}$ to an arithmetic quotient $\Gamma\backslash\mathbb{D}$. Here Γ is an arithmetic group acting properly discontinuously on $\mathbb{D}$ (see the beginning of Section 4C for the definition).

Notice that n' is the dimension of the Hermitian symmetric domain $\mathbb{D}$. We denote by n the dimension of $\mathcal{F}$. The first main theorem of the paper is the following:

Theorem 1.1 (Main Theorem 1). (i) *We have the equality $n' = n$.*

(ii) *The period map $\mathscr{P} : \mathcal{F} \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$ is an algebraic isomorphism. Here $\mathcal{H}_s$ is a Γ-invariant hyperplane arrangement in $\mathbb{D}$.*

(iii) *The period map $\mathscr{P}$ extends naturally to an algebraic isomorphism $\mathcal{F}_1 \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_*)$, where $\mathcal{F}_1$ is a natural partial completion of $\mathcal{F}$, adding cubic fourfolds with at worst* ADE*-singularities, and $\mathcal{H}_*$ is a Γ-invariant hyperplane arrangement contained in $\mathcal{H}_s$.*

Denote by $\overline{\mathcal{F}}$ the GIT compactification of $\mathcal{F}$; see Section 2B. For a Γ-invariant hyperplane arrangement $\mathcal{H}$ in $\mathbb{D}$, we denote by $\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}}$ the Looijenga compactification of $\Gamma\backslash(\mathbb{D} - \mathcal{H})$; see Section A5. We characterize $\overline{\mathcal{F}}$ via:

Theorem 1.2 (Main Theorem 2). (i) *The period map $\mathscr{P}$ extends to an algebraic isomorphism $\overline{\mathcal{F}} \cong \overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$ between the two projective varieties.*

(ii) *There are two pairs* (G_1, λ_1) *and* (G_2, λ_2), *each consists of a subgroup of* $\mathrm{SL}(6, \mathbb{C})$ *and a character of the subgroup* (*see Definition 5.6*), *such that the hyperplane arrangement* $\mathcal{H}_*$ *is empty if and only if for* $i = 1$ *or* 2, *there exists* $h \in \mathrm{GL}(6, \mathbb{C})$ *with* $h^{-1}Ah \subset G_i$ *and* $\lambda(a) = \lambda_i(h^{-1}ah)$ *for any* $a \in A$. *In this case*, *the Looijenga compactification* $\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$ *is the Baily–Borel compactification* $\overline{\Gamma\backslash\mathbb{D}}^{bb}$. *See Theorem 5.7 for a complete statement.*

In the previous works for cubic fourfolds [Looijenga 2009; Laza 2010], cubic threefolds [Allcock et al. 2011; Looijenga and Swierstra 2007] and pairs consisting of a cubic threefold and a hyperplane section [Laza et al. 2018], the extended isomorphisms between the GIT compactifications and Looijenga compactifications rely on the machinery developed in [Looijenga 2003a; 2003b]. The key observation is that the period map also identifies the GIT polarization and the automorphic bundle on the period domain. If the period map can be extended to a Zariski-open subset U such that its complement in the GIT compactification has codimension at least 2, then the coordinate ring of the GIT compactification consists of sections (of the GIT polarization) over U. On the other hand, if each nonempty intersection of members in $\mathcal{H}$ has dimension at least 2, then the Γ-invariant automorphic sections with poles along $\mathcal{H}$ form the coordinate ring of the Looijenga compactification. Therefore, the two compactifications are identified if the two conditions hold.

For each case, the hard work on GIT side is to extend the defining domain of the period map to moduli space of varieties with at worst simple singularities and obtain codimension estimate for the indeterminacy locus. On the period domain side one need to obtain dimension estimate for all possible intersections of members in $\mathcal{H}$. Usually this is achieved by careful lattice analysis. In some cases people also need the correct choice of polarization on the GIT side in order to have such an extension. In our setting for Theorems 1.1 and 1.2, the dimension estimate fails in some cases, for example when A is a cyclic group of order 7; see Remark 6.7. So the previous approach does not work for all symmetry types.

We developed a new approach by considering the functorial properties on both GIT and Hodge theory side. We explain the proof of Theorems 1.1 and 1.2 with the following diagram:

$$\begin{array}{ccc}
\mathcal{F} & \xrightarrow{\cong} & \Gamma\backslash(\mathbb{D} - \mathcal{H}_s) \\
\downarrow & & \downarrow \\
\overline{\mathcal{F}} & \dashrightarrow & \overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*} \\
\downarrow{\scriptstyle j} & & \downarrow{\scriptstyle \pi} \\
\overline{\mathcal{M}} & \xrightarrow{\mathscr{P}} & \overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\mathcal{H}_\infty}
\end{array}$$

Here $\overline{\mathcal{M}}$ is the GIT compactification of the moduli space of cubic fourfolds, $\widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ is the period domain for cubic fourfolds and $\mathcal{H}_\infty$ is a ($\widehat{\Gamma}$-invariant) hyperplane arrangement. These are explained in detail in Section 3. The bottom isomorphism in the above diagram is the main result in [Looijenga 2009; Laza 2010]. The top isomorphism is Proposition 4.10 proved in Section 4, which relies essentially on the global Torelli for cubic fourfolds. The left vertical morphism j is finite due to a classical result

by Luna [1975] (with a modified version for projective GIT quotients; see [Ressayre 2010]). This is included in Proposition 2.7. The right vertical morphism π is also finite, which is proved in the Appendix; see Theorem A.13. After establishing the two finiteness results (of j and π) and the horizontal bimeromorphism between $\overline{\mathcal{F}}$ and $\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$, we show that the period map $\mathscr{P}: \mathcal{F} \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$ extends to an isomorphism $\overline{\mathcal{F}} \cong \overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$ by Lemma 5.4.

Our formalism of the proof does not need the codimension and dimension estimates, and hence avoids complicated GIT and lattice analysis. Finally, we reduce the complication to the proof of the functorial property of Looijenga compactifications. This allows us to deal with the theory uniformly and systematically for all symmetry types of cubic fourfolds. To the best of our knowledge, this formalism is new and may have further applications.

In many cases, the hyperplane arrangement $\mathcal{H}_*$ is empty; hence the Looijenga compactification is simply Baily–Borel compactification (for example, [Laza et al. 2018]). We discuss in Section 5 a criterion (Theorem 5.7) based on the symmetry type for the emptiness of $\mathcal{H}_*$. In particular, we apply the criterion to determine the emptiness of $\mathcal{H}_*$ for all symmetry type with A a prime-order cyclic group.

We end the introduction with a discussion on future works. A closely related question is to classify automorphism groups of cubic fourfolds. There are 13 conjugacy classes of prime-order automorphisms of smooth cubic fourfolds (see [González-Aguilera and Liendo 2011]). For two of them, our main theorems recover some of the main results in [Allcock et al. 2011; Looijenga and Swierstra 2007; Laza et al. 2018]. We will discuss these examples in more detail in Sections 6A and 6B.

Cubic fourfolds have very close relation with hyper-Kähler manifolds; see [Beauville and Donagi 1985; Hassett 2000]. For a smooth cubic fourfold X, its Fano scheme of lines is a polarized hyper-Kähler fourfold of $K3^{[2]}$ type. The automorphism group of a smooth cubic fourfold X is naturally identified with the automorphism group of the associated polarized hyper-Kähler manifold; see [Fu 2016]. The classification of automorphism groups of hyper-Kähler manifolds has appealed to a lot of interests recently. There is a systematic study by Mongardi in his thesis [2012; 2013; 2016]. Höhn and Mason [2019] classified all maximal finite symplectic automorphism groups of hyper-Kähler fourfolds of $K3^{[2]}$ type. Those groups are all subgroups of the Conway group. Recently, Laza and the second author classified all finite symplectic automorphism groups of smooth cubic fourfolds; see [Laza and Zheng 2019] . While related to Höhn and Mason's classification [2019], the main difference in [Laza and Zheng 2019] is that the authors are dealing with "polarized" hyper-Kähler fourfolds. Moreover, in many cases the explicit normal forms for the cubic fourfolds with a specified symplectic automorphism group are given. This classification offers a bunch of examples for Theorems 1.1 and 1.2 with $\mathbb{D}$ being type IV domains. Another closely related problem is to characterize the moduli spaces of symmetric or lattice-polarized hyper-Kähler manifolds. There are works along this direction; see [Dolgachev and Kondō 2007, Section 11; Artebani et al. 2011, Section 9; Joumaah 2016; Camere 2016, Section 3; Boissière et al. 2016, Section 5; Boissière et al. 2019]

The symmetries of the Hodge structures can also arise from degenerations of cubic fourfolds or $K3$ surfaces. For example, we consider a one-parameter degeneration of smooth cubic fourfolds to a singular cubic fourfold with only one node. The monodromy of the family gives a reflection on the primitive middle

cohomology. The Hodge structures fixed by this reflection form a hyperplane in the period domain $\widehat{\mathbb{D}}$, which is a 19-dimensional type IV domain. On the geometric side, such singular cubic fourfolds naturally give rise to $K3$ surfaces of degree 6. So the proof above can also be applied to obtain comparison between GIT compactification of $K3$ surfaces of degree 6 and Baily–Borel compactification of period domain. Following this perspective, we are able to realize moduli of singular sextic curves (regarding as singular $K3$ surfaces of degree 2) as arithmetic quotients of type IV domains and again identify GIT compactification and Looijenga compactification; see [Yu and Zheng 2018].

Structure of the paper. Section 2 is devoted to the GIT construction of symmetric hypersurfaces in general. In Section 3 we review concepts about cubic fourfolds, and introduce the global Torelli theorem. Sections 4 and 5 are the main part of the paper, where we formulate and prove our main theorems. As we have mentioned, one of the key ingredients in the proof is the functorial property of Looijenga compactifications. This is treated in the Appendix. Moduli of cubic fourfolds with specified action by cyclic group is discussed in Section 6.

2. General setup: symmetric hypersurfaces

2A. ***Space of symmetric polynomials.*** Let V be a complex vector space of dimension $k+2$. Denote by $\mathrm{Sym}^d(V^*)$ the space of degree d polynomials on V. We have the natural action of $\mathrm{SL}(V)$ on $\mathrm{Sym}^d(V^*)$, namely, $g(F) = F \circ g^{-1}$ for $g \in \mathrm{SL}(V)$ and $F \in \mathrm{Sym}^d(V^*)$. The center of $\mathrm{SL}(V)$ is the group μ_{k+2} consisting of $(k+2)$-th roots of unity. Let A be a finite subgroup of $\mathrm{SL}(V)$ containing μ_{k+2} and denote by $\bar{A} = A/\mu_{k+2}$ the image of A in $\mathrm{PSL}(V)$. Then $\mathrm{Sym}^d(V^*)$ is a representation of A.

For any $\xi \in \mu_{k+2}$ and $F \in \mathrm{Sym}^d(V^*)$, we have $\xi(F) = \xi^{-d}F$. Let $\lambda : A \to \mathbb{C}^\times$ be a character of A such that $\lambda|_{\mu_{k+2}}$ sends $\xi \in \mu_{k+2}$ to ξ^{-d}. Let $\mathcal{V}_\lambda$ be the λ-eigenspace of $\mathrm{Sym}^d(V^*)$. We write $\mathcal{V} = \mathcal{V}_\lambda$ for short. Geometrically, an element in $\mathcal{V}$ determines a degree d hypersurface (not necessarily smooth) in $\mathbb{P}V$, whose automorphism group contains $\bar{A}$.

Two pairs (A_1, λ_1) and (A_2, λ_2) are called equivalent if and only if there exists $g \in \mathrm{SL}(V)$ such that $gA_1g^{-1} = A_2$ and $\lambda_1(a_1) = \lambda_2(ga_1g^{-1})$ for any $a_1 \in A_1$. We call an equivalence class a symmetry type, denoted by T. There is a poset structure on the space of symmetry types, namely, $T_2 \le T_1$ if T_1, T_2 are represented by (A_1, λ_1), (A_2, λ_2) respectively, such that $A_1 \subset A_2$ and $\lambda_1 = \lambda_2|_{A_1}$. Notice that the space $\mathcal{V}$ depends on the representative (A, λ) of T.

For $F \in \mathcal{V}$, we denote by $Z(F)$ the hypersurface defined by F in $\mathbb{P}V$. For $X = Z(F)$, we denote by $\mathrm{Aut}(X)$ the group of elements in $\mathrm{PSL}(V)$ preserving X, and by $\mathrm{Aut}(F)$ the preimage of $\mathrm{Aut}(X)$ in $\mathrm{SL}(V)$. From [Matsumura and Monsky 1963, Theorems 1 and 2] we have:

Theorem 2.1 (Matsumura–Monsky). *When X is smooth, $d \ge 3$, $k \ge 2$,*

(i) *the group* $\mathrm{Aut}(X)$ *is finite,*

(ii) *if* $(d, k) \ne (4, 2)$, *the group* $\mathrm{Aut}(X)$ *contains all biregular automorphisms of X.*

For any $X = Z(F)$, the group $\bar{A}$ is naturally a subgroup of $\mathrm{Aut}(X)$. We propose the following conditions on the symmetry type T:

Condition 2.2. The linear space $\mathcal{V}$ contains a point F defining a smooth hypersurface.

Condition 2.3. The linear space $\mathcal{V}$ contains a point F with the hypersurface $X = Z(F)$ smooth and $\bar{A} = \mathrm{Aut}(X)$.

Remark 2.4. Condition 2.3 is indeed stronger than Condition 2.2. For example, a smooth cubic fourfold with an automorphism of order 7 can be defined by a polynomial

$$F(x_0, \ldots, x_6) = x_0^2x_4 + x_1^2x_2 + x_0x_2^2 + x_3^2x_5 + x_3x_4^2 + x_1x_5^2 + ax_0x_1x_3 + bx_2x_4x_5$$

with $a, b \in \mathbb{C}$ (see Proposition 6.1). The order 7 automorphism ρ is given by $x_i \mapsto \omega^{i+1}x_i$ for ω a primitive 7-root of unity. On the other hand, such a polynomial always admits an order 3 automorphism given by $(x_0, x_1, x_2, x_3, x_4, x_5) \mapsto (x_1, x_3, x_5, x_0, x_2, x_4)$. If we take $\bar{A} = \langle\rho\rangle$ and take λ trivial, then (A, λ) is a symmetry type satisfying Condition 2.2. However, for a generic member $F \in \mathcal{V}$, the automorphism group $\mathrm{Aut}(Z(F))$ is strictly larger than $\bar{A}$. Thus the symmetry type does not satisfy Condition 2.3. See [Laza and Zheng 2019, Theorem 1.2] for more such examples.

For T satisfying Condition 2.2, a generic point in $\mathcal{V}$ defines a smooth hypersurface. We have a similar result about Condition 2.3.

Proposition 2.5. *If $T = [(A, \lambda)]$ satisfies Condition 2.3, then a generic element in $\mathcal{V}$ defines a smooth hypersurface X with $\bar{A} = \mathrm{Aut}(X)$.*

Proof. Suppose $F \in \mathcal{V}$ with $X = Z(F)$ smooth, and $A = \mathrm{Aut}(X)$. Then any small deformation F_1 of F in $\mathcal{V}$ defines a smooth hypersurface $Z(F_1)$. By Proposition 2.1 in [Zheng 2019], when F_1 is sufficiently close to F, there exists $g \in \mathrm{PSL}(V)$ such that $g\mathrm{Aut}(Z(F_1))g^{-1} \subset \mathrm{Aut}(X) = \bar{A}$. Since $F_1 \in \mathcal{V}$, we have $\bar{A} \subset \mathrm{Aut}(Z(F_1))$; hence $\bar{A} = \mathrm{Aut}(Z(F_1))$. □

2B. ***Geometric invariant theory for symmetric hypersurfaces.*** Now we assume that $d \geq 3$, $k \geq 2$. Given a symmetry type $T = [(A, \lambda)]$ satisfying Condition 2.2, let $C = \{g \in \mathrm{SL}(V) \mid gag^{-1} = a \text{ for all } a \in A\}$ and $N = \{g \in \mathrm{SL}(V) \mid gAg^{-1} = A, \lambda(gag^{-1}) = \lambda(a) \text{ for all } a \in A\}$ be two reductive subgroups of $\mathrm{SL}(V)$. For reductivity, see [Luna and Richardson 1979, Lemma 1.1].

Lemma 2.6. *There is a natural action of N on $\mathcal{V}$, under which the points in $\mathcal{V}$ defining smooth hypersurfaces are stable.*

Proof. For any $g \in N$ and $F \in \mathcal{V}$, we need to show $g(F) \in \mathcal{V}$. For any $a \in A$, we have

$$a(g(F)) = g(g^{-1}ag(F)) = g(\lambda(g^{-1}ag)F) = g\lambda(a)F = \lambda(a)g(F),$$

which implies $g(F) \in \mathcal{V}$ by definition of $\mathcal{V}$. Therefore, there is a natural action of N on $\mathcal{V}$.

Now take $F \in \mathcal{V}$ with $X = Z(F)$ smooth. Then $\mathrm{Aut}(X)$ is finite by Theorem 2.1. Since the stabilizer group of F under the action of N is a subgroup of $\mathrm{Aut}(F)$, it is also finite. Moreover, NF is closed in $\mathrm{SL}(V)F$, and the latter is closed in $\mathrm{Sym}^d(V^*)$ since $Z(F)$ is smooth. Thus NF is closed in $\mathrm{Sym}^d(V^*)$; hence also closed in $\mathcal{V}$. We conclude that F is stable under the action of N. □

Denote $\mathcal{V}^{sm} = \{F \in \mathcal{V} \mid Z(F) \text{ smooth}\}$, by $\mathcal{V}^{ss}$ the set of semistable elements in $\mathcal{V}$ under the action of N, and by $\mathbb{P}\mathcal{V}^{sm}$, $\mathbb{P}\mathcal{V}^{ss}$ their projectivizations. By Lemma 2.6, we can take $\mathcal{F} = N\backslash\backslash\mathbb{P}\mathcal{V}^{sm}$ to be the GIT quotient, with the GIT compactification $\overline{\mathcal{F}} = N\backslash\backslash\mathbb{P}\mathcal{V}^{ss}$. Different representatives of the symmetry type induce canonically isomorphic GIT-quotients. Define $\mathcal{M} = \mathrm{SL}(V)\backslash\backslash\mathbb{P}\mathrm{Sym}^d(V^*)^{sm}$ to be the moduli space of smooth degree d hypersurfaces in $\mathbb{P}(V)$, with the GIT compactification $\overline{\mathcal{M}} = \mathrm{SL}(V)\backslash\backslash\mathbb{P}\mathrm{Sym}^d(V^*)^{ss}$. We have the following proposition:

Proposition 2.7. *There is a natural morphism $j : \overline{\mathcal{F}} \to \overline{\mathcal{M}}$ sending $[F] \in \mathcal{F}$ to $[F] \in \mathcal{M}$ for any $F \in \mathcal{V}^{sm}$. This morphism is finite. When T satisfies Condition 2.3, the morphism j is a normalization of its image.*

Proof. Here we use a projective version of the main theorem in [Luna 1975]. See the argument of Proposition 8 in [Ressayre 2010]. Since A is a finite group, there exists certain symmetric power $\mathrm{Sym}^l(\mathcal{V})$ on which the A-action is trivial. Consider the $\mathrm{SL}(V)$-action on the coordinate ring $\bigoplus_m \mathrm{Sym}^{lm}(\mathrm{Sym}^d(V^*)^*)$ of $(\mathbb{P}(\mathrm{Sym}^d(V^*)), \mathcal{O}(l))$. Notice that N is of finite index in the normalizer of A in $\mathrm{SL}(V)$. By the main theorem in [Luna 1975], we have a finite morphism

$$\tilde{j} : \mathrm{Spec}\Big(\Big(\bigoplus_m \mathrm{Sym}^{lm}(\mathcal{V}^*)\Big)^N\Big) \to \mathrm{Spec}\Big(\Big(\bigoplus_m \mathrm{Sym}^{lm}(\mathrm{Sym}^d(V^*)^*)\Big)^{\mathrm{SL}(V)}\Big)$$

sending semistable points to semistable points, and preserving the cone structures. Thus $\tilde{j}$ does not contract any line; hence descends to a finite morphism $j : \overline{\mathcal{F}} \to \overline{\mathcal{M}}$. The morphism j sends $[F] \in \mathcal{F}$ to $[F] \in \mathcal{M}$ for any $F \in \mathcal{V}^{sm}$.

We claim that when T satisfies Condition 2.3, the morphism j is generically injective. Take generically $F_1, F_2 \in \mathcal{V}$ and assume $[F_1] = [F_2]$ in $\mathcal{M}$. Then there exists $g \in \mathrm{SL}(V)$ with $g(F_1) = F_2$. By the calculation

$$g^{-1}ag(F_1) = g^{-1}a(F_2) = g^{-1}\lambda(a)F_2 = \lambda(a)F_1, \tag{1}$$

we have that $g^{-1}ag \in \mathrm{SL}(V)$ is an automorphism of $Z(F_1)$. By the genericity of F_1, we have $A \cong \mathrm{Aut}(F_1)$, which implies that $g^{-1}ag \in A$. Then by equation (1) and $F_1 \in \mathcal{V}$, we have $\lambda(g^{-1}ag) = \lambda(a)$. This implies that $g \in N$, hence $[F_1] = [F_2]$ in $\mathcal{F}$. Thus j is generically injective.

Moreover, since $\overline{\mathcal{F}}$ is normal and projective, j is a normalization of its image. □

Let $T = [(A, \lambda)]$ be a symmetry type satisfying Condition 2.2. Consider the automorphism groups $\mathrm{Aut}(F)$ for all $F \in \mathcal{V}^{sm}$. There exists $F' \in \mathcal{V}^{sm}$ such that $\#\mathrm{Aut}(F')$ is minimal. Let $A' = \mathrm{Aut}(F')$. For any $a \in A'$, there exists $\lambda'(a) \in \mathbb{C}$ with $a(F') = \lambda'(a)F'$. Then we have a symmetry type $T' = [(A', \lambda')]$. It is straightforward that $T \geq T'$, and T' satisfies Condition 2.3. Similar as T, we have for T' correspondingly N', $\mathcal{V}'$ and $\overline{\mathcal{F}'}$. We have the following proposition:

Proposition 2.8. *There exists a natural finite morphism $\overline{\mathcal{F}} \to \overline{\mathcal{F}'}$.*

Proof. By Proposition 2.7, we have two finite morphisms $j : \overline{\mathcal{F}} \to \overline{\mathcal{M}}$ and $j' : \overline{\mathcal{F}'} \to \overline{\mathcal{M}}$, and the latter one is a normalization of its image. We show that j and j' have the same image. We have $j'(\overline{\mathcal{F}'}) \subset j(\overline{\mathcal{F}})$ since $\mathcal{V}' \subset \mathcal{V}$. By Proposition 2.1 in [Zheng 2019], when $F'' \in \mathcal{V}$ is sufficiently close to F', there exists $g \in \mathrm{SL}(V)$, such that $g\mathrm{Aut}(F'')g^{-1} \subset \mathrm{Aut}(F') = A'$. By minimality of $\#A'$, we have $g\mathrm{Aut}(F'')g^{-1} = A'$.

This implies that $\mathrm{Aut}(g(F'')) = A'$; hence $g(F'') \in \mathcal{V}'$. We then have $\dim(j(\overline{\mathcal{F}})) \leq \dim(j'(\overline{\mathcal{F}'}))$. By irreducibilities of the two images, they are the same.

By universal property of normalization, the morphism j factors through j'. Therefore, we have naturally a finite morphism $\overline{\mathcal{F}} \to \overline{\mathcal{F}'}$. □

Remark 2.9. The fiber of the finite morphism $\overline{\mathcal{F}} \to \overline{\mathcal{F}'}$ over $[F']$ is naturally bijective to the orbit of (A, λ) in the set of subdata of (A', λ') under the action of N'.

2C. ***Universal deformation.*** We fix a type $T = [(A, \lambda)]$ satisfying Condition 2.2, and assume $d \geq 3$ and $k \geq 2$. Next we use Luna's étale slice theorem to describe the local structure of $\mathcal{F}$, and construct the universal family of smooth degree d k-folds of type T. We essentially follow the argument in [Zheng 2019, Section 2]. For Luna's étale slice theorem and its proof, see [Luna 1973] or [Vinberg and Popov 1994].

Denote by G the centralizer of $\bar{A}$ in $\mathrm{PSL}(V)$. Recall that $\mathbb{P}\mathcal{V}^{sm}$ is the space of smooth degree d k-folds of symmetry type (A, λ). As a closed subvariety of the affine variety $\mathbb{P}\mathrm{Sym}^d(V^*)^{sm}$, the variety $\mathbb{P}\mathcal{V}^{sm}$ is also affine. There is a natural action of G on $\mathbb{P}\mathcal{V}^{sm}$. For any $x \in \mathbb{P}\mathcal{V}^{sm}$, we denote by Gx the orbit of x and by G_x the stabilizer of x. By Lemma 2.6, Gx is closed in the affine variety $\mathbb{P}\mathcal{V}^{sm}$ and G_x is finite. For a G_x-invariant subvariety S of X containing x, there is an action of G_x on $G \times S$ given by $g(h, y) = (hg^{-1}, gy)$ for any $g \in G_x$, $h \in G$, $y \in S$. We denote by $G \times^{G_x} S$ the quotient of $G \times S$ by this action. By Luna's étale slice theorem, there exists a smooth, locally closed, G_x-invariant subvariety S containing x, such that

(i) the image of $\kappa : G \times^{G_x} S \to \mathbb{P}\mathcal{V}^{sm}$, denoted by U, is Zariski-open and G-invariant,

(ii) the morphism $\kappa : G \times^{G_x} S \to U$ is étale,

(iii) the morphism $G\backslash\!\backslash\kappa : G_x\backslash\!\backslash S \to G\backslash\!\backslash U$ is étale,

(iv) the above two morphisms induce an isomorphism

$$G \times^{G_x} S \cong U \underset{G\backslash\!\backslash U}{\times} G_x\backslash\!\backslash S. \tag{2}$$

We can shrink S in the analytic category such that

(v) S is G_x-invariant, contractible and contains x, with $U = \kappa(G \times^{G_x} S)$ a G-invariant open subset of $\mathbb{P}\mathcal{V}^{sm}$,

(vi) the morphism between analytic spaces: $G_x\backslash\!\backslash S \to G\backslash\!\backslash U$ is an isomorphism.

From (2), we have an isomorphism between analytic spaces,

$$G \times^{G_x} S \cong U,$$

by which we have a principal G_x-bundle $G \times S \to U$. In particular, $G \times S \to U$ is a covering map.

Definition 2.10. For any symmetry type T, we define a category $\mathcal{C}^T_{d,k}$ as follows. The objects are families of degree d k-folds of type T with a specified central fiber. The morphisms are holomorphic maps between families, sending central fiber to central fiber and compatible with the action of $\bar{A}$.

Proposition 2.11. *The family $\mathscr{X}_S$ of degree d k-folds of type T over S has the following universal property. For any subfamily $\mathscr{X}_{S'} \to S' \subset U$ of degree d k-folds of type T containing a central fiber X' with an isomorphism $f : X' \cong X$ compatible with the actions of $\bar{A}$, we have a unique morphism in the category $\mathcal{C}^T_{d,k}$:*

$$\begin{array}{ccc} \mathscr{X}_{S'} & \xrightarrow{\tilde{f}} & \mathscr{X}_S \\ \downarrow & & \downarrow \\ S' & \longrightarrow & S \end{array}$$

such that the restriction of $\tilde{f}$ to X' is f. Moreover, for any two fibers X_1, X_2 of $\mathscr{X}_S$ with an isomorphism $g : X_1 \to X_2$ compatible with the actions of $\bar{A}$, we can extend g uniquely to a morphism $\widetilde{g} : \mathscr{X}_S \to \mathscr{X}_S$ in $\mathcal{C}^T_{d,k}$.

Proof. The base S' lies in U and is covered by $G \times S$. Thus we have a unique lifting $S' \hookrightarrow G \times S$, sending x' to (f^{-1}, x). In other words, we have uniquely a morphism $\tilde{f} : \mathscr{X}_{S'} \to \mathscr{X}_S$, which restricts to f on X'.

Now suppose X_1, X_2 are two fibers of $\mathscr{X}_S$ with an isomorphism $g : X_1 \cong X_2$. Denote by x_1, x_2 the corresponding base points in S. Then (g, x_1), $(\mathrm{id}, x_2) \in G \times S$ have the same image in U. Since $G \times S \to U$ is a principal G_x-bundle, the two pairs (g, x_1) and (id, x_2) are G_x-equivalent; hence $g \in G_x$. The proposition follows. □

We have the following lemma, which is used in the proof of Proposition 4.8. Since it holds for general degree d k-folds, we state and prove it here.

Lemma 2.12. *Let*

$$\begin{array}{ccc} \mathscr{X} & \hookrightarrow & S \times \mathbb{P}V \\ \downarrow & \swarrow & \\ S & & \end{array}$$

be a family of smooth degree d k-folds, with the base S contractible. Suppose there is a group $\widetilde{A}$, such that for all $s \in S$, the fiber $\mathscr{X}_s$ admits a biregular action of $\widetilde{A}$, with induced actions on $H^n(\mathscr{X}_s, \mathbb{Z})$ compatible with respect to the local trivialization. Then there exists an action of $\widetilde{A}$ on the whole family $\mathscr{X} \to S$ inducing on each fiber the existing action.

To prove this, we need another lemma from [Javanpeykar and Loughran 2017, Proposition 2.12; Matsumura and Monsky 1963]:

Lemma 2.13. *For $d \geq 3$, $k \geq 2$, and a smooth degree d k-fold X, the induced action of* $\mathrm{Aut}(X)$ *on $H^k(X, \mathbb{Z})$ is faithful.*

Proof of Lemma 2.12. Without loss of generality, we can assume the action of $\widetilde{A}$ on each $H^n(\mathscr{X}_s, \mathbb{Z})$ is faithful. Take any $s \in S$. By Proposition 2.1 in [Zheng 2019], there is a universal hypersurface family $\mathscr{X}'$ of $\mathscr{X}_s$, such that any isomorphism between two fibers (may coincide) of $\mathscr{X}'$ comes from an automorphism of the central fiber $\mathscr{X}_s$. There exists an open neighborhood U of s in S, with a unique morphism $\mathscr{X}|_U \to \mathscr{X}'$. Then for any $s' \in U$, the action of $\widetilde{A}$ on $\mathscr{X}_{s'}$ is induced by a subgroup $\widetilde{A}'$ of $\mathrm{Aut}(\mathscr{X}_s)$. By Lemma 2.13, and the compatibility of induced actions of $\widetilde{A}$ on $\mathscr{X}_s$ and $\mathscr{X}_{s'}$, we have $\widetilde{A} = \widetilde{A}'$ as subgroups of $\mathrm{Aut}(\mathscr{X}_s)$. Therefore, the actions of $\widetilde{A}$ on fibers of $\mathscr{X} \to S$ glue to an action of $\widetilde{A}$ on the whole family. □

3. Period map for smooth cubic fourfolds

In this section we recall some fundamental facts on the period map for cubic fourfolds, the main references are [Voisin 1986; Hassett 2000; Looijenga 2009; Laza 2009; 2010].

Take $(d, k) = (3, 4)$. Then we have $\mathcal{M}$ the moduli of smooth cubic fourfolds, as a Zariski-open subset of its GIT compactification $\overline{\mathcal{M}}$. Let X be a smooth cubic fourfold. We denote by φ_X the intersection pairing on $H^4(X, \mathbb{Z})$. Then $(H^4(X, \mathbb{Z}), \varphi_X)$ is an odd unimodular lattice of signature $(21, 2)$. Denote by η_X the square of the hyperplane class of X. Then $H^4_0(X, \mathbb{Z}) := \eta_X^\perp$ is an even lattice of discriminant 3. Now we define $(\Lambda, \Lambda_0, \eta)$ to be an abstract data isomorphic to $(H^4(X, \mathbb{Z}), H^4_0(X, \mathbb{Z}), \eta_X)$. This does not depend on the choice of the cubic fourfold X.

Definition 3.1. A marking of the cubic fourfold X is an isomorphism $\Phi : H^4(X, \mathbb{Z}) \cong \Lambda$ of lattices sending η_X to η.

Two marked cubic fourfolds (X_1, Φ_1) and (X_2, Φ_2) are called equivalent if there exists a linear isomorphism $g : X_1 \to X_2$ such that $\Phi_1 = g^*\Phi_2$. Let $\mathcal{M}^m$ be the set of equivalence classes of marked cubic fourfolds. From [Zheng 2019, Section 3], we have:

Proposition 3.2. *The set $\mathcal{M}^m$ is a complex manifold in a natural way.*

Next we define the period domain and period map for cubic fourfolds. Let

$$\widetilde{\mathbb{D}} := \mathbb{P}\{x \in (\Lambda_0)_{\mathbb{C}} \mid \varphi(x, x) = 0, \varphi(x, \bar{x}) < 0\}.$$

This is an analytically open subset of a quadric hypersurface in $\mathbb{P}(\Lambda_0)_{\mathbb{C}}$, and has two connected components. We have naturally a holomorphic map

$$\widetilde{\mathscr{P}} : \mathcal{M}^m \to \widetilde{\mathbb{D}}$$

sending $(X, \Phi) \in \mathcal{M}^m$ to $\Phi(H^{3,1}(X))$. It is called the local period map for cubic fourfolds.

Let $\widehat{\mathbb{D}}$ be one connected component of $\widetilde{\mathbb{D}}$ and $\widehat{\Gamma}$ be the index 2 subgroup of $\mathrm{Aut}(\Lambda, \varphi, \eta)$ which leaves the component $\widehat{\mathbb{D}}$ stable. Then $\widehat{\Gamma}$ is an arithmetic group acting on $\widehat{\mathbb{D}}$, and $\widetilde{\mathscr{P}}$ descends to

$$\mathscr{P} : \mathcal{M} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}},$$

which is called the (global) period map for cubic fourfolds.

Remark 3.3. The subgroup $\widehat{\Gamma}$ consists of elements in Γ with spinor norm 1. Since there exist vectors in Λ_0 with self intersection -2, the group $\widehat{\Gamma}$ is of index 2 in $\mathrm{Aut}(\Lambda, \varphi, \eta)$.

The global Torelli theorem was originally proved by Voisin [1986], with an erratum based on some work by Laza [2009]:

Theorem 3.4 (Voisin). *The period map $\mathscr{P}$ is an open embedding.*

Remark 3.5. In fact, the period map $\mathscr{P}$ is algebraic; see the discussion in [Hassett 2000, Proposition 2.2.3].

We give a lemma which is constantly used; see [Zheng 2019, Proposition 1.3].

Lemma 3.6. *Take X a smooth cubic fourfold. Then* $\mathrm{Aut}(X) \cong \mathrm{Aut}(H^4(X,\mathbb{Z}), \varphi_X, \eta_X, H^{3,1}(X))$.

We have a refined version of Theorem 3.4:

Proposition 3.7 (Voisin, Hassett, Looijenga, Laza). *The local period map $\widetilde{\mathscr{P}} : \mathcal{M}^m \to \widetilde{\mathbb{D}}$ is an open embedding, with image being the complement of a hyperplane arrangement invariant under the action of* $\mathrm{Aut}(\Lambda, \eta)$ *on* $\widetilde{D}$.

Proof. Combining Theorem 3.4 and Lemma 3.6 we have injectivity. The characterization of the image of $\widetilde{\mathscr{P}}$ is due to Looijenga [2009] and Laza [2010, Theorem 1.1], a more precise version is discussed in Proposition 4.7. □

4. Period maps for symmetric cubic fourfolds

4A. ***Local period map for symmetric cubic fourfolds.*** In this section we are going to discuss the local and global period maps for symmetric cubic fourfolds. Let $(d,k) = (3,4)$, and fix a symmetry type $T = [(A,\lambda)]$ satisfying Condition 2.2. We first introduce the local period domains with actions of arithmetic groups. Take $X = Z(F)$ for a generic point $F \in \mathcal{V}$. Recall that the action of A on X induces an action of A on $H^{3,1}(X)$. This action is a character $\zeta : A \to \mathbb{C}^\times$ with trivial restriction on μ_{k+2}. We denote

$$H^4(X)_\zeta = \{x \in H^4(X) \mid ax = \zeta(a)x \ \text{ for all } a \in A\}.$$

Define a Hermitian form $h : H^4(X)_\zeta \times H^4(X)_\zeta \to \mathbb{C}$ by $h(x,y) = \varphi(x,\bar{y})$. Denote by σ_X the action of A on $H^4(X,\mathbb{Z})$. Let σ be an action of A on Λ, making (Λ,η,σ) isomorphic to $(H^4(X,\mathbb{Z}),\eta_X,\sigma_X)$. Denote by $\Lambda_\zeta \subset \Lambda_0 \otimes \mathbb{C}$ the ζ-eigenspace of the action of A on $(\Lambda_0)_\mathbb{C}$.

Proposition 4.1. *The Hermitian form h has signature $(n',2)$ if $\zeta = \bar{\zeta}$ (this is also equivalent to $\zeta(A) \subset \mu_2$); it has signature $(n',1)$ otherwise. Here n' is a nonnegative integer independent of the choice of X.*

Proof. Notice that the lattice $H^4(X,\mathbb{Z})$ has signature $(21,2)$, with negative part $H^{3,1}(X) \oplus H^{1,3}(X)$. If $\zeta(A)$ is not contained in μ_2, we have $\zeta \neq \bar{\zeta}$. Since $H^{1,3}$ lies in $\bar{\zeta}$-eigenspace, the signature of h is $(n',1)$.

For the case $\zeta(A) \subset \mu_2$, both $H^{3,1}(X)$ and $H^{1,3}(X)$ are contained in H_ζ; thus h has signature $(n',2)$. □

An isomorphism $\Phi : (H^4(X,\mathbb{Z}),\eta_X,\sigma_X) \cong (\Lambda,\eta,\sigma)$ is called a T-marking of X. We consider pairs consisting of a smooth cubic fourfold and its T-marking. Two such pairs (X_1,Φ_1) and (X_2,Φ_2) are equivalent if there exists $g \in G$ such that $\Phi_1 = g^*\Phi_2$. Letting $\mathcal{F}^m$ be the set of equivalence classes of such pairs, we have:

Proposition 4.2. *The set $\mathcal{F}^m$ is naturally a complex manifold.*

Proof. First we describe the local charts on $\mathcal{F}^m$. Take a point $(X,\Phi) \in \mathcal{F}^m$, and take a universal deformation $\mathscr{X}_S \to S$ of X as in Proposition 2.11. Since S is contractible, the local system $R^4\pi_*(\mathbb{Z})$ is trivializable over S and the T-marking Φ of X naturally extends to a T-marking for every fiber of $\mathscr{X}_S \to S$. Thus we have a map

$$\alpha : S \to \mathcal{F}^m.$$

We first show that α is injective. Suppose X_1, X_2 are two fibers of $\mathscr{X}_S$, with Φ_1, Φ_2 the induced T-markings by Φ, such that (X_1, Φ_1) and (X_2, Φ_2) represent the same point in $\mathcal{F}^m$. Then there exists $g : X_1 \cong X_2$ with $\Phi_2 = \Phi_1 \circ g^*$. By Proposition 2.11 we have $g \in G_x$ and $\Phi = \Phi \circ g^*$; hence $g^* = \mathrm{id}$. By Lemma 3.6 we have $g = \mathrm{id}$. Thus α is injective.

By definition, $\mathcal{F}^m$ is covered by countably many such $\alpha(S)$, and they form a basis of a topology. To show $\mathcal{F}^m$ is a complex manifold, we need to prove that the topology is Hausdorff. Suppose not, then we have two nonseparated points (X, Φ), $(X', \Phi') \in \mathcal{F}^m$. Then X and X' are isomorphic (because $\mathcal{F}$ is separated). Without loss of generality, we assume $X' = X$. Take $\mathscr{X}_S \to S$ the universal family as in Proposition 2.11, and

$$\alpha, \alpha' : S \to \mathcal{F}^m$$

induced by Φ and Φ'. Now since (X, Φ) and (X', Φ') are nonseparated, we have $\alpha(S) \cap \alpha'(S) \neq \varnothing$. Thus there exists $x_1 \in S$ with corresponding cubic fourfold X_1, such that the two pairs (X_1, Φ) and (X_1, Φ') represent the same point in $\mathcal{F}^m$. Then there is an automorphism g of X_1, such that $\Phi' = \Phi \circ g^*$. Proposition 2.11 implies that g is also an automorphism of X and satisfies the above relation. Thus $(X, \Phi) = (X, \Phi')$ in $\mathcal{F}^m$, a contradiction. We showed the Hausdorff property. We conclude that $\mathcal{F}^m$ is naturally a complex manifold. □

Remark 4.3. Proposition 4.2 can be generalized to degree d k-folds ($d \geq 3$, $k \geq 2$) with specified automorphism group. The argument is the same.

When h has signature $(n', 1)$, we define $\mathbb{D}_T = \mathbb{P}\{x \in \Lambda_\zeta \mid \varphi(x, \bar{x}) < 0\}$, which is a hyperbolic complex ball of dimension n'; when h has signature $(n', 2)$, define $\mathbb{D}_T$ to be a component of

$$\mathbb{P}\{x \in (\Lambda_0)_\zeta \mid \varphi(x, x) = 0, \varphi(x, \bar{x}) < 0\},$$

which is a type IV symmetric domain of dimension n'.

We define the local period map for symmetric cubic fourfolds of type T as the map from $\mathcal{F}^m$ to $\mathbb{D}_T \sqcup \overline{\mathbb{D}_T}$, sending (X, Φ) to $\Phi(H^{3,1}(X))$, still denoted by $\widetilde{\mathscr{P}}$. Suppose $\mathbb{D}_T$ is a type IV domain and $\mathcal{F}^m$ is connected, then we make the choice of $\mathbb{D}_T$ such that $\widetilde{\mathscr{P}}$ has image in $\mathbb{D}_T$. Actually, the two situations, $\mathcal{F}^m$ being connected or not, both happen. See Proposition 4.9 for a precise argument.

4B. ***Properties of local period maps for symmetric cubic fourfolds.*** We need to review basic works by Laza [2009; 2010]. In [Laza 2009] stable and semistable cubic fourfolds are classified. One of the main theorems is:

Theorem 4.4 [Laza 2009]. *A cubic fourfold with at worst* ADE*-singularities is stable.*

Independently, Looijenga [2009] and Laza [2010] proved that the period map $\mathscr{P} : \mathcal{M} \to \widehat{\Gamma} \backslash \widehat{\mathbb{D}}$ extends to the moduli space $\mathcal{M}_1$ of cubic fourfolds with at worst ADE singularities, and characterized its image. The results are gathered in the following theorem:

Theorem 4.5 [Laza 2010]. *The period map $\mathscr{P} : \mathcal{M} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ has image $\widehat{\Gamma}\backslash(\widehat{\mathbb{D}} - \mathcal{H}_\infty - \mathcal{H}_\Delta)$, and extends holomorphically to*

$$\mathscr{P} : \mathcal{M}_1 \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$$

with image $\widehat{\Gamma}\backslash(\widehat{\mathbb{D}} - \mathcal{H}_\infty)$. Here $\mathcal{H}_\infty, \mathcal{H}_\Delta$ are two $\widehat{\Gamma}$-invariant hyperplane arrangements in $\widehat{\mathbb{D}}$, with the quotients $\widehat{\Gamma}\backslash\mathcal{H}_\infty$ and $\widehat{\Gamma}\backslash\mathcal{H}_\Delta$ irreducible.

Remark 4.6. This characterization of the image $\mathscr{P}(\mathcal{M})$ was conjectured by Hassett [2000]. Hassett defined the special cubic fourfolds, some of which correspond to polarized $K3$ surfaces. The hyperplane arrangements $\mathcal{H}_\Delta$ and $\mathcal{H}_\infty$ are two particular ones, parametrizing nodal cubic fourfolds and secant lines of the determinantal cubic fourfold, and corresponding to $K3$ surfaces of degree 6 and 2 respectively; see [Hassett 2000, Sections 4.2 and 4.4].

We have also the following marked version of Theorem 4.5:

Proposition 4.7. *The local period map $\widetilde{\mathscr{P}} : \mathcal{M}^m \to \widetilde{\mathbb{D}}$ has image $\widetilde{\mathbb{D}} - \mathcal{H}_\infty - \mathcal{H}_\Delta - \overline{\mathcal{H}_\infty} - \overline{\mathcal{H}_\Delta}$.*

Proof. By Theorem 4.5, the image of $\widetilde{\mathscr{P}}$ lies in $\widetilde{\mathbb{D}} - \mathcal{H}_\infty - \mathcal{H}_\Delta - \overline{\mathcal{H}_\infty} - \overline{\mathcal{H}_\Delta}$. Take any point x in $\widetilde{\mathbb{D}} - \mathcal{H}_\infty - \mathcal{H}_\Delta - \overline{\mathcal{H}_\infty} - \overline{\mathcal{H}_\Delta}$. By Theorem 4.5 the point $[x] \in \widehat{\Gamma}\backslash(\widehat{\mathbb{D}} - \mathcal{H}_\infty - \mathcal{H}_\Delta)$ lies in the image of $\mathscr{P} : \mathcal{M} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$. Thus the orbit $\mathrm{Aut}(\Lambda, \eta)x$ intersects with $\widetilde{\mathscr{P}}(\mathcal{M}^m)$. Notice that the set $\widetilde{\mathscr{P}}(\mathcal{M}^m)$ is $\mathrm{Aut}(\Lambda, \eta)$-invariant; hence contains the orbit $\mathrm{Aut}(\Lambda, \eta)x$. We showed the surjectivity. □

For a specified type T, we write $\mathbb{D} = \mathbb{D}_T$ for short. We have a natural embedding $\mathbb{D} \sqcup \overline{\mathbb{D}} \hookrightarrow \widetilde{\mathbb{D}}$. Denote $\mathcal{H}_s = \mathbb{D} \cap (\mathcal{H}_\Delta \cup \mathcal{H}_\infty \cup \overline{\mathcal{H}_\Delta} \cup \overline{\mathcal{H}_\infty})$ and $\mathcal{H}_* = \mathbb{D} \cap (\mathcal{H}_\infty \cup \overline{\mathcal{H}_\infty})$. The local period map $\widetilde{\mathscr{P}} : \mathcal{F}^m \to \mathbb{D} \sqcup \overline{\mathbb{D}}$ has image contained in $\mathbb{D} \sqcup \overline{\mathbb{D}} - \mathcal{H}_s - \overline{\mathcal{H}_s}$.

Proposition 4.8. *The local period map $\widetilde{\mathscr{P}} : \mathcal{F}^m \to \mathbb{D} \sqcup \overline{\mathbb{D}}$ is an open embedding, with image either $\mathbb{D} - \mathcal{H}_s$ or $\mathbb{D} \sqcup \overline{\mathbb{D}} - \mathcal{H}_s - \overline{\mathcal{H}_s}$. In particular, $n' = n$.*

Proof. We have a closed embedding $\pi : \mathbb{D} \sqcup \overline{\mathbb{D}} \hookrightarrow \widetilde{\mathbb{D}}$. There is a natural map $j : \mathcal{F}^m \to \mathcal{M}^m$. Suppose $(X_1, \Phi_1), (X_2, \Phi_2)$ represent the same point in $\mathcal{M}^m$, then there exists a linear isomorphism $g : X_1 \cong X_2$ such that

$$g^* = \Phi_1^{-1} \circ \Phi_2 : H^4(X_2, \mathbb{Z}) \to H^4(X_1, \mathbb{Z})$$

Since Φ_1, Φ_2 are compatible with the actions of A on $H^4(X_1, \mathbb{Z})$, $H^4(X_2, \mathbb{Z})$, so is g^*. Lemma 3.6 implies that g is compatible with the actions of A on X_1, X_2. Thus $(X_1, \Phi_1), (X_2, \Phi_2)$ represent the same point in $\mathcal{F}^m$. We showed the injectivity of j.

Combining this with the commutative diagram

$$\begin{array}{ccc} \mathcal{F}^m & \xrightarrow{\widetilde{\mathscr{P}}} & \mathbb{D} \sqcup \overline{\mathbb{D}} \\ \downarrow{\scriptstyle j} & & \downarrow{\scriptstyle \pi} \\ \mathcal{M}^m & \xrightarrow{\widetilde{\mathscr{P}}} & \widetilde{\mathbb{D}} \end{array}$$

we obtain the injectivity of $\widetilde{\mathscr{P}} : \mathcal{F}^m \to \mathbb{D} \sqcup \overline{\mathbb{D}}$. In particular, $n \leq n'$.

Since the differential of $\widetilde{\mathscr{P}}: \mathcal{M}^m \to \widetilde{\mathbb{D}}$ is injective everywhere, so is the differential of $\widetilde{\mathscr{P}}: \mathcal{F}^m \to \mathbb{D} \sqcup \overline{\mathbb{D}}$.

Take $(X, \Phi) \in \mathcal{F}^m$. Let $x = \Phi(H^{3,1}(X)) \in \mathbb{D} \sqcup \overline{\mathbb{D}}$ and y be any point in the component of $\mathbb{D} \sqcup \overline{\mathbb{D}}$ containing x. Since both $\mathbb{D} - \mathcal{H}_s$ and $\overline{\mathbb{D}} - \overline{\mathcal{H}_s}$ are connected, there exists a path

$$\gamma : [0, 1] \to \mathbb{D} \sqcup \overline{\mathbb{D}} - \mathcal{H}_s - \overline{\mathcal{H}_s}$$

with $\gamma(0) = x$ and $\gamma(1) = y$. The path γ has a unique lifting in $\mathcal{M}^m$. By Proposition 3.7, we can choose a family $\mathscr{X} \to [0, 1]$ of cubic fourfolds, with marking Φ of every fiber, such that $(\mathscr{X}_0, \Phi) = (X, \Phi)$ and $\Phi(H^{3,1}(\mathscr{X}_s)) = \gamma(s)$, for all $s \in [0, 1]$. Since $\gamma(s) \in \mathbb{D} \sqcup \overline{\mathbb{D}}$, the Hodge structure on $H^4(X_s, \mathbb{Z})$ has an action of A induced by Φ. By Lemma 3.6, there exist actions of A on $\mathscr{X}_s$ for any $s \in [0, 1]$, inducing compatible actions on $H^4(\mathscr{X}_s, \mathbb{Z})$. By Lemma 2.12, actions of A are of the same type T. Thus we obtain a lifting of γ in $\mathcal{F}^m$, hence $y \in \widetilde{\mathscr{P}}(\mathcal{F}^m)$.

If $\widetilde{\mathscr{P}}(\mathcal{F}^m) \subset \mathbb{D}$, then $\widetilde{\mathscr{P}}(\mathcal{F}^m) = \mathbb{D} - \mathcal{H}_s$; otherwise $\widetilde{\mathscr{P}}(\mathcal{F}^m)$ intersects with both $\mathbb{D}$ and $\overline{\mathbb{D}}$, which implies that $\widetilde{\mathscr{P}}(\mathcal{F}^m) = \mathbb{D} \sqcup \overline{\mathbb{D}} - \mathcal{H}_s - \overline{\mathcal{H}_s}$. □

We introduce an involution on $\mathcal{M}^m$. Take any smooth cubic fourfold $X = Z(F)$, and a marking $\Phi : H^4(X, \mathbb{Z}) \to \Lambda$. Let $X' = Z(\bar{F})$. There exists a homeomorphism τ from X to X' given by the complex conjugation. Let ι be the involution on $\mathcal{M}^m$ sending (X, Φ) to $(X', \Phi \circ \tau^*)$. Consider a smooth cubic fourfold $X = Z(F)$ such that F has real coefficients. Then τ is a diffeomorphism of X, and τ^* sends $H^{3,1}(X)$ to $H^{1,3}(X)$. Therefore, choosing any marking Φ of X, the points $[(X, \Phi)]$ and $[(X, \Phi \circ \tau^*)]$ lie in different components of $\mathcal{M}^m$. This implies that the involution ι exchanges the two components of $\mathcal{M}^m$.

Next we give criteria on the number of connected components of $\mathcal{F}^m$. For a symmetry type $T = [(A, \lambda)]$, we define the complex conjugate $\overline{T}$ of T to be $[(\widetilde{A}, \widetilde{\lambda})]$, where $\widetilde{A}$ is the complex conjugate of A, and $\widetilde{\lambda}(a) = \lambda(\bar{a})$ for all $a \in \widetilde{A}$. From the definition, the involution ι exchanges the two spaces $\mathcal{F}_T^m$ and $\mathcal{F}_{\overline{T}}^m$.

Proposition 4.9. *Given a symmetry type $T = [(A, \lambda)]$:*

(i) *If ζ is not real, then $\mathcal{F}^m$ is connected.*

(ii) *If $T = \overline{T}$, then $\mathcal{F}^m$ has two components.*

(iii) *If T satisfies Condition 2.3, and $T \neq \overline{T}$, then $\mathcal{F}^m$ is connected.*

Proof. Suppose ζ is not real, then $\widetilde{\mathscr{P}}(\mathcal{F}^m)$ lies in the ball attached to (Λ_ζ, h). Thus $\mathcal{F}^m$ is connected.

Suppose $T = \overline{T}$, then $\mathcal{F}^m$ is preserved by ι. Thus $\mathcal{F}^m$ has two components.

Suppose $\mathcal{F}^m$ has two components, then $\widetilde{\mathscr{P}}(\mathcal{F}^m) = \mathbb{D} \sqcup \overline{\mathbb{D}} - \mathcal{H}_s - \overline{\mathcal{H}_s}$. Thus $\mathcal{F}^m$ is preserved by ι. Thus $\mathcal{F}_{\overline{T}}^m = \mathcal{F}_T^m$. This can not happen if T satisfies Condition 2.3 and $T \neq \overline{T}$. The third part follows. □

4C. ***Global period map.*** In this section we are going to define the global period domain for symmetric cubic fourfolds of type T as an arithmetic quotient of $\mathbb{D}$, and study the global period map.

Let $(d, k) = (3, 4)$ and fix a symmetry type $T = [(A, \lambda)]$ satisfying Condition 2.2. Let $\Gamma = \{\rho \in \widehat{\Gamma} \mid \rho \bar{A} \rho^{-1} = \bar{A}\}$ be the normalizer of $\bar{A}$ in $\widehat{\Gamma}$. Take $\rho \in \widehat{\Gamma}$ and a point $x \in \Lambda_\zeta$. We claim that $\rho x \in \Lambda_\zeta$. In fact, taking any $a \in A$, we have

$$a\rho x = \rho\rho^{-1} a \rho x = \rho\zeta(\rho^{-1} a \rho)x = \zeta(\rho^{-1} a \rho)\rho x.$$

Since $\rho \in \widehat{\Gamma}$, we have $\rho[x] \in \widehat{\mathbb{D}}$. The two characters ζ and $\rho^{-1}\zeta\rho$ both give nondefinite eigensubspaces of $\Lambda_{\mathbb{C}}$. We conclude that $\zeta = \rho^{-1}\zeta\rho$; hence $\rho x \in \Lambda_\zeta$. This gives a natural action of Γ on $\mathbb{D}$.

Let N_A be the normalizer of A in $\mathrm{Aut}((\Lambda_0)_{\mathbb{Q}}, \varphi)$, which is a reductive algebraic subgroup. The group Γ is an arithmetic subgroup of N_A; see the Appendix. The arithmetic quotient $\Gamma\backslash\mathbb{D}$ is a quasiprojective variety thanks to the Baily–Borel compactification (see Section A3 in the Appendix). From our assumption that the local period map $\widetilde{\mathscr{P}}$ for $\mathcal{F}^m$ takes values in $\mathbb{D}$, we can take $(\mathcal{F}^m)^1$ to be the connected component of $\mathcal{F}^m$ such that $\widetilde{\mathscr{P}}((\mathcal{F}^m)^1) = \mathbb{D} - \mathcal{H}_s$. Notice that when $\mathcal{F}^m$ is connected, we have $(\mathcal{F}^m)^1 = \mathcal{F}^m$.

Proposition 4.10. *The local period map $\widetilde{\mathscr{P}} : (\mathcal{F}^m)^1 \to \mathbb{D} - \mathcal{H}_s$ descends to an algebraic isomorphism $\mathscr{P} : \mathcal{F} \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$.*

Proof. There are natural analytic morphisms from $\mathcal{F}^m$ to $\mathcal{F}$, and $\mathbb{D} - \mathcal{H}_s$ to $\Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$. We define the global period map $\mathscr{P} : \mathcal{F} \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$ as follows. Take $F \in \mathcal{V}^{sm}$. We choose a T-marking Φ of $X = Z(F)$, such that $\Phi(H^{3,1}(X)) \in \mathbb{D}$ (this also means that $(F, \Phi) \in (\mathcal{F}^m)^1$). We define

$$\mathscr{P}([F]) = [\widetilde{\mathscr{P}}(X, \Phi)].$$

We show this map is well-defined. Take $F_1, F_2 \in \mathcal{V}^{sm}$ with T-markings Φ_1, Φ_2 respectively. Suppose there exists $g \in N$, such that $g(F_1) = F_2$. We have an induced map

$$g^* : H^4(Z(F_2), \mathbb{Z}) \to H^4(Z(F_1), \mathbb{Z}).$$

Next we show $\rho = \Phi_1 g^* \Phi_2^{-1} \in \Gamma$. Denote $a' = gag^{-1}$. Since $g \in N$, we have $a' \in A$. We have the following commutative diagram:

$$\begin{array}{ccccccc}
\Lambda & \xrightarrow{\Phi_2^{-1}} & H^4(Z(F_2), \mathbb{Z}) & \xrightarrow{g^*} & H^4(Z(F_1), \mathbb{Z}) & \xrightarrow{\Phi_1} & \Lambda \\
\downarrow{\scriptstyle a'} & & \downarrow{\scriptstyle a'^*} & & \downarrow{\scriptstyle a^*} & & \downarrow{\scriptstyle a} \\
\Lambda & \xrightarrow{\Phi_2^{-1}} & H^4(Z(F_2), \mathbb{Z}) & \xrightarrow{g^*} & H^4(Z(F_1), \mathbb{Z}) & \xrightarrow{\Phi_1} & \Lambda
\end{array}$$

This implies that, as automorphisms of Λ, $a' = \rho^{-1} a \rho$. Thus $\rho \in \Gamma$. We then have a well-defined analytic morphism $\mathscr{P} : \mathcal{F} \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$.

By definition we have the following commutative diagram:

$$\begin{array}{ccc}
(\mathcal{F}^m)^1 & \xrightarrow{\widetilde{\mathscr{P}}} & \mathbb{D} - \mathcal{H}_s \\
\downarrow{\scriptstyle j} & & \downarrow{\scriptstyle \pi} \\
\mathcal{F} & \xrightarrow{\mathscr{P}} & \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)
\end{array} \tag{3}$$

We next show that $\mathscr{P} : \mathcal{F} \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$ is an isomorphism.

We first show the injectivity. Suppose that $(F_1, \Phi_1), (F_2, \Phi_2) \in \mathcal{F}^m$, with $\Phi_1(H^{3,1}(Z(F_1)))$ and $\Phi_2(H^{3,1}(Z(F_2)))$ representing the same point in $\Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$. Then there exists $\rho \in \Gamma$, such that $\rho\Phi_1(H^{3,1}(Z(F_1))) = \Phi_2(H^{3,1}(Z(F_2)))$. The map

$$\Phi_2^{-1}\rho\Phi_1 : H^4(Z(F_1), \mathbb{Z}) \to H^4(Z(F_2), \mathbb{Z})$$

preserves the polarized Hodge structures. By Lemma 3.6, we have $g \in \mathrm{SL}(V)$, with gF_2 equals to F_1 after rescaling of F_2, and $g^* = \Phi_2^{-1}\rho\Phi_1$. For any $a \in A$, we have $a^* : H^4(Z(F_1), \mathbb{Z}) \to H^4(Z(F_1), \mathbb{Z})$. The $g^{-1}ag$ acts on $Z(F_2)$, and this induces

$$(g^{-1}ag)^* = g^*a^*g^{*-1} = (\Phi_2^{-1}\rho\Phi_1)(\Phi_1^{-1}a\Phi_1)(\Phi_1^{-1}\rho^{-1}\Phi_2) = \Phi_2^{-1}\rho a\rho^{-1}\Phi_2.$$

Since $\rho \in \Gamma$, we have $\rho a\rho^{-1} \in A$. Again by Lemma 3.6, we have $g^{-1}ag \in A$. Since

$$g^{-1}agF_2 = g^{-1}aF_1 = \lambda(a)g^{-1}F_1 = \lambda(a)F_2,$$

we have $\lambda(g^{-1}ag) = \lambda(a)$. We conclude $g \in N$. Thus $\mathscr{P}$ is injective.

By Proposition 4.8, the composition of

$$(\mathcal{F}^m)^1 \to \mathbb{D} - \mathcal{H}_s \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$$

is surjective. By commutativity of diagram (3), the composition of

$$(\mathcal{F}^m)^1 \to \mathcal{F} \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$$

is also surjective; hence $\mathscr{P} : \mathcal{F} \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$ is surjective.

The algebraicity of $\mathscr{P}$ can be deduced from its extension to certain compactifications on both sides; see Theorem 5.3. An alternative argument follows the proof of Proposition 2.2.3 in [Hassett 2000] using Baily–Borel compactification and the Borel extension theorem. □

5. Compactifications

In this section we are going to study the compactifications of both two sides of $\mathscr{P} : \mathcal{F} \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$. The essential ingredient is the identification between the GIT compactification of the moduli space of cubic fourfolds and the Looijenga compactification of the global period domain, proved by Looijenga [2009] and Laza [2010] independently. Depending on this, we will prove Theorem 1.2(i), and then deduce Theorem 1.1(iii). In Theorem 5.7 (=Theorem 1.2(ii)), we give a criterion when the Looijenga compactification is actually Baily–Borel compactification.

Let $(d, k) = (3, 4)$. Recall that from Theorem 4.5 we have the isomorphism $\mathscr{P} : \mathcal{M}_1 \cong \widehat{\Gamma}\backslash(\widehat{\mathbb{D}} - \mathcal{H}_\infty)$. From [Looijenga 2009; Laza 2010] we have:

Theorem 5.1 (Looijenga, Laza). *The period map $\mathscr{P}$ extends to an isomorphism $\mathscr{P} : \overline{\mathcal{M}} \to \overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\mathcal{H}_\infty}$.*

Recall that $\mathcal{H}_* = \mathbb{D} \cap (\mathcal{H}_\infty \cup \overline{\mathcal{H}}_\infty)$, which is a Γ-invariant hyperplane arrangement in $\mathbb{D}$. We have a morphism between locally symmetric varieties

$$\Gamma\backslash\mathbb{D} \to \mathrm{Aut}(\Lambda, \eta)\backslash\widetilde{\mathbb{D}} \cong \widehat{\Gamma}\backslash\widehat{\mathbb{D}}.$$

We can construct the Looijenga compactification $\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$ of $\Gamma\backslash(\mathbb{D} - \mathcal{H}_*)$ (see the Appendix). From Theorem A.13, we have:

Proposition 5.2. *There exists a finite morphism $\pi : \overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*} \to \overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\mathcal{H}_\infty}$. If T satisfies Condition 2.3, then this morphism is a normalization of its image.*

We now state our main theorem:

Theorem 5.3. *The global period map $\mathscr{P} : \mathcal{F} \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$ extends to an algebraic isomorphism $\mathscr{P} : \overline{\mathcal{F}} \cong \overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$.*

We need the following fact in algebraic geometry. We give the proof for the reader's convenience.

Lemma 5.4. *Let $f_1 : Z_1 \to Y$ and $f_2 : Z_2 \to Y$ be finite morphisms between irreducible algebraic varieties. Suppose Z_1, Z_2 are normal. Moreover, suppose that there exists Zariski-open subset U_i of Z_i, $i = 1$ or 2, with a biholomorphic map $g : U_1 \to U_2$, such that $f_1 = f_2 \circ g$. Then g extends to an algebraic isomorphism $Z_1 \to Z_2$.*

Proof. Without loss of generality, we assume that Y is affine. Let $\mathbb{C}(Z)$ be the field of rational functions on an irreducible algebraic variety Z, and $M(Z)$ the field of meromorphic functions. We claim $g^*\mathbb{C}(Z_2) = \mathbb{C}(Z_1)$. Let $x \in \mathbb{C}(U_2) = \mathbb{C}(Z_2)$. Since $\mathbb{C}(U_2)$ is a finite extension of $\mathbb{C}(Y)$, g^*x is finite over $\mathbb{C}(U_1)$. We can find a Zariski-open subset U_1° of U_1, with a Galois covering $\widetilde{U} \to U_1^\circ$, such that $g^*x \in \mathbb{C}(\widetilde{U})$. Since $g^*x \in M(U_1^\circ)$, it is invariant under the action of Deck transformations. Thus $g^*x \in \mathbb{C}(U_1^\circ) = \mathbb{C}(Z_1)$. The claim follows.

The coordinate ring $\mathbb{C}[Z_i]$ is the integral closure of $\mathbb{C}[Y]$ in $\mathbb{C}(Z_i)$. So $g^*\mathbb{C}[Z_2] = \mathbb{C}[Z_1]$. Thus g extends to an algebraic isomorphism $Z_1 \cong Z_2$. □

Proof of Theorem 5.3. We have the commutative diagram

$$\begin{array}{ccc} \mathcal{F} & \xrightarrow{\cong} & \Gamma\backslash(\mathbb{D} - \mathcal{H}_s) \\ \downarrow & & \downarrow \\ \overline{\mathcal{F}} & \dashrightarrow & \overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*} \\ \downarrow{\scriptstyle j} & & \downarrow{\scriptstyle \pi} \\ \overline{\mathcal{M}} & \xrightarrow{\mathscr{P}} & \overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\mathcal{H}_\infty} \end{array} \tag{4}$$

with both j, π finite morphisms. The commutativity is straightforward from the definitions of the maps. Since $\mathcal{F}$ is Zariski-open in $\overline{\mathcal{F}}$, the image $j(\mathcal{F})$ contains a Zariski-open subset of $j(\overline{\mathcal{F}})$. Thus $j(\overline{\mathcal{F}})$ is the closure of $j(\mathcal{F})$ in $\overline{\mathcal{M}}$. The same argument shows that $\pi(\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*})$ is the closure of $\pi(\Gamma\backslash(\mathbb{D} - \mathcal{H}_s))$ in $\overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\mathcal{H}_\infty}$. By commutativity of diagram (4), the two images $j(\mathcal{F})$ and $\pi(\Gamma\backslash(\mathbb{D} - \mathcal{H}_s))$ are identified via $\mathscr{P}$, so are $j(\overline{\mathcal{F}})$ and $\pi(\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*})$. By Propositions 2.7, 5.2 and Lemma 5.4, we have an identification between $\overline{\mathcal{F}}$ and $\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$ which extends $\mathscr{P} : \mathcal{F} \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$. This identification is the extended global period map $\mathscr{P} : \overline{\mathcal{F}} \cong \overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$. □

The proof of the above theorem does not use algebraicity of $\mathscr{P}$. Actually, we can deduce algebraicity of $\mathscr{P}$ from Theorem 5.3. At this point, we have already finished the proof of Theorem 1.1(i), (ii) and Theorem 1.2(i). In the rest of this section, we prove Theorem 1.1(iii) and Theorem 1.2(ii).

Let $\mathcal{V}_1$ be the subset of $\mathcal{V}$ consisting of cubic forms of type T defining cubic fourfolds with at worst ADE-singularities. The points in $\mathcal{V}_1$ are stable with respect to the action of $\mathrm{SL}(V)$ on $\mathrm{Sym}^3(V^*)$; hence also stable with respect to the action of N on $\mathcal{V}$. Define $\mathcal{F}_1 = N\backslash\backslash\mathbb{P}\mathcal{V}_1$ to be the moduli space of cubic fourfolds of type T with at worst ADE-singularities. We have:

Proposition 5.5. *The period map $\mathscr{P} : \mathcal{F} \to \Gamma\backslash(\mathbb{D} - \mathcal{H}_s)$ extends to an algebraic isomorphism $\mathscr{P} : \mathcal{F}_1 \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_*)$.*

Proof. From the definition we have $j(\mathcal{F}_1) = j(\overline{\mathcal{F}}) \cap \mathcal{M}_1$ and $j^{-1}(j(\mathcal{F}_1)) = \mathcal{F}_1$. From Proposition 2.7, the morphism $j : \mathcal{F}_1 \to \mathcal{M}_1$ is finite. On the other hand, we have

$$\pi(\Gamma\backslash(\mathbb{D} - \mathcal{H}_*)) = \pi(\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}) \cap \widehat{\Gamma}\backslash(\widehat{\mathbb{D}} - \mathcal{H}_\infty)$$

and

$$\pi^{-1}(\pi(\Gamma\backslash(\mathbb{D} - \mathcal{H}_*))) = \Gamma\backslash(\mathbb{D} - \mathcal{H}_*).$$

From Proposition 5.2, the morphism $\pi : \Gamma\backslash(\mathbb{D} - \mathcal{H}_*) \to \widehat{\Gamma}\backslash(\widehat{\mathbb{D}} - \mathcal{H}_\infty)$ is finite. By Theorems 4.5 and 5.3, the two images $j(\mathcal{F}_1)$ and $\pi(\Gamma\backslash(\mathbb{D} - \mathcal{H}_*))$ are identified via $\mathscr{P}$. By Lemma 5.4, we have the algebraic isomorphism $\mathscr{P} : \mathcal{F}_1 \cong \Gamma\backslash(\mathbb{D} - \mathcal{H}_*)$. □

If the hyperplane arrangement $\mathcal{H}_*$ is empty, then the Looijenga compactification of $\Gamma\backslash\mathbb{D}$ is actually the Baily–Borel compactification. Next we give a criterion of emptiness of $\mathcal{H}_*$ from the perspective of GIT. Following Section 6 of [Laza 2009], there is a rational curve χ parametrizing certain semistable cubic fourfolds, given by

$$F_{a,b}(x_0, \ldots, x_5) = \begin{vmatrix} x_0 & x_1 & x_2 + 2ax_5 \\ x_1 & x_2 - ax_5 & x_3 \\ x_2 + 2ax_5 & x_3 & x_4 \end{vmatrix} + bx_5^3$$

where $(a : b) \in \mathrm{WP}(1 : 3)$ with $\mathrm{WP}(1 : 3)$ the weighted projective space of weight $(1 : 3)$. We denote by $X_{(a:b)}$ the cubic fourfold defined by $F_{a,b}$. Denote

$$F_0(x_0, \ldots, x_4) = \begin{vmatrix} x_0 & x_1 & x_2 \\ x_1 & x_2 & x_3 \\ x_2 & x_3 & x_4 \end{vmatrix},$$

then $F_{a,b}(x_0, \ldots, x_5) = F_0(x_0, \ldots, x_4) + ax_5(4x_1x_3 - 3x_2^2 - x_0x_4) + (b - 4a^3)x_5^3$.

We next define two pairs (G_1, λ_1) and (G_2, λ_2) which will be used in Theorem 5.7. Here for $i = 1$ or 2, G_i is a subgroup of $\mathrm{SL}(V)$ and $\lambda_i : G_i \to \mathbb{C}^\times$ is a character of G_i. As we will discuss below, the pairs (G_1, λ_1) and (G_2, λ_2) are essentially symmetries for $F_{1,0}$ and $F_{0,1}$ respectively.

The cubic fourfold $X_{(1:0)}$ is called the determinantal cubic fourfold. The singular locus of $X_{(1:0)}$ is a rational surface. Explicitly, take V_3 to be a complex vector space of dimension 3 and denote by

$[y_0 : y_1 : y_2]$ a homogeneous coordinate for $\mathbb{P}V_3$. Consider an embedding $\mathbb{P}V_3 \hookrightarrow \mathbb{P}V$ defined by $[y_0 : y_1 : y_2] \mapsto [x_0 : \cdots : x_5]$ with $x_0 = y_0^2,\ x_1 = y_0y_1,\ x_2 - x_5 = y_1^2,\ x_0 + 2x_5 = y_0y_2,\ x_3 = y_1y_2,\ x_4 = y_2^2$. This induces a natural morphism from $\mathrm{GL}(V_3)$ to $\mathrm{GL}(V)$. The image of $\mathbb{P}V_3$ in $\mathbb{P}V$ is called the Veronese surface, and it is the singular locus of $X_{(1:0)}$. Actually the singular cubic fourfold $X_{(1:0)}$ is the secant variety of the Veronese surface in $\mathbb{P}^5$, and the linear automorphism group of $X_{(1:0)}$ can be identified with $\mathrm{PSL}(V_3)$. For each $g \in \mathrm{GL}(V_3)$ there is a complex number $\lambda_1(g)$ such that $gF_{1,0} = \lambda_1(g)F_{1,0}$. We hence obtain a character λ_1 of $\mathrm{GL}(V_3)$. By standard theory on general linear group, there exists an integer k such that $\lambda_1(g) = \det(g)^k$ for any $g \in \mathrm{GL}(V_3)$. To know k, we only need to compute $\lambda_1(g)$ for a special g. Take $g : (y_0, y_1, y_2) \mapsto (ty_0, y_1, y_2)$. Then

$$g : (x_0, x_1, x_2 + 2x_5, x_2 - x_5, x_3, x_4) \mapsto (t^2x_0, tx_1, t(x_2 + 2x_5), x_2 - x_5, x_3, x_4).$$

Thus $gF_{1,0} = t^2F_{1,0}$. This implies that $\lambda_1(g) = t^2 = \det(g)^2$ and we have $k = 2$. In conclusion, for any $g \in \mathrm{GL}(V_3)$ we have $\lambda_1(g) = \det(g)^2$.

For $b \neq 0$, the singular locus of the cubic fourfold $X_{(a:b)}$ is a rational curve. Explicitly, take V_2 to be a complex vector space of dimension 2 and denote by $[y_0 : y_1]$ a homogeneous coordinate for $\mathbb{P}V_2$. Let V_5 be the subspace of V defined by $x_5 = 0$. Consider an embedding $\mathbb{P}V_2 \hookrightarrow \mathbb{P}V_5$ defined by $[y_0 : y_1] \mapsto [x_0 : \cdots : x_4]$ with $x_i = y_0^{4-i}y_i$ for $i = 0, 1, 2, 3, 4$. This also induces a natural morphism from $\mathrm{GL}(V_2)$ to $\mathrm{GL}(V)$. Then the singular locus of $X_{(a:b)}$ is the image of $\mathbb{P}V_2 \hookrightarrow \mathbb{P}V_5 \hookrightarrow \mathbb{P}V$. By [Laza 2009, Proposition 6.6 and its proof] the linear automorphism group of $X_{(a:b)}$ for a generic choice $(a : b) \in \mathrm{WP}(1 : 3)$ is $\mathrm{PSL}(V_2)$. For any $g \in \mathrm{GL}(V_2)$, there exists a complex number $\lambda_2(g)$ such that $gF_0 = \lambda_0(g)F_0$. A similar calculation as before gives $\lambda_0(g) = \det(g)^6$.

When $(a, b) = (0, 1)$, we have extra automorphisms of $X_{(a:b)}$ given by taking scalars on x_5. Suppose $(g, u) \in \mathrm{GL}(V_5) \times \mathbb{C}^\times$ is an automorphism of $F_{0,1} = F_0 + x_5^3$. Since $gF_0 = \det(g)^6F_0$ and $u(x_0^3) = u^3x_0^3$, we must have $\det(g)^6 = u^3$. Thus $\det(g)^2/u$ is a third root of unity. The following definition is then natural:

Definition 5.6. (i) Let G_1 be the intersection of $\mathrm{SL}(V)$ with the image of $\mathrm{GL}(V_3) \to \mathrm{GL}(V)$.

(ii) Let $\widetilde{G_2}$ be the subgroup of $\mathrm{GL}(V_2) \times \mathbb{C}^*$ consisting of elements (g, u) such that $(\det g)^2/u$ is a third root of unity. Let G_2 be the intersection of $\mathrm{SL}(V)$ with the image of the natural map $\widetilde{G_2} \to \mathrm{GL}(V)$.

Both G_1 and G_2 contain the center of $\mathrm{SL}(V)$. The restriction of λ_1 to G_1 is still denoted by λ_1. For G_2, we have a character $\lambda_2 : (g, u) \mapsto \lambda_0(g) = \det(g)^6 = u^3$. The next theorem gives a criterion on emptiness of $\mathcal{H}_*$. We will apply this criterion to prime-order groups (Proposition 6.5).

Theorem 5.7. *For a symmetry type (A, λ) satisfying Condition 2.2, the following three statements are equivalent*:

(i) *The hyperplane arrangement $\mathcal{H}_*$ is nonempty.*

(ii) *The space $\mathbb{P}\mathcal{V}_\lambda$ intersects with the orbit $\mathrm{PSL}(V)\chi$ of the rational curve χ in $\mathbb{P}\mathrm{Sym}^3(V^*)$.*

(iii) *For $i = 1$ or 2, there exists $h \in \mathrm{SL}(V)$ such that $h^{-1}Ah \subset G_i$ and for any $a \in A$ we have $\lambda(a) = \lambda_i(h^{-1}ah)$. If this is satisfied, we say that (A, λ) factors through (G_i, λ_i).*

Proof. We first show the equivalence of (i) and (ii). If (ii) holds, the intersection points survive after taking GIT quotients since the $\mathrm{PSL}(V)$ orbits of points in χ are closed. Conversely, suppose $j(\overline{\mathcal{F}})$ intersects with the image of χ at $[F]$ in $\overline{\mathcal{M}}$. We can always take the representative F in $\mathcal{V}_\lambda$ has closed N-orbit. According to the main theorem in [Luna 1975], the $\mathrm{PSL}(V)$-orbit of $[F] \in \mathbb{P}\mathrm{Sym}^3(V^*)$ is also closed. So F represents an element in $\mathrm{PSL}(V)\chi$.

Secondly we recall that the blow-up and blow-down construction in Looijenga compactification $\overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\mathcal{H}_\infty}$ gives a stratum corresponding to χ. We claim that $\mathcal{H}_*$ is nonempty if and only if the image of $\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$ in $\overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\mathcal{H}_\infty}$ intersects with the stratum. From the proof of functoriality of semitoric compactification in Section A4, we know that $\mathbb{D}^\Sigma$ intersects with $\overline{\mathcal{H}}_\infty$ if and only if $\mathbb{D}$ intersects with $\mathcal{H}_\infty$. So the image of $\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$ intersects with the stratum if and only if $\mathbb{D}$ intersects with $\mathcal{H}_\infty$. By diagram (4), the intersection of $j(\overline{\mathcal{F}})$ with the image of χ in $\overline{\mathcal{M}}$ is equivalent to the intersection of the image of $\pi(\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*})$ with the stratum corresponding to χ. The equivalence of (i) and (ii) follows.

Next we show the equivalence of (ii) and (iii). Suppose (iii) is satisfied, then for $i = 1$ or 2, there exists $h \in \mathrm{GL}(V)$ such that $h^{-1}Ah \subset G_i$ and $\lambda(a) = \lambda_i(h^{-1}ah)$ for any $a \in A$. Then $hF_{1,0}$ or $hF_{0,1}$ lies in $\mathcal{V}_\lambda$. This implies (ii).

Suppose (ii) holds. Then there is a member $F_{a,b}$ in χ, and an element $h \in \mathrm{GL}(V)$, such that $hF_{a,b} \in \mathcal{V}_\lambda$. If $b = 0$ or $a = 0$, then (A, λ) factors through (G_1, λ_1) or (G_2, λ_2). Otherwise, we claim that the linear automorphism group of $X_{(a:b)}$ is indeed $\mathrm{PSL}(V_2)$; hence (A, λ) factors through both (G_1, λ_1) and (G_2, λ_2). Let $g \in \mathrm{GL}(V)$ be an automorphism of $F_{a,b}$. Since the singular locus of $F_{a,b}$ is the image $\mathbb{P}V_2 \hookrightarrow \mathbb{P}V$, the automorphism g fixes $\mathbb{P}V_2$ which is the smallest subspace of $\mathbb{P}V$ containing the singular locus. Moreover, g is induced by an element of $\mathrm{GL}(V_2)$. Since $F_{a,b}(x_0, \ldots, x_5)$ contains no monomial with the degree of x_5 equal to 2, the action of g on the coordinate x_5 is by a scalar. By $a \neq 0$ we conclude that g fixes x_5. Therefore, the linear automorphism group of $X_{(a:b)}$ is $\mathrm{PSL}(V_2)$. The discussion above shows that (ii) and (iii) are equivalent. □

6. Examples and related constructions

In this section we apply our theorems to specific examples. We will first review the classification of prime-order automorphisms of smooth cubic fourfolds [González-Aguilera and Liendo 2011, Theorem 3.8] in Section 6A, then in Section 6B we will show how our results recover a main theorem in [Laza et al. 2018].

6A. ***Prime-order automorphisms of smooth cubic fourfolds.*** The classification of prime-order automorphisms of smooth cubic fourfolds was given in [González-Aguilera and Liendo 2011, Theorem 3.8]. For the reader's convenience we present the result in this section. (There was a small mistake in [González-Aguilera and Liendo 2011, Theorem 3.8]. The second example with $p = 5$ contains only singular cubic fourfolds. This is pointed out in [Boissière et al. 2016, Remark 6.3]).

Proposition 6.1 [González-Aguilera and Liendo 2011]. *Let ω be a prime p-th root of unity and $\rho = (m_0, \ldots, m_5)$ be the automorphism of $V \cong \mathbb{C}^6$ given by $(x_0, \ldots, x_5) \mapsto (\omega^{m_0}x_0, \ldots, \omega^{m_5}x_5)$. The list of*

smooth cubic polynomials F preserved by the action under ρ is as follows:

$$T_2^1 : \rho = (0,0,0,0,0,1), \quad n = 14, \quad F = L_3(x_0,\dots,x_4) + x_5^2 L_1(x_0,\dots,x_4).$$

$$T_2^2 : \rho = (0,0,0,0,1,1), \quad n = 12,$$
$$F = L_3(x_0,\dots,x_3) + x_4^2 L_1(x_0,\dots,x_3) + x_4x_5M_1(x_0,\dots,x_3) + x_5^2 N_1(x_0,\dots,x_3).$$

$$T_2^3 : \rho = (0,0,0,1,1,1), \quad n = 10,$$
$$F = L_3(x_0,x_1,x_2) + x_0L_2(x_3,x_4,x_5) + x_1M_2(x_3,x_4,x_5) + x_2N_2(x_3,x_4,x_5).$$

$$T_3^1 : \rho = (0,0,0,0,0,1), \quad n = 10, \quad F = L_3(x_0,\dots,x_4) + x_5^3.$$

$$T_3^2 : \rho = (0,0,0,0,1,1), \quad n = 4, \quad F = L_3(x_0,\dots,x_3) + M_3(x_4,x_5).$$

$$T_3^3 : \rho = (0,0,0,0,1,2), \quad n = 8, \quad F = L_3(x_0,\dots,x_3) + x_4^3 + x_5^3 + x_4x_5M_1(x_0,\dots,x_3).$$

$$T_3^4 : \rho = (0,0,0,1,1,1), \quad n = 2, \quad F = L_3(x_0,x_1,x_2) + M_3(x_3,x_4,x_5).$$

$$T_3^5 : \rho = (0,0,0,1,1,2), \quad n = 7,$$
$$F = L_3(x_0,x_1,x_2) + M_3(x_3,x_4) + x_5^3 + x_3x_5L_1(x_0,x_1,x_2) + x_4x_5M_1(x_0,x_1,x_2).$$

$$T_3^6 : \rho = (0,0,1,1,2,2), \quad n = 8,$$
$$F = L_3(x_0,x_1) + M_3(x_2,x_3) + N_3(x_4,x_5) + \sum_{i=1,2;\, j=3,4;\, k=5,6} a_{ijk}x_ix_jx_k.$$

$$T_3^7 : \rho = (0,0,1,1,2,2), \quad n = 6,$$
$$F = x_2L_2(x_0,x_1) + x_3M_2(x_0,x_1) + x_4^2L_1(x_0,x_1)$$
$$+ x_4x_5M_1(x_0,x_1) + x_5^2N_1(x_0,x_1) + x_4N_2(x_2,x_3) + x_5O_2(x_2,x_3).$$

$$T_5^1 : \rho = (0,0,1,2,3,4), \quad n = 4,$$
$$F = L_3(x_0,x_1) + x_2x_5L_1(x_0,x_1) + x_3x_4M_1(x_0,x_1) + x_2^2x_4 + x_2x_3^2 + x_3x_5^2 + x_4^2x_5.$$

$$T_7^1 : \rho = (1,2,3,4,5,6), \quad n = 2, \quad F = x_0^2x_4 + x_1^2x_2 + x_0x_2^2 + x_3^2x_5 + x_3x_4^2 + x_1x_5^2 + ax_0x_1x_3 + bx_2x_4x_5$$

$$T_{11}^1 : \rho = (0,1,3,4,5,9), \quad n = 0, \quad F = x_0^3 + x_1^2x_5 + x_2^2x_4 + x_2x_3^2 + x_1x_4^2 + x_3x_5^2.$$

Here the lower index is the prime p, the polynomials L_i, M_i, N_i are of degree i, and n is the dimension of the corresponding GIT*-quotient.*

Remark 6.2. This classification offers 13 symmetry types with $\#\bar{A}$ a prime number 2, 3, 5, 7 or 11. Those symmetry types may not satisfy Condition 2.3. See previous discussion in Remark 2.4.

By Griffiths residue calculus [1969a; 1969b], for a smooth cubic fourfold $X = Z(F)$, the complex line $H^{3,1}(X)$ is generated by $\mathrm{Res}_X(\Omega/F^2)$. Here $\Omega = \Sigma_{i=0}^5 (-1)^i x_i dx_1 \wedge \cdots \wedge \widehat{dx_i} \wedge \cdots \wedge dx_5$. By direct calculation, we have:

Proposition 6.3. (i) *For type $T = T_2^2, T_3^3, T_3^4, T_3^6, T_5^1, T_7^1, T_{11}^1$, we have $\zeta = 1$.*

(ii) *For type $T = T_2^1, T_2^3$, we have $\zeta = -1$*

(iii) *For type $T = T_3^1, T_3^2, T_3^5, T_3^7$, we have that $\zeta(\rho)$ is equal to ω or $\overline{\omega}$.*

We already proved that $\mathscr{P}(\mathcal{F}^m)$ is either $\mathbb{D} - \mathcal{H}_s$ or $\mathbb{D} \sqcup \overline{\mathbb{D}} - \mathcal{H}_s - \overline{\mathcal{H}_s}$. From Proposition 4.9, we have:

Proposition 6.4. (i) *If $T = T_3^1, T_3^2, T_3^5, T_3^7$, then $\mathbb{D}$ is a complex hyperbolic ball and $\widetilde{\mathscr{P}}(\mathcal{F}^m) = \mathbb{D} - \mathcal{H}_s$.*

(ii) *If $T = T_2^1, T_2^2, T_2^3, T_3^3, T_3^4, T_3^6$ or T_7^1, then $\mathbb{D}$ is a type* IV *domain and $\widetilde{\mathscr{P}}(\mathcal{F}^m) = \mathbb{D} \sqcup \overline{\mathbb{D}} - \mathcal{H}_s - \overline{\mathcal{H}_s}$.*

Now we apply Theorem 5.7 for prime-order cases.

Proposition 6.5. *For $T = T_2^1, T_2^3, T_3^2, T_3^3, T_3^4, T_3^7, T_{11}^1$, we obtain isomorphisms between* GIT *compactifications $\overline{\mathcal{F}}$ with Baily–Borel compactifications $\overline{\Gamma\backslash\mathbb{D}}^{bb}$. For $T = T_2^2, T_3^1, T_3^5, T_3^6, T_5^1, T_7^1$, the corresponding Looijenga compactifications are not Baily–Borel compactifications.*

Proof. We do the calculation for $p = 2$ and 3; the other cases are similar. Suppose $p = 2$. If (A, λ) factors through (G_1, λ_1), a generator of $\bar{A}$ corresponds (up to conjugate) to $g = \mathrm{diag}(1, 1, -1) \in \mathrm{GL}(V_3)$ with order 2. The image of g in $\mathrm{GL}(V)$ is $\mathrm{diag}(1, 1, 1, 1, -1, -1)$. If (A, λ) factors through (G_2, λ_2), then we take $(g, u) \in \mathrm{GL}(V_2) \times \mathbb{C}^*$ such that $g = \mathrm{diag}(1, -1)$ and u is a third root of unity. The image of (g, u) in $\mathrm{GL}(V)$ is $\mathrm{diag}(1, 1, 1, 1, -1, -1)$. In both two cases, we obtain $\mathrm{diag}(1, 1, 1, 1, -1, -1) \in \mathrm{SL}(V)$ and the values of both λ_1 and λ_2 are equal to 1. By Theorem 5.7, the symmetry type T_2^2 does not give Baily–Borel compactification and T_2^1, T_2^3 give Baily–Borel compactifications.

Suppose $p = 3$. If (A, λ) factors through (G_1, λ_1), then a generator of $\bar{A}$ corresponds to $g = \mathrm{diag}(1, 1, \omega) \in \mathrm{GL}(V_3)$ with order 3. The image of g in $\mathrm{GL}(V)$ is $\mathrm{diag}(1, 1, 1, \omega, \omega, \omega^2)$, with value of λ_1 equal to $\det(g)^2 = \omega^2$. If (A, λ) factors through (G_2, λ_2), then we take $(g, u) \in \mathrm{GL}(V_2) \times \mathbb{C}^\times$ with $g = \mathrm{diag}(1, \omega)$ or $\mathrm{diag}(1, 1)$, and u a third root of unity. The image of (g, u) in $\mathrm{GL}(V)$ is $\mathrm{diag}(1, 1, \omega, \omega, \omega^2, u)$ or $\mathrm{diag}(1, 1, 1, 1, 1, u)$. For these elements, the values of λ_2 equal to $u^3 = 1$. We conclude that for T_3^1, T_3^5, T_3^6 we do not obtain Baily–Borel compactification, and for other T_3^i we do. □

Remark 6.6. Notice that Proposition 6.5 is compatible with results in previous literature. For $T = T_2^1$, we have $\mathcal{H}_*$ is empty and we obtain Baily–Borel compactification. This is proved in [Laza et al. 2018] via a lattice-theoretic argument. For $T = T_3^1$, the arrangement $\mathcal{H}_*$ is not empty and we do not obtain Baily–Borel compactification. This coincides with the work in [Looijenga and Swierstra 2007; Allcock et al. 2011].

Remark 6.7. Notice that for the symmetry type T_7^1, the hyperplane arrangement $\mathcal{H}_*$ is nonempty and the dimension of each member is 1. This is one of the examples in which the approach adopted in previous works does not apply; see the discussion right after Theorem 1.2.

6B. ***Examples revisit.*** Take $T = T_3^1$. Then $T = [(\bar{A} = \mu_3, \lambda = 1)]$ satisfies Condition 2.3. The space $\mathcal{F}$ can be identified with the moduli space of smooth cubic threefolds. The local period domain $\mathbb{D}$ is a complex hyperbolic ball of dimension 10 with an action of an arithmetic group Γ. Then Theorems 1.1 and 1.2 recover the main results in [Looijenga and Swierstra 2007; Allcock et al. 2011]. By Proposition 6.5, the hyperplane arrangement $\mathcal{H}_*$ is nonempty. Actually, from [Looijenga and Swierstra 2007; Allcock et al. 2011], the quotients $\Gamma\backslash\mathcal{H}_s$ has two irreducible components, and $\Gamma\backslash\mathcal{H}_*$ is irreducible.

Take $T = T_2^1$. Then $T = [(\bar{A} = \mu_2, \lambda = 1)]$ satisfies Condition 2.3. In this case, the moduli space $\mathcal{F}$ turns out to be the moduli space of pairs consisting of a cubic threefold and a hyperplane section. This was recently studied in [Laza et al. 2018]. Denote by $\mathcal{W}_1 = H^0(\mathbb{P}^4, \mathcal{O}(3))$ the space of cubic forms in $x_0, \ldots, x_4$ and by $\mathcal{W}_2 = H^0(\mathbb{P}^4, \mathcal{O}(1))$ the space of linear forms in $x_0, \ldots, x_4$. We have

an identification $\mathcal{W}_1 \oplus \mathcal{W}_2 \cong \mathcal{V}$ sending (L_3, L_1) to $L_3 + x_5^2 L_1$. In [Laza et al. 2018], the authors defined $\mathcal{F}$ to be a GIT-quotient of $(\mathbb{P}\mathcal{W}_1 \times \mathbb{P}\mathcal{W}_2, \mathcal{O}(3) \boxtimes \mathcal{O}(1))$ by $\mathrm{SL}(5, \mathbb{C})$. Direct calculation shows that $N = C = \mathrm{SL}(5, \mathbb{C}) \times Z \subset \mathrm{SL}(V)$, where $Z = \{\mathrm{diag}(u, u, u, u, u, u^{-5}) \mid u \in \mathbb{C}^\times\}$ is the center. The following proposition gives the relation of our constructions with that in [Laza et al. 2018]:

Proposition 6.8. *We have identification between polarized projective varieties*:

$$Z\backslash\backslash(\mathbb{P}\mathcal{V}, \mathcal{O}(1)) \cong (\mathbb{P}\mathcal{W}_1 \times \mathbb{P}\mathcal{W}_2, \mathcal{O}(3) \boxtimes \mathcal{O}(1)).$$

Proof. It is equivalent to show

$$\bigoplus_k (H^0(\mathbb{P}\mathcal{V}, \mathcal{O}(k)))^Z \cong \bigoplus_k H^0(\mathbb{P}\mathcal{W}_1 \times \mathbb{P}\mathcal{W}_2, \mathcal{O}(3k) \boxtimes \mathcal{O}(k))$$

as graded algebras. The action of Z on $\mathcal{W}_1$ has weight 3, and on $\mathcal{W}_2$ weight -9.

We have the direct sum decomposition

$$\mathrm{Sym}^m(\mathcal{V}^*) = \bigoplus_{k+l=m} \mathrm{Sym}^k \mathcal{W}_1^* \otimes \mathrm{Sym}^l \mathcal{W}_2^*$$

with Z-action of weight $-3k+9l$. The weight zero part has $k = 3l$ and $m = 4l$. So we obtain identification between the two polarized varieties. □

Moreover, by Proposition 6.5, the hyperplane arrangement $\mathcal{H}_*$ is empty in this case, and we obtain identification between $\overline{\mathcal{F}}$ and Baily–Borel compactification $\overline{\Gamma\backslash\mathbb{D}}^{bb}$. It is straightforward to see that the arithmetic group Γ is exactly the one used in [Laza et al. 2018]. Therefore, we recover the main result in [Laza et al. 2018].

Appendix: Locally symmetric varieties and Looijenga compactifications

It is well-known that the normalization of each stratum in the orbifold loci of a locally Hermitian symmetric variety is still a locally Hermitian symmetric variety. For the reader's convenience, we include a discussion of this fact in Section A1. In the rest of the appendix, we prove that a similar result (Theorem A.13) holds for Looijenga compactifications. This is first observed by Looijenga [2016, p. 72]. We provide the complete formalism and the details of the proof.

We will recall the construction of Looijenga compactifications [2003a; 2003b] of arithmetic quotients $\mathbb{X}$ of complex hyperbolic balls or type IV domains. There are two steps. The first is constructing the semitoric blowup $\overline{\mathbb{X}}^\Sigma$, which is an intermediate compactification of arithmetic quotient $\mathbb{X}$ sitting between Baily–Borel and toroidal compactifications. We will recall the geometric construction of Baily–Borel compactifications of complex hyperbolic balls and type IV domains in Section A3, and recall the semitoric blow-up construction in Section A4. The second step is taking successive blow-up constructions along the hyperplane arrangement in $\overline{\mathbb{X}}^\Sigma$ and blow-down constructions of certain induced strata (we will sketch this in Section A5).

The idea of the proof of Theorem A.13 is that natural morphisms between arithmetic quotients (of balls or type IV domains) can be extended to morphisms between Baily–Borel compactifications, semitoric compactifications and Looijenga compactifications. Moreover, the extensions are finite morphisms. We call this the functorial property. We will prove the functorial property for Baily–Borel compactification in Section A3, for semitoric compactifications in Section A4, and for Looijenga compactification in Section A5. The existence of the extension in the Baily–Borel case is done in [Kiernan and Kobayashi 1972]. Harris [1989] proved the functorial properties for toroidal compactifications of locally symmetric varieties. The other part of our results in Sections A3, A4, A5, up to our knowledge, are new. Our proof follows the same idea in [Harris 1989]. We need to verify that the combinatorial data associated with the ambient hyperplane arrangements induces the same type of combinatorial data for subspaces, and match each stratum accordingly.

A1. ***Orbifold loci of locally symmetric varieties.*** In this section we show the normalization of an orbifold stratum of a locally Hermitian symmetric variety is again a locally Hermitian symmetric variety.

Let G be a real reductive algebraic group with compact center. Let K be a maximal compact subgroup of G. Let $\mathbb{D} = G/K$ be the corresponding symmetric space. Assume $\mathbb{D}$ is Hermitian symmetric and G has a $\mathbb{Q}$-structure. Let $\Gamma \subset G(\mathbb{Q})$ be an arithmetic subgroup. For simplicity, we assume the action of Γ on $\mathbb{D}$ is faithful. Denote by $\mathbb{X} = \Gamma\backslash\mathbb{D}$ the arithmetic quotient. This is naturally a quasiprojective variety due to Baily–Borel compactification [1966]. There is a natural orbifold structure on $\mathbb{X}$. We consider the orbifold locus indexed by certain finite subgroup $A \subset \Gamma$. More precisely, we take $A \subset \Gamma$ fixing some point $x \in \mathbb{D}$. Without loss of generality, we assume K to be the stabilizer of $x \in \mathbb{D}$ under the action of G. Then $A \subset K$. Denote by G_A, K_A and Γ_A the corresponding normalizers of A in G, K and Γ, respectively. Then G_A is again a real reductive algebraic group with compact center and K_A is a maximal compact subgroup (see [Looijenga 2016, pp. 37–38]). There is a natural holomorphic embedding

$$G_A/K_A \hookrightarrow \mathbb{D} = G/K.$$

Define $\mathbb{D}_A := G_A/K_A$. This is a Hermitian symmetric subspace of $\mathbb{D}$. We have the following proposition:

Proposition A.1. *The group Γ_A is an arithmetic subgroup in $G_A(\mathbb{Q})$ and the map $\pi : \Gamma_A\backslash\mathbb{D}_A \to \Gamma\backslash\mathbb{D}$ is finite. Furthermore, if A is the stabilizer of x under the action of Γ, then this map gives a normalization of its image.*

Proof. Due to the extension theorem of Baily–Borel compactifications (see Theorem 2 in [Kiernan and Kobayashi 1972]), the map π is algebraic and proper. We show π is finite. It suffices to show π is quasifinite, namely, having finite fibers. Take any $y \in \mathbb{D}_A$. Suppose we have a point $y' = \rho y$ for $\rho \in \Gamma$. Then $\rho^{-1}A\rho$ is contained in the stabilizer group of y. Actually, the Γ_A-orbits of such points y' are one-to-one corresponding to subgroups with form $\rho^{-1}A\rho$ in the stabilizer group of y, hence finitely many.

If A is the stabilizer group of x, a generic point in $\mathbb{X}_A := \Gamma_A\backslash\mathbb{D}_A$ also has A as stabilizer group. We first show that π is generically injective in this case. Take generically $x_1, x_2 \in \mathbb{D}_A$, and assume they $[x_1] = [x_2]$ in $\Gamma\backslash\mathbb{D}$. Then there exists $\rho \in \Gamma$ such that $\rho x_1 = x_2$. Since both x_1, x_2 have stabilizer group A,

we have $\rho A\rho^{-1} = A$; hence $\rho \in \Gamma_A$. This implies that $[x_1] = [x_2]$ in $\Gamma_A\backslash\mathbb{D}_A$. We have π a finite and birational morphism from a normal variety to its image, hence a normalization of its image. □

Remark A.2. The same construction also works for any finite volume locally Hermitian symmetric varieties. The difference from the arithmetic case is that Γ_A is not automatically a lattice. We need to use the compactification in finite volume case (see Theorem 1 in [Mok and Zhong 1989]) to show that the orbifold locus also admits a compactification, which implies the finiteness of the volume by Yau's Schwarz lemma [1978].

A2. ***Orbifold loci of ball and type* IV *quotients.*** We now focus on arithmetic quotients of balls and type IV domains.

We fix some notation that will be used in the rest of the appendix. Let $(V_{\mathbb{Q}}, \varphi)$ be a vector space over $\mathbb{Q}$ with nondegenerate rational bilinear form φ of signature $(2, N)$. Let $V = V_{\mathbb{Q}} \otimes \mathbb{C}$. Notice that here $V_{\mathbb{Q}}$ is not necessarily the middle cohomology of cubic fourfold. Similar to Section 3, the type IV domain $\widehat{\mathbb{D}}$ attached to $(V_{\mathbb{Q}}, \varphi)$ is a component of

$$\widehat{\mathbb{D}} \sqcup \overline{\widehat{\mathbb{D}}} = \mathbb{P}\{x \in V \mid \varphi(x, x) = 0, \varphi(x, \bar{x}) > 0\}.$$

Denote by $\widehat{G}$ the subgroup of $\mathrm{Aut}(\varphi)(\mathbb{R})$ (of index 2) respecting the component $\widehat{\mathbb{D}}$. Let $\widehat{\Gamma} \subset \widehat{G}$ be an arithmetic subgroup. The corresponding locally Hermitian symmetric variety is $\widehat{\mathbb{X}} = \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$. Let A be a finite subgroup of $\widehat{\Gamma}$. Let ζ be a character of A, such that there exists $x \in V$ with $\varphi(x, x) = 0$ and $\varphi(x, \bar{x}) > 0$, and $a(x) = \zeta(a)x$ for all $a \in A$. Denote by V_ζ the ζ-subspace of V. Then there is a natural Hermitian form h on V_ζ defined by $h(x, y) = \varphi(x, \bar{y})$. If $\zeta = \bar{\zeta}$, this Hermitian form has signature $(2, n)$ and we obtain a type IV subdomain $\mathbb{D}$ of $\widehat{\mathbb{D}}$. Otherwise the signature is $(1, n)$ and we obtain a complex hyperbolic ball $\mathbb{B}$ inside $\widehat{\mathbb{D}}$. Indeed, let

$$G := \{g \in \widehat{G} \mid gAg^{-1} = A\}$$

be an algebraic subgroup over $\mathbb{Q}$. The fixed locus of A in $\mathbb{D}$ is $G(\mathbb{R})/K$, where K is maximal compact subgroup of $G(\mathbb{R})$. Denote $\Gamma = \{\rho \in \widehat{\Gamma} \mid \rho^{-1}A\rho = A\}$. As in Section 4, we have Γ an arithmetic subgroup of $G(\mathbb{Q})$ acting on $\mathbb{B}$ or $\mathbb{D}$. Then we have a natural map $\Gamma\backslash\mathbb{D} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ or $\Gamma\backslash\mathbb{B} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$. We consider the following condition:

Condition A.3. The group A is the stabilizer of a generic point of $\mathbb{D}$ or $\mathbb{B}$.

If A satisfies this condition, Proposition A.1 implies that the morphism $\pi : \Gamma\backslash\mathbb{B} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ or $\pi : \Gamma\backslash\mathbb{D} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ is the normalization of its image.

We will consider a larger set of type IV subdomains. Taking $W_{\mathbb{Q}}$ to be a $\mathbb{Q}$-subspace of $V_{\mathbb{Q}}$ with signature $(2, n)$, we have the associated type IV subdomain $\mathbb{D}$ inside $\widehat{\mathbb{D}}$ with the action of an arithmetic group $\Gamma_W = \{\rho \in \widehat{\Gamma} \mid \rho(W) = W\}$. Take $V_{\mathbb{Z}}$ to be an integral structure on $V_{\mathbb{Q}}$ such that $\Gamma \subset \mathrm{Aut}(V_{\mathbb{Z}})$ has finite index. Denote $W_{\mathbb{Z}} := W_{\mathbb{Q}} \cap V_{\mathbb{Z}}$. For $x \in \mathbb{D}$, define $\mathrm{Pic}(x) := V_x^{1,1} \cap V_{\mathbb{Z}}$ to be the Picard lattice of x where x is viewed as a weight two Hodge structure on $V_{\mathbb{Z}}$. Then for generic $x \in \mathbb{D}$, we have $\mathrm{Pic}(x) = W_{\mathbb{Z}}^{\perp}$.

We have the following lemma:

Lemma A.4. *For A satisfying Condition A.3 and $W = V_\zeta$, we have $\Gamma_A = \Gamma_W$.*

Proof. It is straightforward that $\Gamma_A \subset \Gamma_W$, and they both act on $\mathbb{D}$. Take any $\rho \in \Gamma_W$ and a generic point x in $\mathbb{D}$. Then A is contained in the stabilizer group of ρx. Thus both A and $\rho^{-1}A\rho$ are contained in the stabilizer group of x. Since x is generic, we have $\rho^{-1}A\rho = A$ by Condition A.3. So $\rho \in \Gamma_A$. We showed that $\Gamma_W \subset \Gamma_A$. □

With this lemma, we will simply denote by Γ the arithmetic group acting on $\mathbb{D}$. We have:

Proposition A.5. *For any $\mathbb{Q}$-subspace $W_\mathbb{Q}$ (of $V_\mathbb{Q}$) with signature $(2, n)$, we have a morphism $\pi : \Gamma\backslash\mathbb{D} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$, which is the normalization of its image.*

Proof. Properness is by [Kiernan and Kobayashi 1972]. Take a generic point x in $\mathbb{D}$. Suppose $\rho \in \widehat{\Gamma}$ sends x to $\rho x \in \mathbb{D}$. The Picard lattice $\mathrm{Pic}(\rho x)$ of ρx contains $W_\mathbb{Z}^\perp$; hence $\rho^{-1}(W_\mathbb{Z}^\perp) \subset \mathrm{Pic}(x)$. Since x is generic, we have $\mathrm{Pic}(x) = W_\mathbb{Z}^\perp$. This implies that $\rho(W_\mathbb{Z}^\perp) = W_\mathbb{Z}^\perp$; hence $\rho(W) = W$. Thus $\rho \in \Gamma_W$.

Finally, we show finiteness. Take a point $x \in \mathbb{D}$. For any $\rho \in \widehat{\Gamma}$, we have $\rho^{-1}(W_\mathbb{Z}^\perp)$ contained in the Picard lattice $\mathrm{Pic}(x)$. The set $\widehat{\Gamma}x$ is a disjoint union of some Γ-orbits, each of which corresponds to the image of certain primitive embedding of $W_\mathbb{Z}^\perp$ into $\mathrm{Pic}(x)$. The orthogonal complement of $W_\mathbb{Z}^\perp$ in $\mathrm{Pic}(x)$ is positive definite with discriminant at most $\det(W_\mathbb{Z}^\perp)\det(\mathrm{Pic}(x))$. By reduction theory of lattice, there are finitely many such primitive embeddings. □

A3. ***Functoriality of Baily–Borel compactification.*** In this section we recall Baily–Borel compactifications of arithmetic quotients of complex hyperbolic balls or type IV domains; see [Baily and Borel 1966; Looijenga 2003a; 2003b].

We deal with type IV domain $\widehat{\mathbb{D}}$ first. The boundary components of Baily–Borel compactifications corresponds to $\mathbb{Q}$-isotropic planes J or $\mathbb{Q}$-isotropic lines I. Let

$$\pi_{J^\perp} : \mathbb{P}(V) - \mathbb{P}(J^\perp) \to \mathbb{P}(V/J^\perp) \quad \text{and} \quad \pi_{I^\perp} : \mathbb{P}(V) - \mathbb{P}(I^\perp) \to \mathbb{P}(V/I^\perp)$$

be the natural projections. The image $\pi_{J^\perp}\widehat{\mathbb{D}}$ is isomorphic to upper half plane. The image $\pi_{I^\perp}\widehat{\mathbb{D}}$ is a point. Adding rational boundary components, we have

$$\widehat{\mathbb{D}}^{bb} := \widehat{\mathbb{D}} \sqcup \coprod_J \pi_{J^\perp}\widehat{\mathbb{D}} \sqcup \coprod_I \pi_{I^\perp}\widehat{\mathbb{D}}$$

with suitable topology and ringed space structure. The Baily–Borel compactification is the quotient $\Gamma\backslash\widehat{\mathbb{D}}^{bb}$ as a projective variety.

Given $W_\mathbb{Q} \subset V_\mathbb{Q}$ with signature $(2, n)$. Let $\mathbb{D}$ be the corresponding type IV domain. We have a natural map from $\mathbb{D}$ to $\widehat{\mathbb{D}}$, inducing $\Gamma\backslash\mathbb{D} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$. According to Theorem 2 in [Kiernan and Kobayashi 1972], this holomorphic map can be extended to Baily–Borel compactifications, sending boundary components into boundary components.

Proposition A.6 (type IV to type IV). *There is a natural finite extension of $\pi : \Gamma\backslash\mathbb{D} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ to Baily–Borel compactifications*

$$\pi : \overline{\Gamma\backslash\mathbb{D}}^{bb} \to \overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{bb}.$$

If A satisfies Condition A.3, the map is a normalization of its image.

Proof. Let $W := V_\zeta$ in this proof. The boundary components of $\mathbb{D}^{bb}$ correspond to rational isotropic planes J and rational isotropic lines I in W. From the natural embedding $W \hookrightarrow V$, they also have associated boundary components in $\widehat{\mathbb{D}}^{bb}$. Under the natural commutative diagram

$$\begin{array}{ccc} \mathbb{P}(W) - \mathbb{P}(J^\perp) & \xrightarrow{\pi_{J^\perp}} & \mathbb{P}(W/J^\perp) \\ \downarrow & & \downarrow \\ \mathbb{P}(V) - \mathbb{P}(J^\perp) & \xrightarrow{\pi_{J^\perp}} & \mathbb{P}(V/J^\perp) \end{array}$$

we have isomorphisms $\pi_{J^\perp}\mathbb{D} \to \pi_{J^\perp}\widehat{\mathbb{D}}$, and similar maps $\pi_{I^\perp}\mathbb{D} \to \pi_{I^\perp}\widehat{\mathbb{D}}$, which induce an extension $\mathbb{D}^{bb} \to \widehat{\mathbb{D}}^{bb}$ equivariant under the action of $\Gamma \to \widehat{\Gamma}$. After taking quotients, we have an extension map $\mathbb{X}^{bb} \to \widehat{\mathbb{X}}^{bb}$. By Proposition A.1, this map is generically injective and it is finite over $\widehat{\Gamma}\backslash\widehat{\mathbb{D}}$. Let Γ_J be the stabilizer of J under the action of Γ. The projection of Γ in $\mathrm{GL}(J)$ (or equivalently $\mathrm{GL}(V/J^\perp)$) is arithmetic. The boundary component corresponding to J is the quotient of $\pi_{J^\perp}\widehat{\mathbb{D}}$ by Γ_J, hence a modular curve. The restriction to the boundary component corresponding to each J is a nonconstant map between modular curves, hence finite. The restriction to boundary components corresponding to each I is automatically finite. So we have an algebraic finite morphism between normal varieties $\mathbb{X}^{bb} \to \widehat{\mathbb{X}}^{bb}$. If A satisfies Condition A.3, then this morphism is generically injective by Proposition A.1, hence a normalization of its image. □

We recall the Baily–Borel compactification of Ball quotient. Let K be a CM field and W_K a finite-dimensional vector space over K with

$$h_K : W_K \times W_K \to K$$

a K-valued Hermitian form. For each embedding $\iota : K \hookrightarrow \mathbb{C}$, we define $W_\iota := W_K \otimes_\iota \mathbb{C}$, then we have a Hermitian form $h_\iota : W_\iota \times W_\iota \to \mathbb{C}$. Assume the form h_ι has signature $(1, n)$ under embedding $\iota = \iota_1$ or $\bar{\iota}_1$, and is definite otherwise. The complex ball $\mathbb{B}$ is defined to be the set of positive lines in W_{ι_1}. The boundary components of Baily–Borel compactification correspond to K-isotropic lines I in W_K and we denote $\mathbb{B}^{bb} := \mathbb{B} \sqcup \coprod_I \pi_{I^\perp}\mathbb{B}$. When the totally real part of K is not $\mathbb{Q}$, there exists complex embedding ι such that (W_ι, h_ι) is definite, which implies that any isotropic vector in W_K must be zero. Thus in this case the boundary set is empty.

Now consider the action of A on V with $\zeta \neq \bar{\zeta}$. Let K be the cyclotomic field generated by $\zeta(A)$. Take W_K to be the ζ-eigenspace of $V_K := V_\mathbb{Q} \otimes K$ under the action of A.

Lemma A.7. *The K-vector space W_K is isotropic under φ.*

Proof. Taking any $x, y \in W_K$, we need to show $\varphi(x, y) = 0$. Take $a \in A$ such that $\zeta(a)$ is not real. Then $\zeta(a)^2 \neq 1$. By

$$\varphi(x, y) = \varphi(ax, ay) = \varphi(\zeta(a)x, \zeta(a)y) = \zeta(a)^2\varphi(x, y),$$

we have $\varphi(x, y) = 0$. □

There is a natural Hermitian form h of signature $(1, n)$ on W_K, given by $h(x, y) = \varphi(x, \bar{y})$ for all $x, y \in W_K$. The Galois conjugates of K define eigenspaces of V under the action of A. The sum of all those eigenspaces is a subspace of V defined over $\mathbb{Q}$. Then we have the ball $\mathbb{B}$ consisting of positive lines in W and we denote $\overline{(\Gamma\backslash\mathbb{B})}^{bb} := \Gamma\backslash\mathbb{B}^{bb}$ the Baily–Borel compactification of $\mathbb{X} = \Gamma\backslash\mathbb{B}$.

Proposition A.8 (ball to type IV). *There is a natural finite extension of $\pi : \Gamma\backslash\mathbb{B} \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ to Baily–Borel compactifications*

$$\pi : \overline{\Gamma\backslash\mathbb{B}}^{bb} \to \overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{bb}.$$

If A satisfies Condition A.3, the map is a normalization of its image.

Proof. Similar as the proof for type IV case, we need to identify the boundary components on both sides. The ball and its boundaries are defined as above by W_K. If K is not a quadratic extension of $\mathbb{Q}$, then the boundary set is empty; hence $\Gamma\backslash\mathbb{B}$ is already compact. If K is, then each K-isotropic line I together with its complex conjugate $\bar{I}$ defines a rational isotropic plane in $V_{\mathbb{Q}}$. So there is a natural extension map $\mathbb{B}^{bb} \to \widehat{\mathbb{D}}^{bb}$ which is equivariant under the action of $\Gamma \to \widehat{\Gamma}$. After taking quotient on both sides, we have a finite algebraic map $\pi : \mathbb{X}^{bb} \to \widehat{\mathbb{X}}^{bb}$. If A satisfies Condition A.3, then this morphism is generically injective by Proposition A.1, hence a normalization of its image. □

Remark A.9. Similar constructions of ball quotients appears in the arithmetic examples of Deligne–Mostow theory; see [Deligne and Mostow 1986; Looijenga 2007]. In both constructions, if the cyclotomic field generated by the corresponding characters is not $\mathbb{Q}(\sqrt{-1})$ or $\mathbb{Q}(\sqrt{-3})$, then the Baily–Borel compactification is compact.

A4. ***Functoriality of semitoric compactifications.*** We first briefly sketch the semitoric blow-up constructions of complex hyperbolic balls and type IV domains with respect to certain hyperplane arrangements; see [Looijenga 2003a; 2003b]. Semitoric compactification with respect to a hyperplane arrangement is the minimal blowup of certain boundary components in Baily–Borel compactification, such that the closure of every hypersurface is Cartier at the boundary.

Let $\widehat{\mathcal{H}}$ be a hyperplane arrangement on $\widehat{\mathbb{D}}$ defined by a set of negative vectors $v \in V_{\mathbb{Q}}$, which form finitely many orbits under the action of $\widehat{\Gamma}$. We recall some definitions and notation in [Looijenga 2003b]. Each rational isotropic line I in $V_{\mathbb{Q}}$ realizes $\widehat{\mathbb{D}}$ as a tube domain, with real cone denoted by

$$C_I \subset (I^{\perp}/I \otimes I)(\mathbb{R}).$$

Each rational isotropic plane J determines a half line on the boundaries of the C_I for any $I \subset J$. The union of these cones is called the conical locus of $\widehat{\mathbb{D}}$. Let $C_{I,+}$ be the convex hull of $\overline{C}_I \cap (I^{\perp}/I \otimes I)(\mathbb{Q})$, which is the union of C_I with rational isotropic half lines corresponding to J containing I. The hyperplane arrangement $\widehat{\mathcal{H}}$ determines an admissible decomposition $\Sigma(\widehat{\mathcal{H}})$ of the conical locus. More precisely, it is a Γ-invariant choice of locally rational cone decomposition of $C_{I,+}$ such that the support for isotropic half line corresponding to J is independent of those $I \subset J$; see Section 6 of [Looijenga 2003b] for details.

For each member σ of $\Sigma(\widehat{\mathcal{H}})$ contained in $C_{I,+}$, we define a corresponding vector subspace V_σ of V as follows. When σ is the half line corresponding to an isotropic plane J, then

$$V_\sigma := \big(\bigcap_{J\subset H} H\big)\cap J^\perp.$$

Otherwise V_σ is the span of σ in $I^\perp$, which is also the intersection among $I^\perp$ and those $H\in\widehat{\mathcal{H}}$ containing I. Here we identify $H\subset V$ with $H\in\widehat{\mathcal{H}}$. We have a projection $\pi_\sigma:\mathbb{D}\to\mathbb{P}(V/V_\sigma)$. The semitoric compactification is denoted by $\overline{\mathbb{X}}^\Sigma=\Gamma\backslash\mathbb{D}^\Sigma$. Here $\mathbb{D}^\Sigma:=\coprod_{\sigma\in\Sigma}\pi_\sigma\widehat{\mathbb{D}}$. The space $\overline{\mathbb{X}}^\Sigma$ has a natural map to $\widehat{\mathbb{X}}^{bb}$ respecting the stratifications. We have two different types of boundary components. One is finite quotient of an abelian torsor over the modular curve $\widehat{\Gamma}_J\backslash\pi_{J^\perp}\widehat{D}$. The abelian torsor is modeled over vector group $J^\perp/V_\sigma$ quotient by a lattice. The other is an algebraic torus torsor over a point $\pi_{I^\perp}\widehat{D}$, which is the boundary stratum in the quotient of an infinite-type toric variety induced by the cone decomposition of $C_{I,+}$. In particular, each cone of codimension k corresponds to algebraic torus torsor of dimension k.

Given $W_\mathbb{Q}\subset V_\mathbb{Q}$ a sublattice of signature $(2,n)$, with $\mathbb{D}$ the associated type IV domain, we have the intersection $\mathcal{H}:=\mathbb{D}\cap\widehat{\mathcal{H}}$ a Γ-invariant hyperplane arrangement in $\mathbb{D}$. We also have the semitoric blowup of $\mathbb{D}$ with respect to $\mathcal{H}$.

Theorem A.10 (type IV to type IV). *There is a natural finite extension of $\pi:\Gamma\backslash\mathbb{D}\to\widehat{\Gamma}\backslash\widehat{\mathbb{D}}$ to semitoric compactifications*

$$\pi:\overline{\Gamma\backslash\mathbb{D}}^{\Sigma(\mathcal{H})}\to\overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\Sigma(\widehat{\mathcal{H}})}.$$

If A satisfies Condition A.3, the map is a normalization of its image.

Proof. We first show the existence of $\pi:\overline{\Gamma\backslash\mathbb{D}}^{\Sigma(\mathcal{H})}\to\overline{\widehat{\Gamma}\backslash\widehat{\mathbb{D}}}^{\Sigma(\widehat{\mathcal{H}})}$ as a morphism between two projective varieties, then prove finiteness.

The subdomain is induced by (W,φ). The isotropic lines and planes in W are naturally viewed as boundary data of both $\mathbb{D}$ and $\widehat{\mathbb{D}}$. The conical locus of $\mathbb{D}$ is naturally embedded into that of $\widehat{\mathbb{D}}$.

Suppose $\sigma\in\Sigma(\mathcal{H})$ does not correspond to a rational isotropic plane of W. Then we have a rational isotropic line I, such that σ is contained in $C_{I,W,+}$ and intersects with $C_{I,W}$. For each H containing I, the intersection $H\cap C_{I,W}$ being not empty is equivalent to $H\cap\mathbb{D}$ being not empty. Then there exists $\tau\in\Sigma(\widehat{\mathcal{H}})$ such that $\sigma=\tau\cap W$. We denote by $\widehat{\sigma}$ the minimal element among all such τ. Thus $\sigma=C_{I,W}\cap\widehat{\sigma}$, which implies $W_\sigma=V_{\widehat{\sigma}}\cap W$.

Let $\sigma\in\Sigma(\mathcal{H})$ correspond to an isotropic plane J contained in both W and a hyperplane H. Suppose v is a normal vector of H and $v=w+w^\perp$ the decomposition in $V=W\oplus W^\perp$. We have $\varphi(v,v)<0$. The hyperplane H intersects with $\mathbb{D}$ if and only if $\varphi(w,w)<0$. Since the orthogonal complement of w in $W_\mathbb{Q}$ contains the isotropic plane J, we have either $\varphi(w,w)<0$ or $\varphi(w,w)=0$. Suppose the latter case happens, then $w\in J$ since otherwise $\langle J,w\rangle$ is an isotropic subspace of rank 3 contained in $W_\mathbb{Q}$, which is impossible. Thus in this case $H\supset J^\perp\cap W$. The above argument holds for any $H\in\mathcal{H}$ containing σ; hence $W_\sigma=V_\sigma\cap W$. In this case we also denote $\widehat{\sigma}=\sigma$.

For $\sigma = \{0\} \in \Sigma(\mathcal{H})$, just take $\widehat{\sigma} = \{0\} \in \Sigma(\widehat{\mathcal{H}})$. Then for every $\sigma \in \Sigma(\mathcal{H})$, we have a natural holomorphic map $\pi_\sigma \mathbb{D} \to \pi_{\widehat{\sigma}} \widehat{\mathbb{D}}$ which is apparently injective. Taking union among σ, we have

$$\coprod_{\sigma \in \Sigma(\mathcal{H})} \pi_\sigma \mathbb{D} \to \coprod_{\sigma \in \Sigma(\mathcal{H})} \pi_{\widehat{\sigma}} \widehat{\mathbb{D}} \hookrightarrow \coprod_{\tau \in \Sigma(\widehat{\mathcal{H}})} \pi_\tau \widehat{\mathbb{D}}$$

with the composition continuous. After taking quotients by the equivariant actions on both sides, we obtain a finite map between the boundary components. Actually, for those rational isotropic planes J, we obtain finite morphisms between Abelian torsors; for those rational isotropic lines I, we obtain finite morphisms between algebraic torus torsors. If A satisfies Condition A.3, then π is generically injective by Proposition A.1, hence a normalization of its image. □

Remark A.11. The injectivity of $\coprod_{\sigma \in \Sigma(\mathcal{H})} \pi_\sigma \mathbb{D} \to \coprod_{\tau \in \Sigma(\widehat{\mathcal{H}})} \pi_\tau \widehat{\mathbb{D}}$ is already known by [Looijenga 2003b, paragraph after Lemma 7.1].

For $\zeta \neq \bar{\zeta}$, we have ball $\mathbb{B}$ attached to $W = V_\zeta$. We next describe the semitoric compactification of $\mathbb{B}$ with respect to $\mathcal{H}$. Here we identify elements in $\mathcal{H}$ with hypersurfaces in W. The cusp points correspond to isotropic lines I in W_K. Let

$$j(I) = \Big(\bigcap_{H \in \mathcal{H}, H \supset I} H \Big) \cap I_W^\perp$$

and $\pi_I : \mathbb{P}(W) - \mathbb{P}(j(I)) \to \mathbb{P}(W/j(I))$. Define

$$\overline{\mathbb{X}}^j = \Gamma \backslash \Big(\mathbb{B} \sqcup \coprod_I \pi_{j(I)} \mathbb{B} \Big).$$

It naturally maps to the Baily–Borel compactification. The boundary component over each cusp point is an abelian torsor modeled over the vector space $I_W^\perp / j(I)$ quotient by a lattice.

Theorem A.12 (ball to type IV). *There is a natural finite extension of $\pi : \Gamma \backslash \mathbb{B} \to \widehat{\Gamma} \backslash \widehat{\mathbb{D}}$ to semitoric compactifications*

$$\pi : \overline{\Gamma \backslash \mathbb{B}}^j \to \overline{\widehat{\Gamma} \backslash \widehat{\mathbb{D}}}^{\Sigma(\widehat{\mathcal{H}})}.$$

If A satisfies Condition A.3, the map is a normalization of its image.

Proof. If K is not a quadratic extension of $\mathbb{Q}$, then $\mathbb{X}$ is compact and the theorem holds. Now assume that K is a quadratic extension of $\mathbb{Q}$. Namely, $K = \mathbb{Q}(\sqrt{-D})$ for certain positive integer D. Take any isotropic line I in W_K. Suppose a nonzero generator of I is $e + \sqrt{-D} f$, where $e, f \in V_\mathbb{Q}$. Then $\varphi(e + \sqrt{-D} f, e - \sqrt{-D} f) = 0$. From Lemma A.7 we have $\varphi(e + \sqrt{-D} f, e + \sqrt{-D} f) = 0$. This implies that $J = \langle e, f \rangle$ is an isotropic plane in $V_\mathbb{Q}$.

We claim that $j(I) = W \cap V_J$. Take $H \in \widehat{\mathcal{H}}$ with orthogonal vector $v \in V_\mathbb{Q}$. Under the orthogonal decomposition $V_K = W_K \oplus \overline{W_K} \oplus V'$, we can decompose v as $v = v_W + \overline{v_W} + v'$. Then $\varphi(\mathrm{Re}(v_W), J) = 0$. From Lemma A.7 we have $\varphi(v_W, I) = 0$. Therefore, $\varphi(\mathrm{Im}(v_W), I) = 0$ and hence $\varphi(\mathrm{Im}(v_W), J) = 0$.

Since $(V_{\mathbb{Q}}, \varphi)$ has signature $(2, N)$, the orthogonal complement of J in $V_{\mathbb{Q}}$ is negative semidefinite. Thus $\varphi(\mathrm{Re}(v_W), \mathrm{Re}(v_W)) \leq 0$ and $\varphi(\mathrm{Im}(v_W), \mathrm{Im}(v_W)) \leq 0$. We then have

$$\varphi(v_W, \overline{v_W}) = \varphi(\mathrm{Re}(v_W), \mathrm{Re}(v_W)) + \varphi(\mathrm{Im}(v_W), \mathrm{Im}(v_W)) \leq 0.$$

Suppose $\varphi(v_W, \overline{v_W}) < 0$, then $H \cap \mathbb{B} \neq \varnothing$. Suppose $\varphi(v_W, \overline{v_W}) = 0$, then v_W is an isotropic line in W_K. The vectors $\mathrm{Re}(v_W)$ and $\mathrm{Im}(v_W)$ in $V_{\mathbb{Q}}$ are then isotropic. These two vectors are orthogonal to J; hence they belong to J. We deduce that $H \supset I_W^{\perp}$. By the definition of $j(I)$ and V_J, we conclude the claim.

We now have naturally an injective map $\pi_{j(I)}\mathbb{B} \to \pi_J \widehat{\mathbb{D}}$. Taking the union among those isotropic lines I, we have an injective map

$$\mathbb{B} \sqcup \coprod_I \pi_{j(I)}\mathbb{B} \hookrightarrow \coprod_{\sigma \in \Sigma(\widehat{\mathcal{H}})} \pi_\sigma \widehat{\mathbb{D}}.$$

After taking quotients by the equivariant actions on both sides, we obtain a morphism $\pi : \overline{\Gamma\backslash\mathbb{B}}^j \to \overline{\Gamma\backslash\widehat{\mathbb{D}}}^{\Sigma(\widehat{\mathcal{H}})}$. Actually, the restriction of this π to the boundary component corresponding to I is a finite morphism between Abelian torsors. We conclude that there is natural extension

$$\pi : \overline{(\Gamma\backslash\mathbb{B})}^j \to \overline{(\widehat{\Gamma}\backslash\widehat{\mathbb{D}})}^{\Sigma(\widehat{\mathcal{H}})}$$

which is a finite morphism between projective varieties. If A satisfies Condition A.3, this π is generically injective, hence a normalization of its image. □

A5. ***Main theorem.*** In this section, we first describe the construction of Looijenga compactification $\overline{\mathbb{X}}^{\mathcal{H}}$ of $\mathbb{X}^\circ := \mathbb{X} - \Gamma\backslash\mathcal{H}$. We need to successively blow up nonempty intersections of components of $\Gamma\backslash\mathcal{H}$, and then contract the strict transformations of $\Gamma\backslash\mathcal{H}$ via a natural associated semiample line bundle on the blowup. We then prove existence and finiteness of morphism between Looijenga compactifications on both sides of $\mathbb{X} \to \widehat{\mathbb{X}}$.

The blow-up and blow-down constructions with respect to hyperplane arrangement in any normal analytic variety with a properly given line bundle are discussed in the first three sections in [Looijenga 2003a]. Looijenga applied this general theory to $(\overline{\mathbb{X}}^{\Sigma(\mathcal{H})}, \Gamma\backslash\mathcal{H}, \mathcal{L})$, where $\mathbb{X}$ is either arithmetic quotient of type IV domain $\mathbb{D}$ or ball $\mathbb{B}$, and $\mathcal{L}$ is the natural automorphic line bundle; see [Looijenga 2003a, Theorem 5.7; 2003b, Theorem 7.4].

We now describe the blow-up and blow-down constructions before quotient by the arithmetic groups. The Looijenga compactifications are obtained by the modified spaces quotient by the arithmetic groups. Denote by $\mathrm{PO}(\mathcal{H})$ the set of nonempty intersections of elements in $\mathcal{H}$ as hyperplanes in $\mathbb{D}$ (or $\mathbb{B}$). Let $L \in \mathrm{PO}(\mathcal{H})$ also denote its closure in $\mathbb{D}^\Sigma$ (or $\mathbb{B}^j$). Denote $c(L) := \mathrm{codim}(L) - 1$.

We first look at the semitoric compactification $\mathbb{D}^\Sigma$ of $\mathbb{D}$. Denote by $(\mathbb{D}^\Sigma)^\circ$ the arrangement complement of $\mathcal{H}$ in $\mathbb{D}^\Sigma$. Choose $L \in \mathrm{PO}(\mathcal{H})$ a minimal member. Blowing up along L replaces L by the projectivization of its normal bundle, which is isomorphic to $L \times \mathbb{P}^{c(L)}$. The modified space, denoted by $\mathrm{Bl}_L(\mathbb{D}^\Sigma)$, has natural topology, arrangement (the strict transform of the previous one) and automorphic line bundle. The strict transforms of those hypersurfaces passing through L form a hyperplane arrangement in $\mathbb{P}^{c(L)}$,

and we denote the complement by $(\mathbb{P}^{c(L)})^\circ$. The complement of the new arrangement in $\mathrm{Bl}_L(\mathbb{D}^\Sigma)$ is the disjoint union $(\mathbb{D}^\Sigma)^\circ \sqcup L \times (\mathbb{P}^{c(L)})^\circ$. After blowing up successively until hypersurfaces disjoint, we obtain the final blowup $\widetilde{\mathbb{D}}$. This is a disjoint union of $(\mathbb{D}^\Sigma)^\circ$ with $L \times (\mathbb{P}^{c(L)})^\circ$ for all such minimal L appearing in each step.

Now we can contract $L \times (\mathbb{P}^{c(L)})^\circ$ along the direction of L for all such L, and obtain $\mathbb{D}^*$ with natural quotient topology. Set-theoretically, $L \times (\mathbb{P}^{c(L)})^\circ$ is contracted to $(\mathbb{P}^{c(L)})^\circ$. We have the following description (see [Looijenga 2003b]):

$$\mathbb{D}^* = \coprod_{L \in \mathrm{PO}(\mathcal{H})} \pi_L \mathbb{D}^\circ \sqcup \coprod_{\sigma \in \Sigma(\mathcal{H})} \pi_\sigma \mathbb{D}^\circ.$$

Notice that for σ being the vertex, π_σ is identity and $\pi_\sigma \mathbb{D}^\circ = \mathbb{D}^\circ$.

The spaces $\mathbb{D}^\Sigma$, $\widetilde{\mathbb{D}}$, $\mathbb{D}^*$ constructed above all have natural ringed space structure. Namely, we have the structure sheaves consisting of continuous functions with analytic restriction to each stratum. The group Γ naturally acts on those ringed spaces respecting the stratification. The topological quotient space $\overline{\mathbb{X}}^{\mathcal{H}} := \Gamma \backslash \mathbb{D}^*$ has normal analytic structure respecting the stratification; see [Looijenga 2003b, Theorem 7.4]. According to the Riemann extension theorem, the quotient ringed space structure and the analytic structure on $\overline{\mathbb{X}}^{\mathcal{H}}$ coincide.

For the case of ball, parallel argument gives $\widetilde{\mathbb{B}}$ and $\mathbb{B}^*$. We have

$$\mathbb{B}^* = \mathbb{B}^\circ \sqcup \coprod_{L \in \mathrm{PO}(\mathcal{H})} \pi_L \mathbb{B}^\circ \sqcup \coprod_{I} \pi_{j(I)} \mathbb{B}^\circ.$$

This also has natural ringed structure, and $\overline{\mathbb{X}}^{\mathcal{H}} \cong \Gamma \backslash \mathbb{B}^*$ as analytic spaces.

Theorem A.13 (Main Theorem). *There is a natural extension of $\pi : \Gamma \backslash (\mathbb{D} - \mathcal{H}) \to \widehat{\Gamma} \backslash (\widehat{\mathbb{D}} - \widehat{\mathcal{H}})$ to Looijenga compactifications*

$$\pi : \overline{\Gamma \backslash \mathbb{D}}^{\mathcal{H}} \to \overline{\widehat{\Gamma} \backslash \widehat{\mathbb{D}}}^{\widehat{\mathcal{H}}}$$

which is a finite morphism. If A satisfies Condition A.3, the map is a normalization of its image. The same result holds for ball quotients.

Proof. From Theorem A.10, we have natural morphisms from $\mathbb{D}^\Sigma$ to $\widehat{\mathbb{D}}^\Sigma$. Near each boundary component, there is a contraction map from a neighborhood to the boundary itself. The arrangement in total space is the pullback of smooth arrangement on the boundary. According to the map defined near the boundary components, we know that any $H \in \widehat{\mathcal{H}}$ not intersecting with $\mathbb{D}$ is still away from $\mathbb{D}^\Sigma$ after taking its closure. From Corollary 7.15 in Chapter II in [Hartshorne 1977], we have injective map $\widetilde{\mathbb{D}} \to \widetilde{\widehat{\mathbb{D}}}$ respecting the ringed space structures. Notice that the automorphic line bundle on $\mathbb{D}^\Sigma$ is the pull back of that on $\widehat{\mathbb{D}}^\Sigma$; hence we have an injective map on the stratum $L \times (\mathbb{P}^{c(L)})^\circ$ to $\widehat{L} \times (\mathbb{P}^{c(\widehat{L})})^\circ$ which is linear on the second component. Here $\widehat{L}$ is a minimal member used in certain step of the successive blow-up construction of $\widehat{\mathbb{D}}$, and L is the induced member on the smaller subspace by intersecting with $\widehat{L}$. After blowing down, we have a natural injective map $\mathbb{D}^* \to \widehat{\mathbb{D}}^*$ respecting the ringed space structures.

This morphism descends to a morphism $\pi : \Gamma\backslash\mathbb{D}^* \to \widehat{\Gamma}\backslash\widehat{\mathbb{D}}^*$, still in the category of ringed spaces. We then have an analytic morphism

$$\pi : \overline{\mathbb{X}}^{\mathcal{H}} \to \overline{\widehat{\mathbb{X}}}^{\widehat{\mathcal{H}}}.$$

This analytic morphism extends $\pi : \mathbb{X}^\circ \to \widehat{\mathbb{X}}^\circ$, and sends boundary strata into boundary strata. Combining with Theorem A.10, the extended morphism π here is finite. If A satisfies Condition A.3, it is generically injective and hence a normalization of its image.

The same argument also holds for ball quotients. □

List of symbols

(d, k)	dimension and degree of a hypersurface
V	complex vector space of dimension $k+2$
F	degree d polynomial in $k+2$ variables
X	degree d k-fold; cubic fourfold when $(d, k) = (3, 4)$
A	a finite subgroup of $\mathrm{SL}(V)$, containing the center μ_{k+2} of $\mathrm{SL}(V)$
$\bar{A}$	image of A in $\mathrm{PSL}(V)$
λ	character of A with specified restriction to μ_{k+2}
T	equivalence class of (A, λ), called symmetry type of degree d k-fold
$\mathcal{V}$	eigenspace of $\mathrm{Sym}^d(V^*)$ corresponding to (A, λ)
C	centralizer of A in $\mathrm{SL}(V)$
N	a reductive group acting on $\mathcal{V}$
$\mathcal{V}^{sm}/\mathcal{V}^{ss}$	space of smooth/semistable points in $\mathcal{V}$
$\mathcal{F}$	GIT quotient of $\mathcal{V}^{sm}$ by N
$\mathcal{F}^m$	moduli space of cubic fourfolds with T-markings
$\mathcal{F}_1$	moduli space of cubic fourfolds of type T, which admits at worst ADE singularities
$\overline{\mathcal{F}}$	GIT quotient of $\mathcal{V}^{ss}$ by N, which is a compactification of $\mathcal{F}$
$\mathcal{M}$	moduli space of smooth cubic fourfolds
$\overline{\mathcal{M}}$	GIT compactification of $\mathcal{M}$
$(\Lambda_0)\Lambda$	(primitive) middle cohomology lattice of a smooth cubic fourfold
φ	topological intersection pairing on Λ
η	square of the hyperplane class
$\widehat{\mathbb{D}}$	local period domain for cubic fourfolds
$\widehat{\Gamma}$	monodromy group of the universal family of smooth cubic fourfolds
$\mathcal{H}_\Delta/\mathcal{H}_\infty$	$\widehat{\Gamma}$-invariant hyperplane arrangement in $\widehat{\mathbb{D}}$
ζ	character of A, induced by the action of A on $H^{3,1}(X)$
Λ_ζ	eigenspace of $(\Lambda_0)_{\mathbb{C}}$ corresponding to the character ζ

σ_X/σ	representation of A on $H^4(X,\mathbb{Z})/\Lambda$
h_X/h	Hermitian form on $H^4(X)_\zeta/\Lambda_\zeta$
$\mathbb{D}$	local period domain for cubic fourfolds of symmetry type T
Γ	normalizer of $\bar{A}$ in $\widehat{\Gamma}$
$\mathcal{H}_s/\mathcal{H}_*$	Γ-invariant hyperplane arrangements in $\mathbb{D}$
$\overline{\Gamma\backslash\mathbb{D}}^{\mathcal{H}_*}$	Looijenga compactification of $\Gamma\backslash(\mathbb{D}-\mathcal{H}_*)$
$\widetilde{\mathscr{P}}$	local period map
$\mathscr{P}$	global period map

Acknowledgements

Yu thanks his advisor Prof. S.-T. Yau for his constant support. He is grateful for the support from the Simons Collaboration on Homological Mirror Symmetry 2015. Zheng thanks his advisor E. Looijenga for his support and encouragement. Part of the work was done when he visited Stony Brook. His stay was supported by the Tsinghua Scholarship for Overseas Graduate Studies.

We thank E. Looijenga for stimulating discussion, especially on the appendix. We thank R. Laza and G. Pearlstein for helpful conversation. We thank the reviewer for helpful suggestions. We also had useful communications with Ruijie Yang and Zheng Zhang about this work.

References

[Allcock et al. 2011] D. Allcock, J. A. Carlson, and D. Toledo, *The moduli space of cubic threefolds as a ball quotient*, Mem. Amer. Math. Soc. **985**, Amer. Math. Soc., Providence, RI, 2011. MR Zbl

[Artebani et al. 2011] M. Artebani, A. Sarti, and S. Taki, "$K3$ surfaces with non-symplectic automorphisms of prime order", *Math. Z.* **268**:1-2 (2011), 507–533. MR Zbl

[Baily and Borel 1966] W. L. Baily, Jr. and A. Borel, "Compactification of arithmetic quotients of bounded symmetric domains", *Ann. of Math.* (2) **84** (1966), 442–528. MR Zbl

[Beauville and Donagi 1985] A. Beauville and R. Donagi, "La variété des droites d'une hypersurface cubique de dimension 4", *C. R. Acad. Sci. Paris Sér. I Math.* **301**:14 (1985), 703–706. MR Zbl

[Boissière et al. 2016] S. Boissière, C. Camere, and A. Sarti, "Classification of automorphisms on a deformation family of hyper-Kähler four-folds by p-elementary lattices", *Kyoto J. Math.* **56**:3 (2016), 465–499. MR Zbl

[Boissière et al. 2019] S. Boissière, C. Camere, and A. Sarti, "Complex ball quotients from manifolds of $K3^{[n]}$-type", *J. Pure Appl. Algebra* **223**:3 (2019), 1123–1138. MR Zbl

[Camere 2016] C. Camere, "Lattice polarized irreducible holomorphic symplectic manifolds", *Ann. Inst. Fourier* (*Grenoble*) **66**:2 (2016), 687–709. MR Zbl

[Deligne and Mostow 1986] P. Deligne and G. D. Mostow, "Monodromy of hypergeometric functions and nonlattice integral monodromy", *Inst. Hautes Études Sci. Publ. Math.* **63** (1986), 5–89. MR Zbl

[Dolgachev and Kondō 2007] I. V. Dolgachev and S. Kondō, "Moduli of $K3$ surfaces and complex ball quotients", pp. 43–100 in *Arithmetic and geometry around hypergeometric functions* (Istanbul, 2005), edited by R.-P. Holzapfel et al., Progr. Math. **260**, Birkhäuser, Basel, 2007. MR Zbl

[Fu 2016] L. Fu, "Classification of polarized symplectic automorphisms of Fano varieties of cubic fourfolds", *Glasg. Math. J.* **58**:1 (2016), 17–37. MR Zbl

[González-Aguilera and Liendo 2011] V. González-Aguilera and A. Liendo, "Automorphisms of prime order of smooth cubic n-folds", *Arch. Math.* (*Basel*) **97**:1 (2011), 25–37. MR Zbl

[Griffiths 1969a] P. A. Griffiths, "On the periods of certain rational integrals, I", *Ann. of Math.* (2) **90** (1969), 460–495. MR Zbl

[Griffiths 1969b] P. A. Griffiths, "On the periods of certain rational integrals, II", *Ann. of Math.* (2) **90** (1969), 496–541. MR Zbl

[Harris 1989] M. Harris, "Functorial properties of toroidal compactifications of locally symmetric varieties", *Proc. Lond. Math. Soc.* (3) **59**:1 (1989), 1–22. MR Zbl

[Hartshorne 1977] R. Hartshorne, *Algebraic geometry*, Grad. Texts in Math. **52**, Springer, 1977. MR Zbl

[Hassett 2000] B. Hassett, "Special cubic fourfolds", *Compositio Math.* **120**:1 (2000), 1–23. MR Zbl

[Höhn and Mason 2019] G. Höhn and G. Mason, "Finite groups of symplectic automorphisms of hyper-Kähler manifolds of type $K3^{[2]}$", *Bull. Inst. Math. Acad. Sin.* (*N.S.*) **14**:2 (2019), 189–264. MR Zbl

[Javanpeykar and Loughran 2017] A. Javanpeykar and D. Loughran, "Complete intersections: moduli, Torelli, and good reduction", *Math. Ann.* **368**:3-4 (2017), 1191–1225. MR Zbl

[Joumaah 2016] M. Joumaah, "Non-symplectic involutions of irreducible symplectic manifolds of $K3^{[n]}$-type", *Math. Z.* **283**:3-4 (2016), 761–790. MR Zbl

[Kiernan and Kobayashi 1972] P. Kiernan and S. Kobayashi, "Satake compactification and extension of holomorphic mappings", *Invent. Math.* **16** (1972), 237–248. MR Zbl

[Laza 2009] R. Laza, "The moduli space of cubic fourfolds", *J. Algebraic Geom.* **18**:3 (2009), 511–545. MR Zbl

[Laza 2010] R. Laza, "The moduli space of cubic fourfolds via the period map", *Ann. of Math.* (2) **172**:1 (2010), 673–711. MR Zbl

[Laza and Zheng 2019] R. Laza and Z. Zheng, "Automorphisms and periods of cubic fourfolds", preprint, 2019. arXiv

[Laza et al. 2018] R. Laza, G. Pearlstein, and Z. Zhang, "On the moduli space of pairs consisting of a cubic threefold and a hyperplane", *Adv. Math.* **340** (2018), 684–722. MR Zbl

[Looijenga 2003a] E. Looijenga, "Compactifications defined by arrangements, I: The ball quotient case", *Duke Math. J.* **118**:1 (2003), 151–187. MR Zbl

[Looijenga 2003b] E. Looijenga, "Compactifications defined by arrangements, II: Locally symmetric varieties of type IV", *Duke Math. J.* **119**:3 (2003), 527–588. MR Zbl

[Looijenga 2007] E. Looijenga, "Uniformization by Lauricella functions: an overview of the theory of Deligne–Mostow", pp. 207–244 in *Arithmetic and geometry around hypergeometric functions* (Istanbul, 2005), edited by R.-P. Holzapfel et al., Progr. Math. **260**, Birkhäuser, Basel, 2007. MR Zbl

[Looijenga 2009] E. Looijenga, "The period map for cubic fourfolds", *Invent. Math.* **177**:1 (2009), 213–233. MR Zbl

[Looijenga 2016] E. Looijenga, "Moduli spaces and locally symmetric varieties", pp. 33–75 in *Development of moduli theory* (Kyoto, 2013), edited by O. Fujino et al., Adv. Stud. Pure Math. **69**, Math. Soc. Japan, Tokyo, 2016. MR Zbl

[Looijenga and Swierstra 2007] E. Looijenga and R. Swierstra, "The period map for cubic threefolds", *Compos. Math.* **143**:4 (2007), 1037–1049. MR Zbl

[Luna 1973] D. Luna, "Slices étales", pp. 81–105 in *Sur les groupes algébriques*, Mém. Soc. Math. France **33**, Soc. Math. France, Paris, 1973. MR Zbl

[Luna 1975] D. Luna, "Adhérences d'orbite et invariants", *Invent. Math.* **29**:3 (1975), 231–238. MR Zbl

[Luna and Richardson 1979] D. Luna and R. W. Richardson, "A generalization of the Chevalley restriction theorem", *Duke Math. J.* **46**:3 (1979), 487–496. MR Zbl

[Matsumura and Monsky 1963] H. Matsumura and P. Monsky, "On the automorphisms of hypersurfaces", *J. Math. Kyoto Univ.* **3** (1963), 347–361. MR Zbl

[Mok and Zhong 1989] N. Mok and J. Q. Zhong, "Compactifying complete Kähler–Einstein manifolds of finite topological type and bounded curvature", *Ann. of Math.* (2) **129**:3 (1989), 427–470. MR Zbl

[Mongardi 2012] G. Mongardi, "Symplectic involutions on deformations of $K3^{[2]}$", *Cent. Eur. J. Math.* **10**:4 (2012), 1472–1485. MR Zbl

[Mongardi 2013] G. Mongardi, "On symplectic automorphisms of hyper-Kähler fourfolds of $K3^{[2]}$ type", *Michigan Math. J.* **62**:3 (2013), 537–550. MR Zbl

[Mongardi 2016] G. Mongardi, "Towards a classification of symplectic automorphisms on manifolds of $K3^{[n]}$ type", *Math. Z.* **282**:3-4 (2016), 651–662. MR Zbl

[Ressayre 2010] N. Ressayre, "Geometric invariant theory and the generalized eigenvalue problem", *Invent. Math.* **180**:2 (2010), 389–441. MR Zbl

[Vinberg and Popov 1994] E. B. Vinberg and V. L. Popov, "Invariant theory", pp. 123–278 in *Algebraic geometry, IV*, edited by A. N. Parshin and I. R. Shafarevich, Encyc. Math. Sci. **55**, Springer, 1994.

[Voisin 1986] C. Voisin, "Théorème de Torelli pour les cubiques de $\mathbb{P}^5$", *Invent. Math.* **86**:3 (1986), 577–601. Correction in **172**:2 (2008), 455–458. MR Zbl

[Yau 1978] S. T. Yau, "A general Schwarz lemma for Kähler manifolds", *Amer. J. Math.* **100**:1 (1978), 197–203. MR Zbl

[Yu and Zheng 2018] C. Yu and Z. Zheng, "Moduli of singular sextic curves via periods of $K3$ surfaces", preprint, 2018. arXiv

[Zarhin 1983] Y. G. Zarhin, "Hodge groups of $K3$ surfaces", *J. Reine Angew. Math.* **341** (1983), 193–220. MR Zbl

[Zheng 2019] Z. Zheng, "Orbifold aspects of certain occult period maps", *Nagoya Math. J.* (online publication November 2019).

Communicated by Gavril Farkas
Received 2019-05-04 Revised 2020-04-14 Accepted 2020-06-04

yucl18@math.upenn.edu *University of Pennsylvania, Philadelphia, PA, United States*

zhengzw11@mpim-bonn.mpg.de *Max Planck Institute for Mathematics, Bonn, Germany*

msp

ALGEBRA AND NUMBER THEORY 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2685

Motivic multiple zeta values relative to μ_2

Zhongyu Jin and Jiangtao Li

We establish a short exact sequence about depth-graded motivic double zeta values of even weight relative to μ_2. We find a basis for the depth-graded motivic double zeta values relative to μ_2 of even weight and a basis for the depth-graded motivic triple zeta values relative to μ_2 of odd weight. As an application of our main results, we prove Kaneko and Tasaka's conjectures about the sum odd double zeta values and the classical double zeta values. We also prove an analogue of Kaneko and Tasaka's conjecture in depth three. Finally, we formulate a conjecture which is related to sum odd multiple zeta values in higher depth.

1. Introduction

Multiple zeta values are defined by the convergent series

$$\zeta(n_1,\dots,n_r)=\sum_{0<k_1<\cdots<k_r}\frac{1}{k_1^{n_1}\cdots k_r^{n_r}},\qquad n_1,\dots,n_{r-1}>0,\ n_r>1.$$

In particular, when $r=1$ they are the classical Riemann zeta values. We call r the depth, and $N=n_1+\cdots+n_r$ the weight of the above multiple zeta value.

Denote by $\mathcal{Z}_N$ the $\mathbb{Q}$-vector space generated by all the weight N multiple zeta values for $N>0$, and $\mathcal{Z}_0=\mathbb{Q}$. Then $\mathcal{Z}_{N_1}\cdot\mathcal{Z}_{N_2}\subseteq\mathcal{Z}_{N_1+N_2}$, and

$$\mathcal{Z}=\mathcal{Z}_0+\mathcal{Z}_1+\cdots+\mathcal{Z}_N+\cdots$$

is an algebra with the shuffle product. The weight structure is conjectured to be a grading. There is a depth filtration $\mathcal{D}$ on $\mathcal{Z}_N$,

$$\mathcal{D}_r\mathcal{Z}_N=\langle\zeta(n_1,\dots,n_k)\in\mathcal{Z}_N,k\le r\rangle_{\mathbb{Q}},$$

where for $x_i\in\mathbb{R}, i\in A$, $\langle x_i,\ i\in A\rangle_{\mathbb{Q}}$ mean the $\mathbb{Q}$-linear subspace generated by $x_i, i\in A$ in $\mathbb{R}$.

The double zeta values generate the subspace $\mathcal{D}_2\mathcal{Z}$ of $\mathcal{Z}$. Gangl, Kaneko, Zagier [Gangl et al. 2006] found an interesting connection between period polynomials of $SL_2(\mathbb{Z})$ and the double shuffle relations among $\mathcal{D}_2\mathcal{Z}$.

Brown [2012] defined the motivic multiple zeta values algebra $\mathcal{H}$. Its elements can be written as $\mathbb{Q}$-linear combinations of motivic multiple zeta values $\zeta^{\mathfrak{m}}(n_1,\dots,n_r)$. There is a surjective algebra homomorphism:

$$\eta:\mathcal{H}\to\mathcal{Z},\quad \zeta^{\mathfrak{m}}(n_1,\dots,n_r)\mapsto\zeta(n_1,\dots,n_r).$$

MSC2010: primary 11F32; secondary 11F67.
Keywords: multiple zeta values, period polynomial, mixed Tate motives.

Brown proved that the set $\{\zeta^{\mathfrak{m}}(n_1,\dots,n_r), n_i \in 2,3\}$ is a basis for nonzero weight subspace of $\mathcal{H}$, thus proving that every multiple zeta value is a $\mathbb{Q}$-linear combination of $\zeta(n_1,\dots,n_r),\ n_i \in 2,3$ (Conjecture C in [Hoffman 1997]). Because of the period homomorphism, we can study the multiple zeta values by studying these motivic multiple zeta values.

Motivic multiple zeta values also satisfy the double shuffle relations by [Soudères 2010]. By Gangl, Kaneko, Zagier's results, there are also period polynomial relations among motivic double zeta values of even weight. This fact was reinterpreted as a short exact sequence which involves motivic double zeta values of even weight with a slight modification and period polynomials in the second author's paper [Li 2019].

Furthermore, Li [2019] proposed two exact sequence conjectures which relate the depth-graded version of motivic multiple zeta values and period polynomials of $\mathrm{SL}_2(\mathbb{Z})$. Li verified the two conjectures in low depth. Besides, Li and Liu [2020] established a short exact sequence about motivic double zeta values of odd weight.

After Brown, Glanois [2016] considered the motivic multiple zeta values relative to μ_N, which is a generalization of $\mathcal{H}$ for the cyclotomic field , where μ_N is the set of all N-th roots of unity. She gave a basis of motivic multiple zeta values relative to μ_N for $N = 2, 3, 4, 6, 8$. We will give a brief introduction to Glanois' work in the next section in the case of $N = 2$.

Ma [2015] studied motivic double zeta values relative to μ_N for $N = 2, 3$. He found various connections between some special matrices which come from motivic Galois action on motivic double zeta values relative to μ_N, Hecke operators and newforms of $\Gamma_0(N)$ for $N = 2, 3$.

In the rest of this paper, we only consider the motivic multiple zeta values relative to μ_2, and denote by $\mathcal{H}$ the $\mathbb{Q}$-algebra generated by them rather than the motivic multiple zeta algebra of Brown for convenience.

For positive integers $n_1 \geq 1, n_2 \geq 2$, define

$$\zeta^o(n_1,n_2) = \sum_{0<k_1<k_2,\mathrm{odd}} \frac{1}{k_1^{n_1}k_2^{n_2}}.$$

It is obvious to see

$$\zeta^o(n_1,n_2) = \frac{1}{4}\big(\zeta(n_1,n_2) - \zeta(\overline{n_1},n_2) - \zeta(n_1,\overline{n_2}) + \zeta(\overline{n_1},\overline{n_2})\big),$$

where

$$\zeta(\overline{n_1},n_2) = \sum_{0<k_1<k_2} \frac{(-1)^{k_1}}{k_1^{n_1}k_2^{n_2}}, \qquad \zeta(n_1,\overline{n_2}) = \sum_{0<k_1<k_2} \frac{(-1)^{k_2}}{k_1^{n_1}k_2^{n_2}}, \qquad \zeta(\overline{n_1},\overline{n_2}) = \sum_{0<k_1<k_2} \frac{(-1)^{k_1+k_2}}{k_1^{n_1}k_2^{n_2}}.$$

Denote by $\mathcal{Z}^2$ the $\mathbb{Q}$-vector space generated by

$$\zeta\begin{pmatrix} n_1,\dots,n_r \\ \epsilon_1,\dots,\epsilon_r \end{pmatrix} = \sum_{0<k_1<\cdots<k_r} \frac{\epsilon_1^{k_1}\cdots\epsilon_r^{k_r}}{k_1^{n_1}\cdots k_r^{n_r}}, \qquad \epsilon_i \in \{1,-1\},\ (n_r,\epsilon_r) \neq (1,1),$$

and we call $N = n_1 + \cdots + n_r$ and r its weight and depth respectively. Denote by $\mathcal{D}_r\mathcal{Z}^2$ the subspace of $\mathcal{Z}^2$ spanned by elements of depth $\leq r$. According to the previous equality, we have $\zeta^o(n_1,n_2) \in \mathcal{D}_2\mathcal{Z}^2$.

Similarly we have

$$\zeta^o(n_1, n_2, \dots, n_r) = \sum_{0<k_1<k_2<\cdots<k_r,\ \text{odd}} \frac{1}{k_1^{n_1} k_2^{n_2} \cdots k_r^{n_r}} = \frac{1}{2^r} \sum_{\epsilon_i \in \{\pm 1\}, 1 \le i \le r} \epsilon_1 \epsilon_2 \cdots \epsilon_r \zeta \binom{n_1, \dots, n_r}{\epsilon_1, \dots, \epsilon_r}.$$

We call $\zeta^o(n_1, n_2, \dots, n_r)$ a sum odd multiple zeta value.

In the case of even weight, Kaneko and Tasaka [2013] found the following result:

Theorem 1.1 (Kaneko–Tasaka). *For any even integer $N \ge 4$, denote by $\mathcal{S}_N(2)$ the space of cusp forms for $\Gamma_0(2) = \left\{\gamma \in SL_2(\mathbb{Z});\ \gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right),\ c \equiv 0 \bmod 2\right\}$ of weight N, then we have*

$$\dim_{\mathbb{Q}} \langle \zeta^o(2r, N-2r);\ 1 \le r \le N/2-1 \rangle_{\mathbb{Q}} \le N/2 - 1 - \dim_{\mathbb{C}} \mathcal{S}_N(2).$$

Besides, for N even, they also conjectured that elements

$$\zeta^o(2r-1, N-2r+1), \quad 1 \le r \le N/2-1$$

are $\mathbb{Q}$-linear independent, and each element $\zeta^o(r, N-r)$, $1 \le r \le N-2$ can be written as a $\mathbb{Q}$-linear combination of

$$\zeta^o(2r-1, N-2r+1), \quad 1 \le r \le N/2-1, \zeta(N).$$

In this paper we will reinterpret Kaneko and Tasaka results on the motivic level.

There are weight grading and depth filtration structures on $\mathcal{H}$ which are compatible with the usual weight and depth structures on classical multiple zeta values relative to μ_2. Denote by $gr_r^{\mathcal{D}}\mathcal{H}_N$ the weight N depth r part of the depth-graded motivic multiple zeta values, and

$$gr_\bullet \mathcal{H} = \mathbb{Q} \oplus \bigoplus_{r \ge 1} \mathcal{D}_r\mathcal{H}/\mathcal{D}_{r-1}\mathcal{H}.$$

Denote by $\zeta^{o,\mathfrak{m}}(n_1, n_2)$ the motivic sum odd double zeta value, which we will introduce later. Let $\mathcal{P}_{ev}^{o,\mathfrak{m}}$ be the space generated by the images of

$$\{\zeta^{o,\mathfrak{m}}(n_1, n_2),\ n_1, n_2 \ge 2, \text{even}\}$$

in the quotient $gr_2^{\mathcal{D}}\mathcal{H}$.

Our first main result (in a rough version) is the following:

Theorem 1.2. *There is an exact sequence with respect to sum odd motivic double zeta values*

$$0 \to \mathcal{P}_{ev}^{o,\mathfrak{m}} \to (gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \otimes gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}})^b \to \mathbb{P}(\Gamma_0(2))^\vee \to 0,$$

where $\mathcal{H}$ is the algebra of motivic multiple zeta values relative to μ_2.

Details of the above notations will be introduced in Sections 3 and 4. Theorem 1.2 gives a description of the space of motivic sum odd multiple zeta values of the form

$$\zeta^{\mathfrak{m},o}(n_1, n_2), \quad n_1, n_2 \ge 2, \text{even},$$

and from it we recover Theorem 1.1 immediately.

We can also discuss the case of odd n_1, n_2, and obtain the following theorem:

Theorem 1.3. (i) *For an even integer $N \geq 4$, the set of the images of*

$$\{\zeta^{o,\mathfrak{m}}(n_1, n_2), n_1 + n_2 = N,\ 1 \leq n_i \leq N-1, \textit{odd}\}$$

in $gr_2^{\mathcal{D}}\mathcal{H}$ is a basis for $gr_2^{\mathcal{D}}\mathcal{H}_N$.

(ii) *For an odd integer $N \geq 5$, the set of the images of*

$$\{\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3), n_1 + n_2 + n_3 = N,\ 1 \leq n_i \leq N-2, \textit{odd}\}$$

in $gr_3^{\mathcal{D}}\mathcal{H}$ is a basis for $gr_3^{\mathcal{D}}\mathcal{H}_N$.

In the above theorem, $\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3)$ means the motivic version of the sum odd multiple zeta value $\zeta^o(n_1, n_2, n_3)$. Its definition will be given in Section 2.

From the explicit calculations in the proof of Theorems 1.2 and 1.3, we also obtain the following theorem, which was conjectured in [Kaneko and Tasaka 2013, Section 3.2, Remark 2].

Theorem 1.4. (i) *For an even integer $N \geq 4$, the space*

$$\langle \zeta^o(r, N-r);\ 1 \leq r \leq N-2 \rangle_{\mathbb{Q}}$$

is spanned by

$$\{\zeta(N), \zeta^o(r, N-r);\ 1 \leq r \leq N-3, \textit{odd}\}.$$

(ii) *For an even integer $N \geq 6$, we have*

$$\big\langle \zeta(n_1, n_2);\ n_1 + n_2 = N, n_2 \geq 2 \rangle_{\mathbb{Q}} \subseteq \langle \zeta^o(n_1, n_2);\ n_1 + n_2 = N, n_2 \geq 2 \rangle_{\mathbb{Q}},$$
$$\langle \zeta^o(n_1, n_2);\ n_1 + n_2 = N, n_i \textit{ even} \rangle_{\mathbb{Q}} \subseteq \langle \zeta(n_1, n_2);\ n_1 + n_2 = N, n_2 \geq 2 \rangle_{\mathbb{Q}}.$$

We can also give some information in higher depth cases, as in the case of depth 3:

Theorem 1.5. *For a given odd integer $N \geq 5$, and $n_1 + n_2 + n_3 = N$, $n_1, n_2 \geq 1$, $n_3 \geq 2$, the element*

$$\zeta^o(n_1, n_2, n_3)$$

can be written as a $\mathbb{Q}$-linear combination of

$$\zeta^o(m_1, m_2, m_3), \quad m_1 + m_2 + m_3 = N, \quad m_1, m_2 \geq 1, m_3 \geq 3, m_i \textit{ odd}$$

and lower depth multiple zeta values relative to μ_2.

It seems that Theorem 1.3 should also be true for higher depth. We calculate the depth-graded motivic Galois action for sum odd motivic multiple zeta values explicitly in higher depth. Assuming the invertibility of a specific matrix, we can prove the higher depth analogue of Theorem 1.3.

Our paper is divided as follows. In Section 2A, we introduce mixed Tate motives over $\mathbb{Z}\big[\frac{1}{2}\big]$. In Section 2B, we introduce motivic multiple zeta values relative to μ_2, which was considered by Glanois [2016], following Deligne and Goncharov's work [2005]. We consider the motivic Galois action and show the way to do the calculation in Section 2C. Then we give a brief introduction to period polynomials in Section 3. The proofs of our main results will be given in Sections 4 and 5.

2. Motivic multiple zeta values relative to μ_2

As said in the introduction, the motivic multiple zeta values relative to μ_N are the generalization of Brown's motivic multiple zeta values. In this section, we only define them in the case of $N = 2$. The main references for this section are [Deligne 2010; Glanois 2016; Gil and Fresán 2018].

2A. ***Mixed Tate motives over*** $\mathbb{Z}\left[\frac{1}{2}\right]$**.** Consider the category of mixed Tate motives over $\mathbb{Z}\left[\frac{1}{2}\right]$; denote it by $\mathcal{MT}_2$. It is a Tannakian category with the natural fiber functor

$$\omega : \mathcal{MT}_2 \to \mathrm{Vec}_{\mathbb{Q}};\ M \mapsto \oplus \omega_r(M),$$

where

$$\omega_r(M) = \mathrm{Hom}_{\mathcal{MT}_2}(\mathbb{Q}(r),\, gr^{\omega}_{-2r}(M)).$$

Let $\mathcal{G}^{\mathcal{MT}_2}$ be the Tannakian fundamental group (the motivic Galois group) of $\mathcal{MT}_2$ with respect to this fiber functor ω, and $\mathcal{U}^{\mathcal{MT}_2}$ be the pro-unipotent radical of $\mathcal{G}^{\mathcal{MT}_2}$. We have

$$\mathcal{G}^{\mathcal{MT}_2} \cong \mathbb{G}_m \ltimes \mathcal{U}^{\mathcal{MT}_2}.$$

By Proposition 1.9 in [Deligne and Goncharov 2005], the extension group $\mathrm{Ext}^1_{\mathcal{MT}_2}(\mathbb{Q}(0), \mathbb{Q}(n))$ is nontrivial only when $n \geq 1$, odd and

$$\begin{aligned} &\mathrm{Ext}^1_{\mathcal{MT}_2}(\mathbb{Q}(0), \mathbb{Q}(n)) \cong \mathbb{Q}, && n \geq 1, \text{ odd}, \\ &\mathrm{Ext}^2_{\mathcal{MT}_2}(\mathbb{Q}(0), \mathbb{Q}(n)) = 0, && \text{for all } n. \end{aligned}$$

By Appendix A in [Deligne and Goncharov 2005], there is a set of symbols $\{f_{2n+1}; n \geq 0\}$ such that (noncanonical isomorphism)

$$\mathcal{O}(\mathcal{U}^{\mathcal{MT}_2}) \cong \mathbb{Q}\langle f_1, f_3, \ldots, f_{2n+1}, \ldots\rangle,$$

where $\mathbb{Q}\langle f_1, f_3, \ldots, f_{2n+1}, \ldots\rangle$ denotes the noncommutative polynomial ring with variables $f_1, f_3, \ldots, f_{2n+1}, \ldots$ under the shuffle product.

Let $\mathfrak{g}$ be the Lie algebra of $\mathcal{U}^{\mathcal{MT}_2}$, then $\mathfrak{g} = (m/m^2)^\vee$, where $m \subseteq \mathcal{O}(\mathcal{U}^{\mathcal{MT}_2})$ is the maximal ideal. It is a free Lie algebra with a set of generators $\{\sigma_{2n+1}; n \geq 0\}$.

Denote by ${}_0\Pi_1 = \pi_1^{dR}(\mathbb{P}^1 - \{0, 1, -1, \infty\}, \overrightarrow{1_0}, \overrightarrow{-1_1})$ the motivic torsor of paths from 0 to 1 on $\mathbb{P}^1 - \{0, \pm 1, \infty\}$, with tangential base point given by the tangent vectors 1 at 0 and -1 at 1. It is a functor. For any $\mathbb{Q}$-algebra R, denote by $R\langle\langle e_0, e_{-1}, e_1\rangle\rangle$ the noncommutative formal power series in e_0, e_{-1}, e_1 with coefficients in R and

$$\triangle : R\langle\langle e_0, e_{-1}, e_1\rangle\rangle \to R\langle\langle e_0, e_{-1}, e_1\rangle\rangle \otimes_R R\langle\langle e_0, e_{-1}, e_1\rangle\rangle$$

the co-product on $R\langle\langle e_0, e_{-1}, e_1\rangle\rangle$ satisfying $\triangle e_i = e_i \otimes 1 + 1 \otimes e_i$ for $i \in \{0, \pm 1\}$. Letting $R\langle\langle e_0, e_{-1}, e_1\rangle\rangle^\times$ be the set of nonzero elements of $R\langle\langle e_0, e_{-1}, e_1\rangle\rangle$, we have

$${}_0\Pi_1(R) = \{S \in R\langle\langle e_0, e_{-1}, e_1\rangle\rangle^\times;\ \triangle S = S \otimes S\},$$

i.e., ${}_0\Pi_1(R)$ is the set of group-like elements in $R\langle\langle e_0, e_{-1}, e_1\rangle\rangle^\times$.

Denote by e^i the canonical dual of e_i for $i \in \{0, 1, -1\}$. The affine ring of regular functions of ${}_0\Pi_1$ is the graded algebra with the shuffle product

$$\mathcal{O}({}_0\Pi_1) \cong \mathbb{Q}\langle e^0, e^1, e^{-1}\rangle.$$

The symbol ${}_01_1$ is the point ${}_01_1 : \operatorname{Spec} \mathbb{Q} \to {}_0\Pi_1$ whose function ring homomorphism maps every nonempty word in e^0, e^1, e^{-1} to 0.

More generally, for $x, y \in \{0, \pm 1\}$, denote by ${}_x\Pi_y$ the motivic fundamental groupoid from the tangential point $\vec{1}$ at x to the tangential point $\vec{1}$ at y.

Let V be the automorphism subgroup of the motivic fundamental groupoid (all basepoints are tangential points at $\{0, \pm 1\}$) of $\mathbb{P}^1 - \{0, \pm 1, \infty\}$ satisfying the following properties:

(i) Elements of V are compatible with the composition law on the motivic groupoid of $\mathbb{P}^1 - \{0, \pm 1, \infty\}$.

(ii) Elements of V fix $\exp(e_i) \in {}_i\Pi_i$ for $i \in \{0, \pm 1\}$.

(iii) Elements of V are equivariant with the $\{\pm 1\}$-action on the motivic groupoid.

By Proposition 5.11 in [Deligne 2010], the map

$$\xi : V \to {}_0\Pi_1, \quad a \mapsto a({}_01_1)$$

is an isomorphism of schemes and

$$\operatorname{Lie} V = (\mathbb{L}(e_0, e_1, e_{-1}), \{\ ,\ \}),$$

where $\mathbb{L}(e_0, e_1, e_{-1})$ is the free Lie algebra generated by the three symbols e_0, e_1, e_{-1}, and $\{\ ,\ \}$ denotes the Ihara Lie bracket on $\mathbb{L}(e_0, e_1, e_{-1})$.

The action of $\mathcal{U}^{\mathcal{MT}_2}$ on ${}_x\Pi_y$, $x, y \in \{0, \pm 1\}$ factors through V. So there is a natural Lie algebra homomorphism

$$i : \mathfrak{g} \to \operatorname{Lie} V = (\mathbb{L}(e_0, e_1, e_{-1}), \{\ ,\ \}).$$

By the main results of [Deligne 2010], the map i is injective.

For any element w in $\mathbb{L}(e_0, e_1, e_{-1})$, let $\operatorname{depth}(w)$ be the smallest number of total occurrences of e_1 and e_{-1} in w. It induces a depth filtration $\mathcal{D}$ on $\mathbb{L}(e_0, e_1, e_{-1})$ as follows:

$$\mathcal{D}^r\mathbb{L}(e_0, e_1, e_{-1}) = \{w \in \mathbb{L}(e_0, e_1, e_{-1}); \operatorname{depth}(w) \geq r\}.$$

According to [Deligne 2010], the map i satisfies

$$i(\sigma_1) = e_{-1}, \quad i(\sigma_{2n+1}) = (1 - 2^{2n})\operatorname{ad}(e_0)^{2n}e_{-1} + 2^{2n}\operatorname{ad}(e_0)^{2n}e_1 + \mathrm{HDT},$$

where HDT means the higher depth term.

The motivic Lie algebra $\mathfrak{g}$ has an induced depth filtration $\mathcal{D}^r\mathfrak{g}$ from the injective map i. Since the Ihara bracket is compatible with the depth filtration, we know that the depth-graded space

$$\mathfrak{dg} = \bigoplus_{r \geq 1} \mathcal{D}^r\mathfrak{g}/\mathcal{D}^{r+1}\mathfrak{g}$$

is a Lie algebra with the induced Ihara bracket. Furthermore, from the main results of [Deligne 2010], $\mathfrak{dg}$ is a free Lie algebra with generators

$$\overline{i(\sigma_1)} = e_{-1}, \quad \overline{i(\sigma_{2n+1})} = (1-2^{2n})\mathrm{ad}(e_0)^{2n}e_{-1} + 2^{2n}\mathrm{ad}(e_0)^{2n}e_1$$

in the depth one part.

We will use them in the style of Lie polynomial in $\mathbb{Q}\langle e_0, e_1, e_{-1}\rangle$ rather than Lie words in the rest of this paper for convenience:

$$i(\sigma_{2n+1}) = (1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{2n-r}e_{-1}e_0^r + 2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{2n-r}e_1e_0^r + \mathrm{HDT}. \tag{1}$$

2B. ***Motivic multiple zeta values.*** Similar to Brown's work, Glanois [2016] considered motivic iterated integral $I^{\mathfrak{m}}$ and motivic multiple zeta values $\zeta^{\mathfrak{m}}\binom{x_1,x_2,\ldots,x_p}{\epsilon_1,\epsilon_2,\ldots,\epsilon_p}$, $\epsilon_i \in \mu_N$ relative to the set of N-th roots of unity μ_N from [Deligne and Goncharov 2005]. We denote by $\mathcal{H} = \mathcal{H}^2$ the $\mathbb{Q}$-vector space of motivic multiple zeta values relative to $\mu_2 = \{1, -1\}$. Here we only give the definition in the case of $N = 2$.

Let us consider the map

$$dch : \mathbb{Q}\langle e^0, e^1, e^{-1}\rangle \to \mathbb{R}.$$

For words $u_i \in \{e^0, e^1, e^{-1}\}$, $i = 1, \ldots, k$ satisfying $u_1 \neq e^0$, $u_k \neq e^1$, define

$$dch(u_1\cdots u_k) = \int_{0<t_1<\cdots<t_k<1}\omega_{u_1}(t_1)\cdots\omega_{u_k}(t_k),$$

where $\omega_{e^0}(t) = dt/t$, $\omega_{e^i}(t) = dt/(i-t)$, $i \in \{1, -1\}$. By Appendix A in [Le and Murakami 1996], we know that

$$\int_{\varepsilon<t_1<\cdots<t_k<1-\eta}\omega_{u_1}(t_1)\cdots\omega_{u_k}(t_k) = P(\log(\varepsilon), \log(\eta)) + O(\sup(\varepsilon|\log(\epsilon)|^A + \eta|\log(\eta)|^B)),$$

where P is a polynomial. For a general word sequence $u_1\cdots u_k$, define

$$dch(u_1, \ldots, u_k) = P(0, 0).$$

By the shuffle product of iterated integral, dch is a $\mathbb{Q}$-algebra homomorphism. Since $\mathcal{O}({}_0\Pi_1)$ is an inductive object in the category of mixed Tate motives $\mathcal{MT}_2$, there is a natural action of $\mathcal{G}^{\mathcal{MT}_2}$ on $\mathbb{Q}\langle e^0, e^1, e^{-1}\rangle$. Denote by $\mathcal{I}$ the largest graded subideal of $\operatorname{Ker} dch$ which is stable under the action of $\mathcal{G}^{\mathcal{MT}_2}$. The motivic multiple zeta algebra $\mathcal{H}$ is $\mathcal{O}({}_0\Pi_1)/\mathcal{I}$.

Denote by $\mathcal{I}^{\mathfrak{m}}$ the natural quotient map

$$\mathcal{I}^{\mathfrak{m}} : \mathcal{O}({}_0\Pi_1) = \mathbb{Q}\langle e^0, e^1, e^{-1}\rangle \to \mathcal{H}$$

and by per the map $per : \mathcal{H} \to \mathbb{R}$ satisfying $per \circ \mathcal{I}^{\mathfrak{m}} = dch$.

The motivic multiple zeta value $\zeta^{\mathfrak{m}}\binom{x_1,x_2,\ldots,x_p}{\epsilon_1,\epsilon_2,\ldots,\epsilon_p}$ is

$$\mathcal{I}^{\mathfrak{m}}\big(e^{(\epsilon_1\cdots\epsilon_p)^{-1}}(e^0)^{x_1-1}e^{(\epsilon_2\cdots\epsilon_p)^{-1}}(e^0)^{x_2-1}\cdots e^{(\epsilon_p)^{-1}}(e^0)^{x_p-1}\big).$$

It's obvious to check that

$$per\left(\zeta^{\mathfrak{m}}\binom{x_1,\dots,x_p}{\epsilon_1,\dots,\epsilon_p}\right)=\zeta\binom{x_1,\dots,x_p}{\epsilon_1,\dots,\epsilon_p}.$$

Define $\zeta^{o,\mathfrak{m}}(n_1,\dots,n_r)$ as

$$\zeta^{o,\mathfrak{m}}(n_1,\dots,n_r)=\frac{1}{2^r}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\epsilon_1\cdots\epsilon_r\zeta^{\mathfrak{m}}\binom{n_1,\dots,n_r}{\epsilon_1,\dots,\epsilon_r}.$$

It's clear that the image of $\zeta^{o,\mathfrak{m}}(n_1,\dots,n_r)$ under the period map *per* is the sum odd multiple zeta values $\zeta^{o}(n_1,\dots,n_r)$.

In $\mathcal{O}({}_0\Pi_1)=\mathbb{Q}\langle e^0,e^1,e^{-1}\rangle$, for any word $u_1\cdots u_k$, $u_i\in\{e^0,e^1,e^{-1}\}$, k is called its weight and the total number of occurrences of e^1 and e^{-1} is called its depth. Denote by $\mathcal{D}_r\mathbb{Q}\langle e^0,e^1,e^{-1}\rangle$ the subspace spanned by elements of depth $\le r$.

Since the depth filtration on $\mathcal{O}({}_0\Pi_1)$ is motivic [Deligne and Goncharov 2005], it induces a natural depth filtration on $\mathcal{H}$. The depth filtration on $\mathcal{H}$ is compatible with the depth filtration on $\mathcal{Z}^2$ through the map *per*.

Denote by $gr_r^{\mathcal{D}}\mathcal{H}=\mathcal{D}_r\mathcal{H}/\mathcal{D}_{r-1}\mathcal{H}$. The following formula in the case of depth 1 follows from the main results in [Deligne and Goncharov 2005]. From this formula we can deduce a basis for $gr_1^{\mathcal{D}}\mathcal{H}$:

Lemma 2.1. (*Deligne–Goncharov*) *We have the distribution formula*

$$\zeta^{\mathfrak{m}}\binom{n}{-1}=(2^{-n+1}-1)\zeta^{m}\binom{n}{1},\quad \textit{for all } n\ge 2.$$

Lemma 2.2. (*Deligne–Goncharov*) *There is a basis of* $gr_1^{\mathcal{D}}\mathcal{H}$: $\left\{\zeta^{\mathfrak{m}}\binom{r}{-1},\ r\ge 1 \textit{ odd}\right\}$.

Remark 2.3. We will always write $\zeta^{\mathfrak{m}}\binom{n_1,n_2}{-1,1}$ as $\zeta^{\mathfrak{m}}(\overline{n_1},n_2)$, similarly $\zeta^{\mathfrak{m}}(n_1,\overline{n_2})$, $\zeta^{\mathfrak{m}}(\overline{n_1},\overline{n_2})$, $\zeta^{\mathfrak{m}}(\bar{k})$ for convenience.

2C. ***Motivic Galois action.*** In this subsection we explain how to calculate the depth-graded motivic Galois action of the motivic Lie algebra of $\mathcal{MT}\left(\mathbb{Z}\left[\frac{1}{2}\right]\right)$ on the motivic multiple zeta values relative to μ_2. Then we give the definition of the map ∂ and deduce its injectivity from the results of [Brown 2012; Deligne 2010; Glanois 2016].

Since the expression of $i(\sigma_{2n+1})$ in $(\mathbb{L}(e_0,e_1,e_{-1}),\{\ ,\ \})$ has canonical depth one part, σ_{2n+1} in $\mathfrak{g}=\operatorname{Lie}\mathcal{U}^{\mathcal{MT}_2}$ induces a well-defined derivation

$$\partial_{2n+1}: gr_r^{\mathcal{D}}\mathcal{H}\to gr_{r-1}^{\mathcal{D}}\mathcal{H}.$$

Denote by $gr_1^{\mathcal{D}}\mathcal{H}^{\mathrm{odd}}$ the subspace of $gr_1^{\mathcal{D}}\mathcal{H}$ generated by the images of

$$\zeta^{\mathfrak{m}}(\overline{1}),\zeta^{\mathfrak{m}}(\overline{3}),\dots,\zeta^{\mathfrak{m}}(\overline{2n+1}),\dots$$

in the quotient space $gr_1^{\mathcal{D}}\mathcal{H}$. In this section we will show how to calculate the map ∂_{2n+1} explicitly.

Since $\mathcal{O}({}_0\Pi_1)$ is an ind-object in the category $\mathcal{MT}_2$, there is an action of the motivic Lie algebra

$$\mathfrak{g}\times\mathcal{O}({}_0\Pi_1)\to\mathcal{O}({}_0\Pi_1).$$

Denote by $\mathfrak{h} = \operatorname{Lie} V = (\mathbb{L}(e_0, e_1, e_{-1}), \{\ ,\ \})$. The action of $\mathfrak{g}$ on $\mathcal{O}({}_0\Pi_1)$ factors through the action of $\mathfrak{h}$ on $\mathcal{O}({}_0\Pi_1)$.

Denote by $\mathcal{U}\mathfrak{h}$ the universal enveloping algebra of $\mathfrak{h}$. Then

$$\mathcal{U}\mathfrak{h} = (\mathbb{Q}\langle e_0, e_1, e_{-1}\rangle, \circ),$$

where $\circ$ denotes the new product on $\mathbb{Q}\langle e_0, e_1, e_{-1}\rangle$ transformed from the natural concatenation product on $\mathcal{U}\mathfrak{h}$.

The product $\circ$ is difficult to calculate in general, but by the same reasoning as Proposition 2.2 in [Brown 2013], for any $a \in \mathfrak{h}$, any words w in e_0, e_1, e_{-1}, and any $n \geq 0$, we have

$$a \circ (e_0^n e_i w) = e_0^n\big(([i]a)e_i + e_i([i]a)^*\big)w + e_0^n e_i(a \circ w), \quad i \in \{1, -1\}, \tag{2}$$

where

$$a \circ e_0^n = e_0^n a, \quad [i](e_0^a e_i^b e_{-i}^c \cdots) = e_0^a e_{-1}^b e_1^c \cdots, \quad (a_1 \cdots a_n)^* = (-1)^n (a_n \cdots a_1), \qquad a_i \in \{e_0, e_1, e_{-1}\}.$$

From the correspondence between unipotent algebraic group and nilpotent Lie algebra, we know that for any $a \in \mathfrak{h}$, the natural action of a on $\mathcal{O}({}_0\Pi_1)$,

$$\mathcal{O}({}_0\Pi_1) = \mathbb{Q}\langle e^0, e^1, e^{-1}\rangle \xrightarrow{a} \mathcal{O}({}_0\Pi_1) = \mathbb{Q}\langle e^0, e^1, e^{-1}\rangle, \quad x \mapsto a(x),$$

is dual to the following action of a on $\mathcal{U}\mathfrak{h}$:

$$\mathcal{U}\mathfrak{h} = \mathbb{Q}\langle e_0, e_1, e_{-1}\rangle \xrightarrow{a} \mathcal{U}\mathfrak{h} = \mathbb{Q}\langle e_0, e_1, e_{-1}\rangle, \quad y \mapsto a \circ y.$$

By the definition of $\mathcal{H}$ and ∂_{2n+1}, we have the commutative diagram

$$\begin{array}{ccc} gr_r^{\mathcal{D}}\mathbb{Q}\langle e^0, e^1, e^{-1}\rangle & \xrightarrow{\overline{\partial_{2n+1}}} & gr_{r-1}^{\mathcal{D}}\mathbb{Q}\langle e^0, e^1, e^{-1}\rangle \\ \downarrow & & \downarrow \\ gr_r^{\mathcal{D}}\mathcal{H} & \xrightarrow{\partial_{2n+1}} & gr_{r-1}^{\mathcal{D}}\mathcal{H} \end{array}$$

where $\overline{\partial_{2n+1}}$ is the depth-graded version of the action of $i(\sigma_{2n+1})$ on $\mathbb{Q}\langle e^0, e^1, e^{-1}\rangle$. Thus in order to write out the maps $\overline{\partial_{2n+1}}$ and ∂_{2n+1} clearly, we need to compute the action $\circ : \mathfrak{h} \times \mathcal{U}\mathfrak{h} \to \mathcal{U}\mathfrak{h}$ first.

There is a well-defined map

$$\partial : gr_r^{\mathcal{D}}\mathcal{H} \to gr_1^{\mathcal{D}}\mathcal{H}^{\mathrm{odd}} \otimes gr_{r-1}^{\mathcal{D}}\mathcal{H}; \ \ \partial = \sum_{n\geq 0} \zeta^{\mathfrak{m}}(\overline{2n+1}) \otimes \partial_{2n+1}.$$

The following proposition is crucial to our analysis.

Proposition 2.4. *For $r \geq 2$, the map ∂ is injective.*

Proof. By exactly the same method as that in [Brown 2012, Section 2.3], it follows that $\mathcal{H} \cong \mathcal{O}(\mathcal{U}^{\mathcal{MT}_2})[t]$ (t is a weight 2, depth 1 element with trivial action of $\mathfrak{g}$) as a $\mathfrak{g}$-module. Moreover $t^n, n \geq 1$ are all depth 1 elements.

So we have

$$gr_r\mathcal{H} \cong gr_r\mathcal{O}(\mathcal{U}^{\mathcal{MT}_2}) \oplus \bigoplus_{n\geq 1} gr_{r-1}\mathcal{O}(\mathcal{U}^{\mathcal{MT}_2})t^n.$$

It suffices to prove that $\partial|_{gr_r\mathcal{O}(\mathcal{U}^{\mathcal{MT}_2})}$ is injective. By the main results of [Deligne 2010], the depth-graded motivic Lie algebra $\mathfrak{dg}$ is a free Lie algebra with generators in the depth one part. By the correspondence between nilpotent Lie algebra and unipotent algebraic group, $\partial|_{gr_r\mathcal{O}(\mathcal{U}^{\mathcal{MT}_2})}$ is injective. □

Remark 2.5. Proposition 2.4 is not true for Brown's original motivic multiple zeta values, since, in that case, the depth-graded motivic Lie algebra of $\mathcal{MT}(\mathbb{Z})$ is not a free Lie algebra and it has generators in higher depth part. See [Brown 2013; Enriquez and Lochak 2016; Li 2020] for some conjectural descriptions of the depth-graded motivic Lie algebra of $\mathcal{MT}(\mathbb{Z})$.

3. Period polynomials

In this section, we review the theory of period polynomials, and define $\mathbb{P}(\Gamma_0(2))^\vee$ in Theorem 1.2. The main reference is [Kaneko and Tasaka 2013].

As we know, $\Gamma_0(2) = \{\gamma \in \mathrm{SL}_2(\mathbb{Z});\ \gamma \equiv 0 \bmod 2\}$ is generated by two elements

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad M = \begin{pmatrix} -1 & -1 \\ 2 & 1 \end{pmatrix}$$

For a positive even integer N, denote by V_N the space of homogeneous polynomials with two variables $X,\ Y$ of degree $N-2$:

$$V_N = \left\{ P(X,Y) \in \mathbb{Q}[X,Y];\ P(X,Y) = \sum_{i=0}^{N-2} a_i X^i Y^{N-2-i} \right\}.$$

The group $\Gamma = \Gamma_0(2)$ acts on V_N naturally: for any polynomial $P(X,Y) \in V_N$ and $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(2)$,

$$\gamma \circ P(X,Y) = P(aX + bY, cX + dY).$$

We write this action as $P(X,Y)|\gamma$ for convenience. Consider the subspace W_N of V_N as follows:

$$W_N = \{P(X,Y) \in V_N[X,Y];\ P|(1-T)(1+M) = 0\}.$$

Denote by $\mathcal{S}_N(2)$ the space of cusp forms of weight N for $\Gamma_0(2)$. For $f \in \mathcal{S}_N(2)$, the period polynomial $r_f(X,Y)$ of f is given by

$$r_f(X,Y) = \int_0^\infty f(z)(X - zY)^{N-2}\, dz.$$

It can be shown that

$$r_f(X,Y) \in W_N \otimes \mathbb{C}.$$

Now we decompose W_N into two parts. Let $\varepsilon = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. It is obvious to see that $P|(1 \pm \varepsilon) \in W_N^{\pm}$ for any $P \in W_N$, thus we have the direct sum decomposition

$$W_N = W_N^+ \oplus W_N^-,$$

where W_N^+ (resp. W_N^-) is the even (resp. odd) part of W_N,

$$W_N^{\pm} = \{P \in V_N;\ P|\varepsilon = \pm P,\ P|(1-T)(1+M) = 0\}.$$

For $f \in S_N(2)$, denote by $r_f^{\pm}$ the even and odd parts of the map r_f,

$$r^{\pm} : S_N(2) \to W_N^{\pm} \otimes \mathbb{C};\ f \mapsto r_f^{\pm}(X, Y).$$

We can decompose W_N^+ further as

$$W_N^+ = \mathbb{Q} \cdot X^{N-2} \oplus W_N^{+,0} \oplus \mathbb{Q} \cdot Y^{N-2},$$

where

$$W_N^{+,0} = \left\{ P(X,Y) = \sum_{i=2\ even}^{N-4} a_i X^i Y^{N-2-i} \in V_N;\ P|(1-T)(1+M) = 0 \right\}.$$

Kaneko and Tasaka [2013] proved the following two propositions which describe the structure of $W_N^{+,0}$:

Proposition 3.1. *For any even integer N, there are two isomorphisms of vector spaces*

$$r^+ : S_N(2) \to W_N^{+,0} \otimes \mathbb{C} \quad \textit{and} \quad r^- : S_N(2) \to W_N^- \otimes \mathbb{C}.$$

Proposition 3.2. *For any even integer n, k, denote by $\binom{n}{k}$ the binomial coefficient. The space $W_N^{+,0}$ is of the form*

$$\left\{ \sum_{i=2\ even}^{N-4} a_i X^i Y^{N-2-i};\ \sum_{i=2\ even}^{N-4} \left(\binom{i}{j} - \binom{i}{N-2-j} \right) a_{N-2-j} = 0,\ 1 \le j \le N-3, \textit{odd} \right\}.$$

Remark 3.3. We let $\binom{n}{k} = 0$ when $k > n$ or $k < 0$.

Denote by $\mathbb{P}(\Gamma_0(2)) = \bigoplus_{N\ \mathrm{even}} W_N^{+,0}$. Then $\mathbb{P}(\Gamma_0(2))^\vee$ in Theorem 1.2 is the compact dual of $\mathbb{P}(\Gamma_0(2))$.

4. The depth two case

In this section we calculate the map

$$\partial : gr_r^{\mathcal{D}} \mathcal{H} \to gr_1^{\mathcal{D}} \mathcal{H}^{\mathrm{odd}} \otimes gr_{r-1}^{\mathcal{D}} \mathcal{H}$$

in the case of $r = 2$ explicitly. Then we establish a short exact sequence about sum odd motivic double zeta values and we find a basis for the depth-graded motivic double zeta values relative to μ_2 by the explicit expression of the map ∂ in the case of $r = 2$. As an application of our results, we prove Kaneko and Tasaka's conjectures [2013, Remark 2].

4A. ***The calculation in depth two.*** The following formulas come from direct calculation. We write $\overline{i(\sigma_{2n+1})}$ as $\overline{\sigma_{2n+1}}$ for short. When $n=0,\ \overline{\sigma_{2n+1}}=\overline{\sigma_1}$, $a_1\geq 0$, we have

$$\overline{\sigma_1}\circ(e_0^{a_0}e_1e_0^{a_1})=e_0^{a_0}e_{-1}e_1e_0^{a_1}-e_0^{a_0}e_1e_{-1}e_0^{a_1}+e_0^{a_0}e_1e_0^{a_1}e_{-1}, \tag{3}$$

$$\overline{\sigma_1}\circ(e_0^{a_0}e_{-1}e_0^{a_1})=e_0^{a_0}e_1e_{-1}e_0^{a_1}-e_0^{a_0}e_{-1}e_1e_0^{a_1}+e_0^{a_0}e_{-1}e_0^{a_1}e_{-1}. \tag{4}$$

When $n>0,\ a_1\geq 0$, we have

$$\begin{aligned}
&\overline{\sigma_{2n+1}}\circ(e_0^{a_0}e_1e_0^{a_1})\\
&=e_0^{a_0}(\overline{\sigma_{2n+1}}e_1-e_1\overline{\sigma_{2n+1}})e_0^{a_1}+e_0^{a_0}e_1e_0^{a_1}\overline{\sigma_{2n+1}}\\
&=(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_0^{2n-r}e_{-1}e_0^re_1e_0^{a_1}+2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_0^{2n-r}e_1e_0^re_1e_0^{a_1}\\
&\quad-(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_1e_0^re_{-1}e_0^{2n-r}e_0^{a_1}-2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_1e_0^re_1e_0^{2n-r}e_0^{a_1}\\
&\quad+(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_1e_0^{a_1}e_0^{2n-r}e_{-1}e_0^r+2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_1e_0^{a_1}e_0^{2n-r}e_1e_0^r,
\end{aligned} \tag{5}$$

and

$$\begin{aligned}
&\overline{\sigma_{2n+1}}\circ(e_0^{a_0}e_{-1}e_0^{a_1})\\
&=e_0^{a_0}\big([-1](\overline{\sigma_{2n+1}})e_{-1}-e_{-1}[-1](\overline{\sigma_{2n+1}})\big)e_0^{a_1}+e_0^{a_0}e_{-1}e_0^{a_1}\overline{\sigma_{2n+1}}\\
&=(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_0^{2n-r}e_1e_0^re_{-1}e_0^{a_1}+2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_0^{2n-r}e_{-1}e_0^re_{-1}e_0^{a_1}\\
&\quad-(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_{-1}e_0^re_1e_0^{2n-r}e_0^{a_1}-2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_{-1}e_0^re_{-1}e_0^{2n-r}e_0^{a_1}\\
&\quad+(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_{-1}e_0^{a_1}e_0^{2n-r}e_{-1}e_0^r+2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{a_0}e_{-1}e_0^{a_1}e_0^{2n-r}e_1e_0^r.
\end{aligned} \tag{6}$$

By taking dual of formula (1) and (2) we have the following result:

Lemma 4.1. *For positive even integers n_1, n_2 and $\epsilon_1,\epsilon_2\in\{1,-1\}$,*

$$\overline{\partial_1}(e^{\epsilon_1}(e^0)^{n_1-1}e^{\epsilon_2}(e^0)^{n_2-1})=0.$$

Proof. We calculate the map by taking dual of the action $\overline{\sigma_{2n+1}}$, thus we only need to find the terms $e_{\epsilon_1}(e_0)^{n_1-1}e_{\epsilon_2}(e_0)^{n_2-1}$ on the right-hand sides of equations (3) and (4). However, there are no such terms because n_1, n_2 are both even and thus $n_1-1,\ n_2-1$ are odd, it means that there is at least one e_0 between e_{ϵ_1} and e_{ϵ_2}, and one e_0 after e_{ϵ_2}. It follows that $\partial_1(e^{\epsilon_1}(e^0)^{n_1-1}e^{\epsilon_2}(e^0)^{n_2-1})=0$ for all n_1, n_2 even. □

Lemma 4.2. *For positive even integers n_1, n_2 and $n > 0$, writing the words*

$$\binom{2n}{n_1-1}e^1(e^0)^{n_1+n_2-2n-2}, \qquad \binom{2n}{n_1-1}e^{-1}(e^0)^{n_1+n_2-2n-2},$$

$$\binom{2n}{n_2-1}e^1(e^0)^{n_1+n_2-2n-2}, \qquad \binom{2n}{n_2-1}e^{-1}(e^0)^{n_1+n_2-2n-2},$$

as $\Theta_1^{n_1}$, $\Theta_{-1}^{n_1}$, $\Theta_1^{n_2}$, $\Theta_{-1}^{n_2}$ respectively for convenience, we have

$$\overline{\partial_{2n+1}}(e^1(e^0)^{n_1-1}e^1(e^0)^{n_2-1}) = 2^{2n}(\Theta_1^{n_1} - \Theta_1^{n_2}), \tag{7}$$

$$\overline{\partial_{2n+1}}(e^1(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}) = (1-2^{2n})(\Theta_1^{n_1} - \Theta_1^{n_2}), \tag{8}$$

$$\overline{\partial_{2n+1}}(e^{-1}(e^0)^{n_1-1}e^1(e^0)^{n_2-1}) = (1-2^{2n})\Theta_{-1}^{n_1} - 2^{2n}\Theta_{-1}^{n_2}, \tag{9}$$

$$\overline{\partial_{2n+1}}(e^{-1}(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}) = 2^{2n}\Theta_{-1}^{n_1} - (1-2^{2n})\Theta_{-1}^{n_2}. \tag{10}$$

Proof. To calculate the term

$$\overline{\partial_{2n+1}}(e^1(e^0)^{n_1-1}e^1(e^0)^{n_2-1}),$$

we only need to find the term

$$e_1(e_0)^{n_1-1}e_1(e_0)^{n_2-1}$$

on the right-hand sides of equations (3) and (4). The only two possibilities are $a_0 = 0, r = n_1 - 1$ or $a_0 = 0, r = n_2 - 1$. Thus we have

$$\overline{\partial_{2n+1}}(e^1(e^0)^{n_1-1}e^1(e^0)^{n_2-1}) = 2^{2n}\left[\binom{2n}{n_1-1} - \binom{2n}{n_2-1}\right]e^1(e^0)^{n_1+n_2-2n-2}.$$

Formula (5) is proved. By the same method we have (6)–(8). □

It is also useful for us to determine the case that n_1, n_2 are both odd. We use the same argument here and the result is a little different.

Lemma 4.3. *For positive odd integers $n_1 \geq 1, n_2 \geq 1$, we have*

$$\overline{\partial_1}(e^{i_1}(e^0)^{n_1-1}e^{i_2}(e^0)^{n_2-1}) = \begin{cases} e^{-1}, & \text{if } (i_1,i_2) = (1,-1),\ n_1 = 1, n_2 = 1, \\ -e^1(e^0)^{n_2-1} + e^{-1}(e^0)^{n_2-1}, & \text{if } (i_1,i_2) = (1,-1),\ n_1 = 1, n_2 \geq 3, \\ e^1(e^0)^{n_2-1} - e^{-1}(e^0)^{n_2-1}, & \text{if } (i_1,i_2) = (-1,1),\ n_1 = 1, n_2 \geq 1, \\ e^1(e^0)^{n_1-1}, & \text{if } (i_1,i_2) = (1,-1),\ n_1 \geq 3, n_2 = 1, \\ e^{-1}(e^0)^{n_1-1}, & \text{if } (i_1,i_2) = (-1,-1),\ n_1 \geq 1, n_2 = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Define $\delta\binom{m}{n} = 1$ if $m = n$, $\delta\binom{m}{n} = 0$ if $m \neq n$.

Lemma 4.4. *For positive odd integers n_1, n_2, let $\Theta_1^{n_1}$, $\Theta_{-1}^{n_1}$, $\Theta_1^{n_2}$, $\Theta_{-1}^{n_2}$ be as above and $n \geq 1$, we have*

$$\overline{\partial_{2n+1}}(e^1(e^0)^{n_1-1}e^1(e^0)^{n_2-1}) = -2^{2n}\Big((\Theta_1^{n_1} - \Theta_1^{n_2}) - \delta\binom{2n}{n_1-1}\Theta_1^{n_1}\Big),$$

$$\overline{\partial_{2n+1}}(e^1(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}) = -(1-2^{2n})\Big((\Theta_1^{n_1} - \Theta_1^{n_2}) - \delta\binom{2n}{n_1-1}\Theta_{-1}^{n_1}\Big),$$

$$\overline{\partial_{2n+1}}(e^{-1}(e^0)^{n_1-1}e^1(e^0)^{n_2-1}) = -(1-2^{2n})\Big(\Theta_{-1}^{n_1} - \delta\binom{2n}{n_1-1}\Theta_1^{n_1}\Big) + 2^{2n}\Theta_{-1}^{n_2},$$

$$\overline{\partial_{2n+1}}(e^{-1}(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}) = -2^{2n}\Big(\Theta_{-1}^{n_1} - \delta\binom{2n}{n_1-1}\Theta_{-1}^{n_1}\Big) + (1-2^{2n})\Theta_{-1}^{n_2}.$$

With the above lemmas, we can calculate the maps $\bar{\partial}$ and ∂ in the case of $r = 2$.

4B. ***Proofs of the main results.*** Now we are ready to state our main results. We have already defined the map ∂_{2n+1} for $n \geq 0$ in Section 2C and the space $\mathcal{P}_{ev}^{o,\mathfrak{m}}$ in Section 1, which is the subspace of $gr_2^{\mathcal{D}}\mathcal{H}$ generated by the set of images of $\{\zeta^{o,\mathfrak{m}}(n_1, n_2),\ n_1, n_2 \geq 2, \text{even}\}$. Define

$$D : gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \otimes gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \to (\mathbb{P}(\Gamma_0(2)))^{\vee},$$

$$\zeta^{\mathfrak{m}}(\overline{2n_1+1}) \otimes \zeta^{\mathfrak{m}}(\overline{2n_2+1}) \mapsto \frac{2^{2n_2}-1}{2^{2n_2+1}-1} \cdot v(2n_1+1, 2n_2+1), \quad n_1, n_2 \geq 0,$$

where $v(2n_1+1, 2n_2+1)$ is a linear functional on $\mathbb{P}(\Gamma_0(2))$ satisfying

$$v(2n_1+1, 2n_2+1)(p) = p_{2n_1,2n_2}$$

for

$$p = \sum p_{2m_1,2m_2} X^{2m_1} Y^{2m_2} \in \mathbb{P}(\Gamma_0(2)).$$

Theorem 4.5. *Denote by $(gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \otimes gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}})^b$ the subspace of $gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \otimes gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}}$ which is generated by $\zeta^{\mathfrak{m}}(\bar{n}_1) \otimes \zeta^{\mathfrak{m}}(\bar{n}_2),\ n_1, n_2 \geq 3$, odd. Then*

$$\partial(\mathcal{P}_{ev}^{o,\mathfrak{m}}) \subseteq (gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \otimes gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}})^b$$

and there is an exact sequence

$$0 \to \mathcal{P}_{ev}^{o,\mathfrak{m}} \xrightarrow{\tilde{\partial}} (gr_1^{\mathcal{D}}\mathcal{H}^{odd} \otimes gr_1^{\mathcal{D}}\mathcal{H}^{odd})^b \xrightarrow{\widetilde{D}} \mathbb{P}(\Gamma_0(2))^{\vee} \to 0,$$

where the second map $\tilde{\partial}$ is induced from $\partial|_{\mathcal{P}_{ev}^{o,\mathfrak{m}}}$ and the third map $\widetilde{D}$ is induced from D defined as above.

Proof. By Lemmas 4.1 and 4.2 it's obvious to check that

$$\partial(\mathcal{P}_{ev}^{o,\mathfrak{m}}) \subseteq (gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \otimes gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}})^b.$$

The map $\tilde{\partial}$ is injective by Proposition 2.4. The surjectivity of $\widetilde{D}$ is trivial. We only need to show that $\operatorname{Im}\tilde{\partial} = \operatorname{Ker}\widetilde{D}$.

The following diagram is commutative:

$$\begin{array}{ccc} gr_2^{\mathcal{D}}\mathbb{Q}\langle e^0,e^1,e^{-1}\rangle & \xrightarrow{\bar{\partial}} & (\mathfrak{dg}_1)^{\vee}\otimes gr_1^{\mathcal{D}}\mathbb{Q}\langle e^0,e^1,e^{-1}\rangle \\ \downarrow & & \downarrow \\ gr_2^{\mathcal{D}}\mathcal{H} & \xrightarrow{\partial} & gr_1^{\mathcal{D}}\mathcal{H}\otimes gr_1^{\mathcal{D}}\mathcal{H} \end{array}$$

where

$$\bar{\partial}=\sum_{n\geq 0}(\overline{\sigma_{2n+1}})^{\vee}\otimes\bar{\partial}_{2n+1},$$

and $(\overline{\sigma_{2n+1}})^{\vee}$, $n\geq 0$, is the dual basis of $\overline{\sigma_{2n+1}}$, $n\geq 0$, in $(\mathfrak{dg}_1)^{\vee}$. The second column map transforms $(\overline{\sigma_{2n+1}})^{\vee}\otimes\bar{\partial}_{2n+1}(x)$ to $\zeta^{\mathfrak{m}}(\overline{2n+1})\otimes\partial_{2n+1}(\mathcal{I}^{\mathfrak{m}}(x))$.

Thus we can calculate the image of $\mathcal{P}_{ev}^{o,\mathfrak{m}}$ under ∂ by calculating its lift on

$$gr_2\mathbb{Q}\langle e^0,e^1,e^{-1}\rangle.$$

For even integers n_1,n_2, the motivic double zeta value $\zeta^{o,\mathfrak{m}}(n_1,n_2)$ regarded as an element of $gr_2\mathcal{H}$ is equal to

$$\frac{1}{4}\mathcal{I}^{\mathfrak{m}}\big(e^1(e^0)^{n_1-1}e^1(e^0)^{n_2-1}-e^{-1}(e^0)^{n_1-1}e^1(e^0)^{n_2-1}-e^{-1}(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}+e^1(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}\big).$$

Define $\Lambda(n_1,n_2)$ to be

$$\frac{1}{4}\big[e^1(e^0)^{n_1-1}e^1(e^0)^{n_2-1}-e^{-1}(e^0)^{n_1-1}e^1(e^0)^{n_2-1}-e^{-1}(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}+e^1(e^0)^{n_1-1}e^{-1}(e^0)^{n_2-1}\big].$$

Let $\Theta_i^{n_k}$ be as above, $n>0$ and $s=n_1+n_2-2n-2$. According to Lemmas 4.1 and 4.2 we have

$$\overline{\partial_{2n+1}}(\Lambda(n_1,n_2))=\frac{1}{4}\big(\Theta_1^{n_1}-\Theta_1^{n_2}-\Theta_{-1}^{n_1}+\Theta_{-1}^{n_2}\big)=\frac{1}{4}\Big(\binom{2n}{n_1-1}-\binom{2n}{n_2-1}\Big)(e^1(e^0)^s-e^{-1}(e^0)^s).$$

By Lemmas 2.1 and 2.2, if $s>0$, we have

$$\partial_{2n+1}(\zeta^{o,\mathfrak{m}}(n_1,n_2))=\frac{1-2^{s+1}}{4(2^s-1)}\Big(\binom{2n}{n_1-1}-\binom{2n}{n_2-1}\Big)\zeta^{\mathfrak{m}}(\overline{s+1}).$$

Combining with the definition of $\widetilde{D}$ and Proposition 3.2 , it is obvious that $\mathrm{Im}\,(\tilde{\partial})=\mathrm{Ker}\,(\widetilde{D})$. □

Kaneko and Tasaka [2013] proved that there are at least $\dim\mathcal{S}_N(\Gamma_0(2))$-linear independent relations among the numbers $\{\zeta^o(n_1,n_2),n_1+n_2=N,n_1,n_2\geq 2,\text{even}\}$. From Theorem 4.5 we obtain

$$\dim_{\mathbb{Q}}\langle\zeta^o(n_1,n_2);n_1+n_2=N,n_1,n_2\geq 2,\text{even}\rangle_{\mathbb{Q}}\leq\frac{N}{2}-1-\dim\mathcal{S}_N(\Gamma_0(2))$$

immediately. The above inequality is compatible with Kaneko and Tasaka's result.

The next theorem gives an affirmative answer for part of Kaneko and Tasaka's conjectures in the motivic setting.

Theorem 4.6. *For an even integer $N \geq 4$. The elements*

$$\{\zeta^{o,\mathfrak{m}}(k, N-k), 1 \leq k \leq N-1 \ odd\}$$

are $\mathbb{Q}$-linear independent. Moreover, the set of their images in $gr_2^{\mathcal{D}}\mathcal{H}$ is a basis of $gr_2^{\mathcal{D}}\mathcal{H}_N$.

Proof. We will make use of the above calculations again. The case $N=4$ is easy to check. Given an even integer $N \geq 6$, according to Lemmas 4.3 and 4.4, for any odd n_1, n_2 such that $n_1+n_2=N$, we have for all $n_1, n_2 > 1$, $\partial_1\big(\zeta^{\mathfrak{m}}\binom{n_1,n_2}{\epsilon_1,\epsilon_2}\big)=0$ and

$$\partial_1(\zeta^{o,\mathfrak{m}}(1,n_2))=\frac{1}{2}\mathcal{I}^{\mathfrak{m}}\big[-e^1(e^0)^{n_2-1}+e^{-1}(e^0)^{n_2-1}\big], \quad \partial_1(\zeta^{o,\mathfrak{m}}(n_1,1))=\frac{1}{4}\mathcal{I}^{\mathfrak{m}}\big[e^1(e^0)^{n_1-1}-e^{-1}(e^0)^{n_1-1}\big].$$

Thus by the distribution formula, we have

$$\partial_1(\zeta^{o,\mathfrak{m}}(1,n_2))=\frac{1-2^{n_2}}{2-2^{n_2}}\zeta^{\mathfrak{m}}(\overline{n_2}), \quad \partial_1(\zeta^{o,\mathfrak{m}}(n_1,1))=-\frac{1-2^{n_1}}{4(1-2^{n_1-1})}\zeta^{\mathfrak{m}}(\overline{n_1}).$$

For the same reason, if $n \geq 1$ and $s=n_1+n_2-2n-2>0$, the following formula holds:

$$\partial_{2n+1}(\zeta^{o,\mathfrak{m}}(n_1,n_2))=\frac{2^{s+1}-1}{4(2^s-1)}\Big[\binom{2n}{n_1-1}-\binom{2n}{n_2-1}+\delta\binom{2n}{n_1-1}(1-2^{2n+1})\Big]\zeta^{\mathfrak{m}}(\overline{s+1}).$$

If $n \geq 1$ and $s=n_1+n_2-2n-2=0$, we have

$$\partial_{2n+1}(\zeta^{o,\mathfrak{m}}(n_1,n_2))=-\frac{1}{4}\delta\binom{2n}{n_1-1}(2^{2n+1}-1)\zeta^{\mathfrak{m}}(\overline{1}).$$

In conclusion, we can write the map ∂ in the following form in the case of $n_1+n_2=N$:

$$\partial\begin{pmatrix}\zeta^{o,\mathfrak{m}}(1,N-1)\\ \zeta^{o,\mathfrak{m}}(3,N-3)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-1,1)\end{pmatrix}=\widetilde{M}B\begin{pmatrix}\zeta^{\mathfrak{m}}(\overline{1})\otimes\zeta^{\mathfrak{m}}(\overline{N-1})\\ \zeta^{\mathfrak{m}}(\overline{3})\otimes\zeta^{\mathfrak{m}}(\overline{N-3})\\ \vdots\\ \zeta^{\mathfrak{m}}(\overline{N-1})\otimes\zeta^{\mathfrak{m}}(\overline{1})\end{pmatrix}.$$

In the above formula,

$$B=\mathrm{diag}\left(\frac{1-2^{N-1}}{2-2^{N-1}},\frac{2^{N-3}-1}{4(2^{N-4}-1)},\dots,\frac{2^3-1}{4(2^2-1)},-\frac{(2^{N-1}-1)}{4}\right)$$

is a $\big(\frac{N}{2}\big)$-th invertible diagonal matrix, $\widetilde{M}$ is a square matrix of order $\frac{N}{2}$ in the form

$$\widetilde{M}=\begin{pmatrix}1 & \cdots & 0\\ 0 & & 0\\ \vdots & M & \vdots\\ 0 & & 0\\ c & \cdots & 1\end{pmatrix},$$

where M is an $\big(\frac{N}{2}-2\big)$-th square matrix in the middle of $\widetilde{M}$. The matrix M equals $(a_{i,j})_{1\leq i,j\leq\frac{N}{2}-2}$, where

$$a_{i,j}=\binom{2j}{2i}-\binom{2j}{N-2-2i}+\delta\binom{2i}{2j}(1-2^{2j+1}).$$

The theorem holds if M is invertible by the fact that ∂ is injective. M can be written as the form $D+A$, where $t=\frac{N}{2}-1$,

$$D=\operatorname{diag}(d_1,\dots,d_{t-1}),\quad A=(b_{i,j})_{1\le i,j\le t-1},\quad \text{and}\quad d_i=1-2^{2i+1},\quad b_{i,j}=\binom{2j}{2i}-\binom{2j}{2t-2i}.$$

Given j, notice that $b_{i,j}+b_{t+1-i,j}=0$ and $b_{i,j}=0$ for $j<i<t-j$, it's obvious to check that

$$\sum_{i=1}^{t-1}|b_{i,j}|=2\sum_{i=1}^{\min\{\frac{t-1}{2},\ j-1\}}|b_{i,j}|\le 2\sum_{i=1}^{\min\{\frac{t-1}{2},\ j-1\}}\binom{2j}{2i}\le 2\sum_{i=1}^{j-1}\binom{2j}{2i}<2^{2j+1}-1.$$

So clearly for $j=1,\dots,t-1$, we have

$$\sum_{i=1,i\neq j}^{t-1}|b_{i,j}|=\sum_{i=1}^{t-1}|b_{i,j}|-|b_{j,j}|<|d_j+b_{j,j}|.$$

By the following lemma, the matrix M, and furthermore $\widetilde{M}$ are invertible. □

Lemma 4.7. *For a real matrix $A=(a_{i,j})_{1\le i,j\le n}$, if $|a_{i,i}|>\sum_{i\neq j}|a_{i,j}|$ for $i=1,2,\dots,n$, then $|A|\neq 0$.*

Proof. Denote by α_i the i-th column vector of A, if $|A|=0$, there exist $\{k_1,\dots,k_n\}\neq\{0\}$ such that $k_1\alpha_1+\cdots+k_n\alpha_n=0$ is the zero column vector.

Let

$$|k_l|=\max\{|k_1|,\dots,|k_n|\}.$$

Now consider the l-th variable in the above zero column vector. Because we have that $|a_{l,l}|>\sum_{l\neq j}|a_{l,j}|$, $k_1a_{l,1}+\cdots+k_na_{l,n}\neq 0$, we get a contradiction. □

Remark 4.8. Kaneko and Tasaka [2013] conjectured that for given $N\ge 4$, elements

$$\{\zeta^o(n_1,n_2);\ n_1\ge 1, n_2>1, \text{odd},\ n_1+n_2=N\}$$

are $\mathbb{Q}$-linear independent. Theorem 4.6 gives a proof of the motivic version of Kaneko and Tasaka's conjecture.

As we know, for odd $n>1$, the double zeta value $\zeta^o(n,1)$ is not well-defined. However, the motivic sum odd double zeta value $\zeta^{o,\mathfrak{m}}(n,1)$ is well-defined. We will calculate the period of $\zeta^{o,\mathfrak{m}}(n,1)$. Recall that

$$\zeta^{o,\mathfrak{m}}(n,1)=\frac{1}{4}\big[\zeta^{\mathfrak{m}}(n,1)-\zeta^{\mathfrak{m}}(\bar{n},1)-\zeta^{\mathfrak{m}}(n,\bar{1})+\zeta^{\mathfrak{m}}(\bar{n},\bar{1})\big]. \tag{11}$$

Lemma 4.9. *For $n>1$, odd, the period of $\zeta^{o,\mathfrak{m}}(n,1)$ is*

$$\begin{aligned}per(\zeta^{o,\mathfrak{m}}(n,1))&=\frac{1}{4}\big[-\zeta(1,n)-\zeta(n+1)-\zeta(\bar{1},n)+\zeta(\overline{n+1})-\zeta(n,\bar{1})+\zeta(\bar{n},\bar{1})\big]\\&=\frac{1}{4}\big[-\zeta(1,n)+\zeta(1,\bar{n})-\zeta(n,\bar{1})+\zeta(\bar{n},\bar{1})+(2^{-n}-2)\zeta(n+1)\big].\end{aligned}$$

Proof. It is direct to get the periods of $\zeta^{\mathfrak{m}}(n,\overline{1})$ and $\zeta^{\mathfrak{m}}(\bar{n},\overline{1})$, we only need to determine the other two terms in (11). Consider the following regularized integral:

$$\int_{0<t_1<\cdots<t_{n+1}<1-\eta}\frac{dt_1}{1-t_1}\frac{dt_2}{t_2}\cdots\frac{dt_n}{t_n}\frac{dt_{n+1}}{1-t_{n+1}}=\sum_{0<s<r}\frac{(1-\eta)^r}{s^n r}=\sum_{s=1}^{\infty}\frac{1}{s^n}\bigg(-\log(\eta)-\sum_{r=1}^{s}\frac{(1-\eta)^r}{r}\bigg).$$

Let $\log(\eta)=0$. The above integral is equal to $-\sum_{s=1}^{\infty}\sum_{r=1}^{s}(1-\eta)^r/(rs^n)$. Letting $\eta\to 0$, we have

$$-\sum_{s=1}^{\infty}\sum_{r=1}^{s}\frac{1}{rs^n}=-\sum_{0<r<s}\frac{1}{rs^n}-\sum_{s=1}^{\infty}\frac{1}{s^{n+1}}=-\zeta(1,n)-\zeta(n+1).$$

By the definition of *per*, we have

$$per(\zeta^{\mathfrak{m}}(n,1))=-\zeta(1,n)-\zeta(n+1).$$

Similarly, we have

$$per(\zeta^{\mathfrak{m}}(\bar{n},1))=-\zeta(1,\bar{n})-\zeta(\overline{n+1}).$$

Combined with (11), this proves the lemma. □

The following remark follows from Theorem 4.6 and Lemma 4.9 immediately.

Remark 4.10. Every element $\zeta\binom{n_1,n_2}{\epsilon_1,\epsilon_2}$, $n_1+n_2=N$, N even, $(n_2,\epsilon_2)\neq(1,1)$ can be written as a $\mathbb{Q}$-linear combination of $\zeta^o(\text{odd},\text{odd})$, $\zeta(N)$ and $\text{per}(\zeta^{o,\mathfrak{m}}(N-1,1))$ as above.

4C. ***Kaneko and Tasaka's three conjectures.*** Kaneko and Tasaka [2013] additionally conjectured that $\langle\zeta^o(r,N-r);\,1\le r\le N-2\rangle_{\mathbb{Q}}$ is spanned by $\zeta^o(\text{odd},\text{odd})$ and $\zeta(N)$. We will prove this statement as an application of the motivic method.

Theorem 4.11. *For a given even integer $N\ge 4$, the space $\langle\zeta^o(r,N-r);\,1\le r\le N-2\rangle_{\mathbb{Q}}$ is spanned by*

$$\{\zeta(N),\,\zeta^o(r,N-r);\,1\le r\le N-3,\,odd\}.$$

Proof. Denote by $gr_2^{\mathcal{D}}\mathcal{H}_N$ the weight N part of $gr_2^{\mathcal{D}}\mathcal{H}$. According to the property of the period map *per*, we only need to prove that

$$\langle\zeta^{o,\mathfrak{m}}(r,N-r);\,1\le r\le N-2\rangle_{\mathbb{Q}}=\text{span}\{\zeta^{o,\mathfrak{m}}(r,N-r);\,1\le r\le N-3,\,\text{odd}\}$$

in $gr_2^{\mathcal{D}}\mathcal{H}_N$. (Be aware that $\zeta^{o,\mathfrak{m}}(r,N-r);\,1\le r\le N-3$, odd, are elements of $\mathcal{H}$, in the above formula we mean their natural images in $gr_2^{\mathcal{D}}\mathcal{H}_N$.)

We use the same notation as in the proof of Theorem 4.6, there is a matrix E such that

$$\partial\begin{pmatrix}\zeta^{o,\mathfrak{m}}(1,N-1)\\ \zeta^{o,\mathfrak{m}}(3,N-3)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-1,1)\end{pmatrix}=E\begin{pmatrix}\zeta^{\mathfrak{m}}(\overline{1})\otimes\zeta^{\mathfrak{m}}(\overline{N-1})\\ \zeta^{\mathfrak{m}}(\overline{3})\otimes\zeta^{\mathfrak{m}}(\overline{N-3})\\ \vdots\\ \zeta^{\mathfrak{m}}(\overline{N-1})\otimes\zeta^{\mathfrak{m}}(\overline{1})\end{pmatrix},$$

where $E=\widetilde{M}B$ is invertible.

Thus we have

$$\partial E^{-1}\begin{pmatrix}\zeta^{o,\mathfrak{m}}(1,N-1)\\ \zeta^{o,\mathfrak{m}}(3,N-3)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-1,1)\end{pmatrix}=\begin{pmatrix}\zeta^{\mathfrak{m}}(\overline{1})\otimes\zeta^{\mathfrak{m}}(\overline{N-1})\\ \zeta^{\mathfrak{m}}(\overline{3})\otimes\zeta^{\mathfrak{m}}(\overline{N-3})\\ \vdots\\ \zeta^{\mathfrak{m}}(\overline{N-1})\otimes\zeta^{\mathfrak{m}}(\overline{1})\end{pmatrix}.$$

On the other hand, according to Lemmas 4.1 and 4.2, we have

$$\partial\begin{pmatrix}\zeta^{o,\mathfrak{m}}(2,N-2)\\ \zeta^{o,\mathfrak{m}}(4,N-4)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-2,2)\end{pmatrix}=F\begin{pmatrix}\zeta^{\mathfrak{m}}(\overline{1})\otimes\zeta^{\mathfrak{m}}(\overline{N-1})\\ \zeta^{\mathfrak{m}}(\overline{3})\otimes\zeta^{\mathfrak{m}}(\overline{N-3})\\ \vdots\\ \zeta^{\mathfrak{m}}(\overline{N-1})\otimes\zeta^{\mathfrak{m}}(\overline{1})\end{pmatrix},$$

where F is a matrix of order $\left(\frac{N}{2}-1,\frac{N}{2}\right)$, thus

$$\partial\begin{pmatrix}\zeta^{o,\mathfrak{m}}(2,N-2)\\ \zeta^{o,\mathfrak{m}}(4,N-4)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-2,2)\end{pmatrix}=\partial FE^{-1}\begin{pmatrix}\zeta^{o,\mathfrak{m}}(1,N-1)\\ \zeta^{o,\mathfrak{m}}(3,N-3)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-1,1)\end{pmatrix}.$$

By the injectivity of ∂, we have

$$\begin{pmatrix}\zeta^{o,\mathfrak{m}}(2,N-2)\\ \zeta^{o,\mathfrak{m}}(4,N-4)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-2,2)\end{pmatrix}=FE^{-1}\begin{pmatrix}\zeta^{o,\mathfrak{m}}(1,N-1)\\ \zeta^{o,\mathfrak{m}}(3,N-3)\\ \vdots\\ \zeta^{o,\mathfrak{m}}(N-1,1)\end{pmatrix}$$

in $gr_2^{\mathcal{D}}\mathcal{H}_N$. From the explicit calculation in Theorems 4.5 and 4.6, it's obvious to check that the last column of the matrix FE^{-1} is 0. By using the period map, the theorem is proved. □

Kaneko and Tasaka [2013] gave some other conjectures and we can prove them by the same motivic method as above.

Proposition 4.12. *For even integer $N\geq 6$, we have*

$$\langle\zeta(n_1,n_2);n_1+n_2=N,n_2\geq 2\rangle_{\mathbb{Q}}\subseteq\langle\zeta^o(n_1,n_2);n_1+n_2=N,n_i\geq 2\rangle_{\mathbb{Q}},$$
$$\langle\zeta^o(n_1,n_2);n_1+n_2=N,n_i\ \textit{even}\rangle_{\mathbb{Q}}\subseteq\langle\zeta(n_1,n_2);n_1+n_2=N,n_2\geq 2\rangle_{\mathbb{Q}}.$$

Proof. We only need to prove this proposition in the motivic version. According to our calculations above, for $n\geq 0$, letting $s=N-2n-2$, we have

$$\partial_1(\zeta^{\mathfrak{m}}(n_1,n_2))=0,\quad \partial_{2n+1}(\zeta^{\mathfrak{m}}(n_1,n_2))=2^{2n}\left((-1)^{n_1}\binom{2n}{n_1-1}-(-1)^{n_2}\binom{2n}{n_2-1}+\delta\binom{2n}{n_1-1}\right)\zeta^{\mathfrak{m}}(s+1).$$

By the distribution formula, when $s \neq 0$ we have

$$\partial_{2n+1}(\zeta^{\mathfrak{m}}(n_1, n_2)) = \frac{2^{N-2}}{1-2^s}\Big((-1)^{n_1}\binom{2n}{n_1-1} - (-1)^{n_2}\binom{2n}{n_2-1} + \delta\binom{2n}{n_1-1}\Big)\zeta^{\mathfrak{m}}(\overline{s+1})$$

and when $s = 0$ we have

$$\partial_{2n+1}(\zeta^{\mathfrak{m}}(n_1, n_2)) = 0.$$

We have shown there is an invertible matrix S such that

$$\partial S^{-1}\begin{pmatrix}\zeta^{o,\mathfrak{m}}(3, N-3)\\ \vdots \\ \zeta^{o,\mathfrak{m}}(N-3, 3)\end{pmatrix} = \begin{pmatrix}\zeta^{\mathfrak{m}}(\overline{3}) \otimes \zeta^{\mathfrak{m}}(\overline{N-3})\\ \vdots \\ \zeta^{\mathfrak{m}}(\overline{N-3}) \otimes \zeta^{\mathfrak{m}}(\overline{3})\end{pmatrix}.$$

By the injectivity of ∂, we have

$$\langle \zeta^{\mathfrak{m}}(n_1, n_2); n_1 + n_2 = N, n_2 \geq 2\rangle_{\mathbb{Q}} \subseteq \langle \zeta^{o,\mathfrak{m}}(n_1, n_2); n_1 + n_2 = N, 2 \leq n_i \leq k-2\rangle_{\mathbb{Q}}.$$

For $n_1, n_2 \geq 2$, even, if $s = 0$ or $n = 0$,

$$\partial_{2n+1}(\zeta^{o,\mathfrak{m}}(n_1, n_2)) = 0,$$

and if $s, n > 0$,

$$\partial_{2n+1}(\zeta^{o,\mathfrak{m}}(n_1, n_2)) = \frac{1}{4}\frac{1-2^{s+1}}{2^s-1}\Big[\binom{2n}{n_1-1} - \binom{2n}{n_2-1}\Big]\zeta^{\mathfrak{m}}(\overline{s+1}).$$

Since the map ∂ is injective, to prove

$$\langle \zeta^{o}(n_1, n_2); n_1 + n_2 = N, n_i \text{ even}\rangle_{\mathbb{Q}} \subseteq \langle \zeta(n_1, n_2); n_1 + n_2 = N, n_2 \geq 2\rangle_{\mathbb{Q}},$$

it suffices to prove that there are numbers $d\binom{m_1,m_2}{n_1,n_2}$ which satisfy

$$\frac{1}{2^{2n}}\Big[\binom{2n}{n_1-1} - \binom{2n}{n_2-1}\Big] = \sum_{\substack{m_1+m_2=N\\ m_i\geq 1}} d\binom{m_1, m_2}{n_1, n_2}\Big[(-1)^{m_1}\binom{2n}{m_1-1} - (-1)^{m_2}\binom{2n}{m_2-1} + \delta\binom{2n}{m_1-1}\Big]$$

for all $n_1 + n_2 = N,\ n_1, n_2 \geq 2$, even, $3 \leq 2n+1 \leq N-3$. The above statement follows from Lemma 4.13 and Remark 4.14 below. □

Denote by

$$\begin{aligned} V_{N,2} &= \langle x_1^{n_1-1}x_2^{n_2-1}; n_1 + n_2 = N,\ n_1, n_2 \geq 3, \text{odd}\rangle_{\mathbb{Q}},\\ \mathbb{P}_{N,2} &= \langle x_1^{n_1-1}x_2^{n_2-1}; n_1 + n_2 = N,\ n_1, n_2 \geq 1\rangle_{\mathbb{Q}},\\ \mathbb{P}^{od}_{N,2} &= \langle x_1^{n_1-1}x_2^{n_2-1}; n_1 + n_2 = N,\ n_1, n_2 \geq 2, \text{even}\rangle_{\mathbb{Q}}. \end{aligned}$$

For $p(x_1, x_2) \in V_{N,2}$, define

$$\begin{aligned} L_{1,1}(p)(x_1, x_2) &= p(x_1, x_2) + p(x_1 - x_2, x_1) - p(x_1 - x_2, x_2),\\ L_{\frac{1}{2},1}(p)(x_1, x_2) &= p\Big(\frac{x_1}{2}, x_2\Big) + p\Big(\frac{x_1-x_2}{2}, x_1\Big) - p\Big(\frac{x_1-x_2}{2}, x_2\Big). \end{aligned}$$

Lemma 4.13. *Denote by $i^{od}:\mathbb{P}_{N,2}\to\mathbb{P}^{od}_{N,2}$ the natural map which satisfies for $p(x_1,x_2)\in\mathbb{P}_{N,2}$,*

$$i^{od}(p)(x_1,x_2)=p(x_1,x_2)-p(-x_1,x_2).$$

There is a linear map $j:\mathbb{P}_{N,2}\to\mathbb{P}^{od}_{N,2}$ such that the following diagram is commutative:

$$\begin{array}{ccc} V_{N,2} & \xrightarrow{L_{1,1}} & \mathbb{P}_{N,2} \\ \Big\downarrow{\scriptstyle L_{\frac{1}{2},1}} & & \Big\downarrow{\scriptstyle j} \\ \mathbb{P}_{N,2} & \xrightarrow{i^{od}} & \mathbb{P}^{od}_{N,2} \end{array}$$

Proof. Define $j_1:\mathbb{P}_{N,2}\to\mathbb{P}_{N,2}$ as the $\mathbb{Q}$-linear map which is induced by

$$x_1\mapsto\frac{x_1+x_2}{2},\quad x_2\mapsto x_2.$$

Define $j_2:\mathbb{P}_{N,2}\to\mathbb{P}_{N,2}$ as the $\mathbb{Q}$-linear map which is induced by

$$x_1\mapsto\frac{x_1+x_2}{2},\quad x_2\mapsto x_1.$$

Define $j=\frac{1}{2}i^{od}\circ(j_1-j_2)$.

For $p\in V_{N,2}$,

$$\begin{aligned}
i^{od}\circ L_{\frac{1}{2},1}(p)(x_1,x_2)&=L_{\frac{1}{2},1}(p)(x_1,x_2)-L_{\frac{1}{2},1}(p)(-x_1,x_2)\\
&=p\Big(\frac{x_1-x_2}{2},x_1\Big)-p\Big(\frac{x_1-x_2}{2},x_2\Big)-p\Big(\frac{x_1+x_2}{2},x_1\Big)+p\Big(\frac{x_1+x_2}{2},x_2\Big),\\
&j\circ L_{1,1}(p)(x_1,x_2)\\
&=\frac{1}{2}\big[(j_1-j_2)\circ L_{1,1}(p)(x_1,x_2)-(j_1-j_2)\circ L_{1,1}(p)(-x_1,x_2)\big]\\
&=\frac{1}{2}\big(j_1\circ L_{1,1}(p)(x_1,x_2)-j_2\circ L_{1,1}(p)(x_1,x_2)-j_1\circ L_{1,1}(p)(-x_1,x_2)+j_2\circ L_{1,1}(p)(-x_1,x_2)\big)\\
&=\frac{1}{2}\Big(L_{1,1}(p)\Big(\frac{x_1+x_2}{2},x_2\Big)-L_{1,1}(p)\Big(\frac{x_1+x_2}{2},x_1\Big)-L_{1,1}(p)\Big(\frac{-x_1+x_2}{2},x_2\Big)+L_{1,1}(p)\Big(\frac{-x_1+x_2}{2},-x_1\Big)\Big)\\
&=\frac{1}{2}\Big[p\Big(\frac{x_1+x_2}{2},x_2\Big)+p\Big(\frac{x_1-x_2}{2},\frac{x_1+x_2}{2}\Big)-p\Big(\frac{x_1-x_2}{2},x_2\Big)\Big]\\
&\quad-\frac{1}{2}\Big[p\Big(\frac{x_1+x_2}{2},x_1\Big)+p\Big(\frac{x_1-x_2}{2},\frac{x_1+x_2}{2}\Big)-p\Big(\frac{x_1-x_2}{2},x_1\Big)\Big]\\
&\quad-\frac{1}{2}\Big[p\Big(\frac{-x_1+x_2}{2},x_2\Big)+p\Big(\frac{x_1+x_2}{2},\frac{x_1-x_2}{2}\Big)-p\Big(\frac{x_1+x_2}{2},x_2\Big)\Big]\\
&\quad+\frac{1}{2}\Big[p\Big(\frac{-x_1+x_2}{2},x_1\Big)+p\Big(\frac{x_1+x_2}{2},\frac{x_1-x_2}{2}\Big)-p\Big(\frac{x_1+x_2}{2},x_1\Big)\Big]\\
&=p\Big(\frac{x_1-x_2}{2},x_1\Big)-p\Big(\frac{x_1-x_2}{2},x_2\Big)-p\Big(\frac{x_1+x_2}{2},x_1\Big)+p\Big(\frac{x_1+x_2}{2},x_2\Big).
\end{aligned}$$

As a result of the above calculations, the lemma is proved. □

Remark 4.14. Define $d\binom{m_1,m_2}{n_1,n_2}$ to be the coefficient of $x_1^{n_1-1}x_2^{n_2-1}$ in $\frac{1}{2}j(x_1^{m_1-1}x_2^{m_2-1})$, i.e.,

$$\frac{1}{2}j(x_1^{m_1-1}x_2^{m_2-1}) = \sum_{\substack{n_1+n_2=N\\ n_i\geq 2,\text{even}}} d\binom{m_1,m_2}{n_1,n_2}x_1^{n_1-1}x_2^{n_2-1}.$$

For $3 \leq 2n+1 \leq N-3$, by running the commutative diagram in Lemma 4.13 on

$$p = x_1^{2n}x_2^{N-2-2n} \in V_{N,2},$$

we have

$$\frac{1}{2^{2n}}\Big[\binom{2n}{n_1-1}-\binom{2n}{n_2-1}\Big] = \sum_{\substack{m_1+m_2=N\\ m_i\geq 1}} d\binom{m_1,m_2}{n_1,n_2}\Big[(-1)^{m_1}\binom{2n}{m_1-1}-(-1)^{m_2}\binom{2n}{m_2-1}+\delta\binom{2n}{m_1-1}\Big].$$

Remark 4.15. Assuming Grothendieck period conjecture, neither of the two inclusions in Theorem 1.4(ii) is an equality. To see this, we can count the dimensions of two sides on motivic level by the injective map ∂.

By the motivic method we can prove the following, which was proved by Kaneko and Tasaka [2013]:

Proposition 4.16. *For odd integer $N > 6$, we have*

$$\langle \zeta^o(n_1,n_2); n_1+n_2 = N, n_i \geq 2\rangle_{\mathbb{Q}} \subseteq \langle \zeta(n_1,n_2); n_1+n_2 = N, n_2 \geq 2\rangle_{\mathbb{Q}}.$$

5. The higher depth case

In this section we calculate the map

$$\partial : gr_r^{\mathcal{D}}\mathcal{H} \to gr_1^{\mathcal{D}}\mathcal{H}^{\text{odd}} \otimes gr_{r-1}^{\mathcal{D}}\mathcal{H}$$

in the case of $r \geq 3$ for sum odd motivic multiple zeta values explicitly. As a corollary we obtain a basis for the depth-graded motivic triple zeta values of odd weight. What's more, all elements of this basis are the natural images of sum odd motivic multiple zeta values in the depth-graded motivic triple zeta values of odd weight. At last we conjecture that a matrix appearing in the explicit calculation of ∂ on the sum odd motivic multiple zeta values is invertible.

Denote by

$$T_{N,r} = \{(n_1,\ldots,n_r) \in \mathbb{Z}^r; n_1+\cdots+n_r = N, n_i \geq 1, \text{odd}, 1 \leq i \leq r\}.$$

Define $\delta\binom{m_1,\ldots,m_r}{n_1,\ldots,n_r} = 1$ if $(m_1,\ldots,m_r) = (n_1,\ldots,n_r)$, $\delta\binom{m_1,\ldots,m_r}{n_1,\ldots,n_r} = 0$ if $(m_1,\ldots,m_r) \neq (n_1,\ldots,n_r)$.

Proposition 5.1. *Let $N \equiv r \mod 2$ and $N \geq r+2$. For $(k_1,\ldots,k_r) \in T_{N,r}$ we have*

$$\partial(\zeta^{o,\mathfrak{m}}(k_1,\ldots,k_r)) = \sum_{(n_1,\ldots,n_r)\in T_{N,r}} e\binom{k_1,k_2,\ldots,k_r}{n_1,n_2,\ldots,n_r}\zeta^{\mathfrak{m}}(\overline{n_1}) \otimes \zeta^{o,\mathfrak{m}}(n_2,\ldots,n_r),$$

where for $n_1 \geq 3$, odd,

$$e\binom{k_1,k_2,\ldots,k_r}{n_1,n_2,\ldots,n_r} = \Big(2^{n_1-1}-\frac{1}{2}\Big)\delta\binom{k_1,k_2,\ldots,k_r}{n_1,n_2,\ldots,n_r} + \frac{1}{2}\sum_{i=1}^{r-1}\Big(\binom{n_1-1}{k_{i+1}-1}-\binom{n_1-1}{k_i-1}\Big)\delta\binom{k_1,\ldots,k_{i-1},k_{i+2},\ldots,k_r}{n_2,\ldots,n_i,n_{i+2},\ldots,n_r},$$

$$e\binom{k_1,k_2,\ldots,k_r}{1,n_2,\ldots,n_r} = -\delta\binom{k_1,k_2,\ldots,k_r}{1,n_2,\ldots,n_r} + \frac{1}{2}\delta\binom{k_1,\ldots,k_{r-1},k_r}{n_2,\ldots,n_r,1}.$$

Proof. Notice the following calculation:

$$\begin{aligned}
&\zeta^{o,\mathfrak{m}}(n_1,\dots,n_r)\\
&=\frac{1}{2^r}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\epsilon_1\cdots\epsilon_r\zeta^{\mathfrak{m}}\binom{n_1,\dots,n_r}{\epsilon_1,\dots,\epsilon_r}\\
&=\frac{1}{2^r}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\epsilon_1\cdots\epsilon_r\mathcal{I}^{\mathfrak{m}}(e^{\epsilon_1\cdots\epsilon_r}(e^0)^{n_1-1}e^{\epsilon_2\cdots\epsilon_r}(e^0)^{n_2-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1})\\
&=\frac{1}{2^r}\sum_{\epsilon_i\in\{\pm1\},2\le i\le r}\mathcal{I}^{\mathfrak{m}}\big[(e^1(e^0)^{n_1-1}e^{\epsilon_2\cdots\epsilon_r}(e^0)^{n_2-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1})\\
&\qquad\qquad -e^{-1}(e^0)^{n_1-1}e^{\epsilon_2\cdots\epsilon_r}(e^0)^{n_2-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1})\big]\\
&=\frac{1}{2^r}\sum_{\epsilon_i\in\{\pm1\},2\le i\le r}\mathcal{I}^{\mathfrak{m}}[(e^1(e^0)^{n_1-1}e^{\epsilon_2}(e^0)^{n_2-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1})-e^{-1}(e^0)^{n_1-1}e^{\epsilon_2}(e^0)^{n_2-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1})].
\end{aligned}$$

We have

$$\overline{\sigma_{2n+1}}\circ(e_0^{a_0}e_{i_1}e_0^{a_1}\cdots e_{i_r}e_0^{a_r})=\sum_{j=1}^{r}e_0^{a_0}\cdots(\overline{\sigma_{2n+1}}\circ e_{i_j})e_0^{a_j}\cdots e_{i_r}e_0^{a_r}+e_0^{a_0}\cdots e_{i_r}e_0^{a_r}\overline{\sigma_{2n+1}}.$$

Since, when $n=0$, we have

$$\bar{\sigma}_1=e_{-1},\ e_{-1}\circ e_1=e_{-1}e_1-e_1e_{-1},\ e_{-1}\circ e_{-1}=e_1e_{-1}-e_{-1}e_1,$$

it follows that

$$\begin{aligned}
\bar{\partial}_1\big(e^{i_1}(e^0)^{a_1}\cdots e^{i_s}(e^0)^{a_s}\big)=\delta\binom{a_1}{0}\delta\binom{i_1i_2}{-1}i_1(e^{-1}-e^1)(e^0)^{a_2}\cdots e^{i_s}(e^0)^{a_s}+\cdots\\
+\delta\binom{a_{s-1}}{0}\delta\binom{i_{s-1}i_s}{-1}i_{s-1}e^{i_1}(e^0)^{a_1}\cdots e^{i_{s-2}}(e^0)^{a_{s-2}}(e^{-1}-e^1)(e^0)^{a_s}\\
+\delta\binom{a_s}{0}\delta\binom{i_s}{-1}e^{i_1}(e^0)^{a_1}\cdots e^{i_{s-1}}(e^0)^{a_{s-1}}.
\end{aligned}$$

As a result, we have

$$\begin{aligned}
&\partial_1\big(\zeta^{o,\mathfrak{m}}(n_1,n_2,\dots,n_r)\big)\\
&=\frac{1}{2^r}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\mathcal{I}^{\mathfrak{m}}\big[\bar{\partial}_1\big(\epsilon_1e^{\epsilon_1}(e^0)^{n_1-1}e^{\epsilon_2}(e^0)^{n_2-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1}\big)\big]\\
&=\frac{1}{2^r}\delta\binom{n_1}{1}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\mathcal{I}^{\mathfrak{m}}\Big(\delta\binom{\epsilon_1\epsilon_2}{-1}\epsilon_1^2(e^{-1}-e^1)(e^0)^{n_2-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1}\Big)+\cdots\\
&\quad+\frac{1}{2^r}\delta\binom{n_{r-1}}{1}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\mathcal{I}^{\mathfrak{m}}\Big[\delta\binom{\epsilon_{r-1}\epsilon_r}{-1}\epsilon_1\epsilon_{r-1}e^{\epsilon_1}(e^0)^{n_1-1}\cdots e^{\epsilon_{r-2}}(e^0)^{n_{r-2}-1}(e^{-1}-e^1)(e^0)^{n_r-1}\Big]\\
&\qquad+\frac{1}{2^r}\delta\binom{n_r}{1}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\mathcal{I}^{\mathfrak{m}}\Big(\delta\binom{\epsilon_r}{-1}\epsilon_1e^{\epsilon_1}(e^0)^{n_1-1}\cdots e^{\epsilon_{r-1}}(e^0)^{n_{r-1}-1}\Big)\\
&=-\delta\binom{n_1}{1}\zeta^{o,\mathfrak{m}}(n_2,\dots,n_r)+\frac{1}{2}\delta\binom{n_r}{1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-1}).
\end{aligned}$$

In the above calculation, the last equality is due to the fact that

$$\sum_{\epsilon_1,\epsilon_2\in\{\pm1\}} \delta\binom{\epsilon_1\epsilon_2}{-1}\epsilon_1 = 0.$$

Similarly for $n \geq 1$, from

$$\overline{\sigma_{2n+1}} = (1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{2n-r}e_{-1}e_0^r + 2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}e_0^{2n-r}e_1e_0^r,$$

$$\begin{aligned}\overline{\sigma_{2n+1}}\circ e_1 = {}&(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}\big(e_0^{2n-r}e_{-1}e_0^re_1 - e_1e_0^{2n-r}e_{-1}e_0^r\big)\\ &+2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}\big(e_0^{2n-r}e_1e_0^re_1 - e_1e_0^{2n-r}e_1e_0^r\big),\end{aligned}$$

$$\begin{aligned}\overline{\sigma_{2n+1}}\circ e_{-1} = {}&(1-2^{2n})\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}\big(e_0^{2n-r}e_1e_0^re_{-1} - e_{-1}e_0^{2n-r}e_1e_0^r\big)\\ &+2^{2n}\sum_{r=0}^{2n}(-1)^r\binom{2n}{r}\big(e_0^{2n-r}e_{-1}e_0^re_{-1} - e_{-1}e_0^{2n-r}e_{-1}e_0^r\big),\end{aligned}$$

we have

$$\begin{aligned}&\bar\partial_{2n+1}\big(e^{i_1}(e^0)^{a_1}\cdots e^{i_s}(e^0)^{a_s}\big)\\
&\quad=(1-2^{2n})\delta\binom{a_1}{2n}\delta\binom{i_1i_2}{-1}e^{-i_1}(e^0)^{a_2}e^{i_3}(e^0)^{a_3}\cdots e^{i_s}(e^0)^{a_s}\\
&\qquad-(1-2^{2n})(-1)^{a_1}\binom{2n}{a_1}\delta\binom{i_1i_2}{-1}e^{i_1}(e^0)^{a_1+a_2-2n}e^{i_3}(e^0)^{a_3}\cdots e^{i_s}(e^0)^{a_s}\\
&\qquad+2^{2n}\delta\binom{a_1}{2n}\delta\binom{i_1i_2}{1}e^{i_1}(e^0)^{a_2}e^{i_3}(e^0)^{a_3}\cdots e^{i_s}(e^0)^{a_s}\\
&\qquad-2^{2n}(-1)^{a_1}\binom{2n}{a_1}\delta\binom{i_1i_2}{1}e^{i_1}(e^0)^{a_1+a_2-2n}e^{i_3}(e^0)^{a_3}\cdots e^{i_s}(e^0)^{a_s}\\
&\qquad+\cdots+(1-2^{2n})(-1)^{a_{s-1}}\binom{2n}{a_{s-1}}\delta\binom{i_{s-1}i_s}{-1}e^{i_1}(e^0)^{a_1}\cdots e^{i_{s-2}}(e^0)^{a_{s-2}+a_{s-1}-2n}e^{-i_{s-1}}(e^0)^{a_s}\\
&\qquad-(1-2^{2n})(-1)^{a_{s-1}}\binom{2n}{a_{s-1}}\delta\binom{i_{s-1}i_s}{-1}e^{i_1}(e^0)^{a_1}\cdots e^{i_{s-2}}(e^0)^{a_{s-2}}e^{i_{s-1}}(e^0)^{a_{s-1}+a_s-2n}\\
&\qquad+2^{2n}(-1)^{a_{s-1}}\binom{2n}{a_{s-1}}\delta\binom{i_{s-1}i_s}{1}e^{i_1}(e^0)^{a_1}\cdots e^{i_{s-2}}(e^0)^{a_{s-2}+a_{s-1}-2n}e^{i_{s-1}}(e^0)^{a_s}\\
&\qquad+2^{2n}(-1)^{a_{s-1}}\binom{2n}{a_{s-1}}\delta\binom{i_{s-1}i_s}{1}e^{i_1}(e^0)^{a_1}\cdots e^{i_{s-2}}(e^0)^{a_{s-2}}e^{i_{s-1}}(e^0)^{a_{s-1}+a_s-2n}\\
&\qquad+(1-2^{2n})(-1)^{a_s}\binom{2n}{a_s}\delta\binom{i_s}{-1}e^{i_1}(e^0)^{a_1}\cdots(e^0)^{a_{s-2}}e^{i_{s-1}}(e^0)^{a_{s-1}+a_s-2n}\\
&\qquad+2^{2n}(-1)^{a_s}\binom{2n}{a_s}\delta\binom{i_s}{1}e^{i_1}(e^0)^{a_1}\cdots(e^0)^{a_{s-2}}e^{i_{s-1}}(e^0)^{a_{s-1}+a_s-2n}.\end{aligned}$$

Thus for $(n_1, n_2, \dots, n_r) \in T_{N,r}$, we have

$$
\begin{aligned}
&\partial_{2n+1}\big(\zeta^{o,\mathfrak{m}}(n_1,\dots,n_r)\big)\\
&= \frac{1}{2^r}\sum_{\epsilon_i\in\{\pm1\},1\le i\le r}\epsilon_1\mathcal{I}^{\mathfrak{m}}\big[\bar{\partial}_{2n+1}\big(e^{\epsilon_1}(e^0)^{n_1-1}\cdots e^{\epsilon_r}(e^0)^{n_r-1}\big)\big]\\
&= \frac{1}{2}(2^{2n}-1)\delta\tbinom{n_1}{2n+1}\zeta^{o,\mathfrak{m}}(n_2,\dots,n_r)\\
&\quad+\frac{(-1)^{n_1-1}}{2}(2^{2n}-1)\tbinom{2n}{n_1-1}\zeta^{o,\mathfrak{m}}(n_1+n_2-2n-1,n_3,\dots,n_r)\\
&\qquad+\frac{2^{2n}}{2}\delta\tbinom{n_1}{2n+1}\zeta^{o,\mathfrak{m}}(n_2,\dots,n_r)-\frac{2^{2n}}{2}(-1)^{n_1-1}\tbinom{2n}{n_1-1}\zeta^{o,\mathfrak{m}}(n_1+n_2-2n-1,n_3,\dots,n_r)\\
&\qquad\quad+\cdots+\frac{1-2^{2n}}{2}(-1)^{n_{r-1}-1}\tbinom{2n}{n_{r-1}-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-3},n_{r-2}+n_{r-1}-2n-1,n_r)\\
&\qquad\qquad-\frac{1-2^{2n}}{2}(-1)^{n_{r-1}-1}\tbinom{2n}{n_{r-1}-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-2},n_{r-1}+n_r-2n-1)\\
&\qquad\qquad\quad+\frac{2^{2n}}{2}(-1)^{n_{r-1}-1}\tbinom{2n}{n_{r-1}-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-3},n_{r-2}+n_{r-1}-2n-1,n_r)\\
&\qquad\qquad\qquad-\frac{2^{2n}}{2}(-1)^{n_{r-1}-1}\tbinom{2n}{n_{r-1}-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-2},n_{r-1}+n_r-2n-1)\\
&\qquad\qquad\qquad\quad+\frac{1-2^{2n}}{2}(-1)^{n_r-1}\tbinom{2n}{n_r-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-2},n_{r-1}+n_r-2n-1)\\
&\qquad\qquad\qquad\qquad+\frac{2^{2n}}{2}(-1)^{n_r-1}\tbinom{2n}{n_r-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-2},n_{r-1}+n_r-2n-1)\\
&= \Big(1-\frac{1}{2}\Big)\delta\tbinom{n_1}{2n+1}\zeta^{o,\mathfrak{m}}(n_2,\dots,n_r)-\frac{1}{2}\tbinom{2n}{n_1-1}\zeta^{o,\mathfrak{m}}(n_1+n_2-2n-1,n_3,\dots,n_r)\\
&\quad+\cdots+\frac{1}{2}\tbinom{2n}{n_{r-1}-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-3},n_{r-2}+n_{r-1}-2n-1,n_r)\\
&\qquad-\frac{1}{2}\tbinom{2n}{n_{r-1}-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-2},n_{r-1}+n_r-2n-1)\\
&\qquad\quad+\frac{1}{2}\tbinom{2n}{n_r-1}\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{r-2},n_{r-1}+n_r-2n-1)\\
&= (2^{n_1-1}-\frac{1}{2})\delta\tbinom{n_1}{2n+1}\zeta^{o,\mathfrak{m}}(n_2,\dots,n_r)\\
&\quad+\frac{1}{2}\sum_{i=1}^{r-1}\Big(\tbinom{2n}{n_{i+1}-1}-\tbinom{2n}{n_i-1}\Big)\cdot\zeta^{o,\mathfrak{m}}(n_1,\dots,n_{i-1},n_i+n_{i+1}-2n-1,n_{i+2},\dots,n_r).
\end{aligned}
$$

Thus the proposition holds. □

With the help of the above proposition, we can generalize Theorem 4.6 to the case of depth 3.

Theorem 5.2. *For $r = 3$, $N \geq 5$ odd.*

(i) *The set of the images of elements*

$$\{\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3); n_1 + n_2 + n_3 = N, n_i \text{ odd}\}$$

in $gr_3^{\mathcal{D}}\mathcal{H}_N$ is a basis of the total space $gr_3^{\mathcal{D}}\mathcal{H}_N$.

(ii) *Every element in*

$$\mathcal{P}_{N,3}^{o} = \langle \zeta^{o}(n_1, n_2, n_3); n_1 + n_2 + n_3 = N, n_3 > 1\rangle_{\mathbb{Q}}$$

can be written as a $\mathbb{Q}$-linear combination of some sum odd multiple zeta values of weight N, depth 3 and multiple zeta values relative to μ_2 of weight N, depth less than 3.

Proof. We have known that the set of elements

$$\{\zeta^{o,\mathfrak{m}}(n_1, n_2); n_1 + n_2 = k,\ n_i \geq 1, \text{odd}\}$$

is a basis of the space $gr_2^{\mathcal{D}}\mathcal{H}_N$. Similar to the proofs of Theorems 4.6 and 4.11 we will use the above proposition to prove the first part. Using Lemma 4.7, we only need to prove that for any given $(n_1, n_2, n_3) \in T_{N,3}$,

$$\sum_{(k_1,k_2,\ldots,k_r)\neq(n_1,n_2,\ldots,n_r)} \left|e\binom{k_1, k_2, \ldots, k_r}{n_1, n_2, \ldots, n_r}\right| < \left|e\binom{n_1, n_2, \ldots, n_r}{n_1, n_2, \ldots, n_r}\right|.$$

When $n_1 = 1$, the above inequality is trivial.

When $n_1 \geq 3$, we have

$$\begin{aligned}
&\sum_{(k_1,k_2,k_3)\in S_{N,3}} \left|\delta\binom{k_3}{n_3}\left(\binom{n_1-1}{k_1-1} - \binom{n_1-1}{k_2-1}\right) + \delta\binom{k_1}{n_2}\left(\binom{n_1-1}{k_2-1} - \binom{n_1-1}{k_3-1}\right)\right| \\
&\quad\leq \sum_{(k_1,k_2,k_3)\in S_{N,3}} \left|\delta\binom{k_3}{n_3}\left(\binom{n_1-1}{k_1-1} - \binom{n_1-1}{k_2-1}\right)\right| + \left|\delta\binom{k_1}{n_2}\left(\binom{n_1-1}{k_2-1} - \binom{n_1-1}{k_3-1}\right)\right| \\
&\quad\leq \sum_{(k_1,k_2,k_3)\in S_{N,3}} \delta\binom{k_3}{n_3}\left(\left|\binom{n_1-1}{k_1-1}\right| + \left|\binom{n_1-1}{k_2-1}\right|\right) + \delta\binom{k_1}{n_2}\left(\left|\binom{n_1-1}{k_2-1}\right| + \left|\binom{n_1-1}{k_3-1}\right|\right) - 2 \\
&\quad\leq 4\sum_{i\geq 0}\binom{n_1-1}{2i} - 2 < 2^{n_1} - 1.
\end{aligned}$$

Thus the first statement holds.

As for the second part of this theorem, denote by

$$\mathcal{C} = \{\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3); n_1 + n_2 + n_3 = N\} \setminus \{\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3); n_1 + n_2 + n_3 = N, n_i \text{ odd}\}.$$

Assume that there is a lexicographical order on $T_{N,r}$, it induces an order on $\mathcal{C}$ and $\{\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3); n_1 + n_2 + n_3 = N\}$. Let α (resp. β) be the column vector whose i-th element is the i-th element in $\mathcal{C}$

(resp. $\{\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3); n_1+n_2+n_3 = N\}$). The argument above and Proposition 5.1 show that there is a matrix P and an invertible matrix Q such that

$$\partial(\alpha) = P\gamma, \quad \partial(\beta) = Q\gamma,$$

where $\gamma = (\zeta^{\mathfrak{m}}(\overline{1}) \otimes \zeta^{o,\mathfrak{m}}(1, N-2), \ldots, \zeta^{\mathfrak{m}}(\overline{N-2}) \otimes \zeta^{o,\mathfrak{m}}(1,1))^T$.

The last column of Q is $\left(0, \ldots, 0, 2^{N-3} - \frac{1}{2}\right)^T$ obviously, and the last column of P is 0 because of the following equation:

$$\partial_{N-2}(\zeta^{o,\mathfrak{m}}(n_1, n_2, n_3)) = 0, \quad \text{for all } \zeta^{o,\mathfrak{m}}(n_1, n_2, n_3) \in \mathcal{C}.$$

By the injectivity of ∂ we have

$$\alpha = PQ^{-1}\beta,$$

and that the last row of PQ^{-1} is 0. Thus the theorem holds. □

Furthermore we can put forward the following conjecture:

Conjecture 5.3. *For any $r \geq 4$, $N \geq r+2$, $N - r \equiv 0 \bmod 2$, the order $|T_{N,r}|$ matrix*

$$E = \left(e\binom{k_1, k_2, \ldots, k_r}{n_1, n_2, \ldots, n_r}\right)$$

as in Proposition 5.1 is invertible.

Remark 5.4. If this conjecture is true we can directly generalize Theorem 5.2 to cases of higher depth by induction. Unfortunately in depth ≥ 4, the matrix E is usually not a strictly diagonal dominant matrix any more. Thus Lemma 4.7 is not helpful in cases of higher depth. By explicit calculation we have checked that Conjecture 5.3 is true for $r = 4$, $N = 6, 8, 10$.

Remark 5.5. The motivic approach in this paper can also be used to study cyclotomic multiple zeta values for other roots of unity. By explicit calculation, the analogue of Theorem 1.3 in other roots of unity (at least for μ_3, μ_4, μ_6 and μ_8) is true. Unfortunately we can't find any short exact sequence either for other roots of unity or for μ_2 in higher depth.

Acknowledgements

We express our sincere gratitude to Koji Tasaka for pointing out an error in the early version of this manuscript. We also thank the referee for detailed comments and suggestions to improve this paper.

References

[Brown 2012] F. Brown, "Mixed Tate motives over $\mathbb{Z}$", *Ann. of Math.* (2) **175**:2 (2012), 949–976. MR Zbl

[Brown 2013] F. Brown, "Depth-graded motivic multiple zeta values", preprint, 2013. arXiv

[Deligne 2010] P. Deligne, "Le groupe fondamental unipotent motivique de $\mathbb{G}_m - \mu_N$, pour $N = 2, 3, 4, 6$ ou 8", *Publ. Math. Inst. Hautes Études Sci.* **112** (2010), 101–141. MR Zbl

[Deligne and Goncharov 2005] P. Deligne and A. B. Goncharov, "Groupes fondamentaux motiviques de Tate mixte", *Ann. Sci. École Norm. Sup.* (4) **38**:1 (2005), 1–56. MR Zbl

[Enriquez and Lochak 2016] B. Enriquez and P. Lochak, "Homology of depth-graded motivic Lie algebras and Koszulity", *J. Théor. Nombres Bordeaux* **28**:3 (2016), 829–850. MR Zbl

[Gangl et al. 2006] H. Gangl, M. Kaneko, and D. Zagier, "Double zeta values and modular forms", pp. 71–106 in *Automorphic forms and zeta functions* (Tokyo, 2004), edited by S. Böcherer et al., World Sci., Hackensack, NJ, 2006. MR Zbl

[Gil and Fresán 2018] J. B. Gil and J. Fresán, "Multiple zeta values: from numbers to motives", 2018, available at http://javier.fresan.perso.math.cnrs.fr/mzv.pdf. To appear in Clay Math. Proc.

[Glanois 2016] C. Glanois, "Motivic unipotent fundamental groupoid of $\mathbb{G}_m - \mu_N$ for $N = 2, 3, 4, 6, 8$ and Galois descents", *J. Number Theory* **160** (2016), 334–384. MR Zbl

[Hoffman 1997] M. E. Hoffman, "The algebra of multiple harmonic series", *J. Algebra* **194**:2 (1997), 477–495. MR Zbl

[Kaneko and Tasaka 2013] M. Kaneko and K. Tasaka, "Double zeta values, double Eisenstein series, and modular forms of level 2", *Math. Ann.* **357**:3 (2013), 1091–1118. MR Zbl

[Le and Murakami 1996] T. T. Q. Le and J. Murakami, "Kontsevich's integral for the Kauffman polynomial", *Nagoya Math. J.* **142** (1996), 39–65. MR Zbl

[Li 2019] J. Li, "The depth structure of motivic multiple zeta values", *Math. Ann.* **374**:1-2 (2019), 179–209. MR Zbl

[Li 2020] J. Li, "Depth-graded motivic Lie algebra", *J. Number Theory* **214** (2020), 38–55. MR Zbl

[Li and Liu 2020] J. Li and F. Liu, "Motivic double zeta values of odd weight", *Manuscripta Math.* (2020).

[Ma 2015] D. Ma, "Connections between double zeta values relative to μ_N, Hecke operators T_N, and newforms of level $\Gamma_0(N)$ for $N = 2, 3$", preprint, 2015. arXiv

[Soudères 2010] I. Soudères, "Motivic double shuffle", *Int. J. Number Theory* **6**:2 (2010), 339–370. MR Zbl

Communicated by Hélène Esnault
Received 2019-05-15 Revised 2019-09-08 Accepted 2020-06-13

zyjin@pku.edu.cn *School of Mathematical Sciences, Peking University, Beijing, China*

lijiangtao@amss.ac.cn *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*

msp

ALGEBRA AND NUMBER THEORY 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2713

An intriguing hyperelliptic Shimura curve quotient of genus 16

Lassina Dembélé

Let F be the maximal totally real subfield of $\mathbb{Q}(\zeta_{32})$, the cyclotomic field of 32-nd roots of unity. Let D be the quaternion algebra over F ramified exactly at the unique prime above 2 and 7 of the real places of F. Let $\mathcal{O}$ be a maximal order in D, and $X_0^D(1)$ the Shimura curve attached to $\mathcal{O}$. Let $C = X_0^D(1)/\langle w_D\rangle$, where w_D is the unique Atkin–Lehner involution on $X_0^D(1)$. We show that the curve C has several striking features. First, it is a hyperelliptic curve of genus 16, whose hyperelliptic involution is exceptional. Second, there are 34 Weierstrass points on C, and exactly half of these points are CM points; they are defined over the Hilbert class field of the unique CM extension E/F of class number 17 contained in $\mathbb{Q}(\zeta_{64})$, the cyclotomic field of 64-th roots of unity. Third, the normal closure of the field of 2-torsion of the Jacobian of C is the Harbater field N, the unique Galois number field $N/\mathbb{Q}$ unramified outside 2 and ∞, with Galois group $\mathrm{Gal}(N/\mathbb{Q}) \simeq F_{17} = \mathbb{Z}/17\mathbb{Z} \rtimes (\mathbb{Z}/17\mathbb{Z})^\times$. In fact, the Jacobian $\mathrm{Jac}(X_0^D(1))$ has the remarkable property that each of its simple factors has a 2-torsion field whose normal closure is the field N. Finally, and perhaps the most striking fact about C, is that it is also hyperelliptic over $\mathbb{Q}$.

1. Introduction

Let F be the maximal totally real subfield of $\mathbb{Q}(\zeta_{32})$, the cyclotomic field of 32-nd roots of unity. Recall that 2 is totally ramified in F, and let $\mathfrak{p}$ be the unique prime above it. Let D be the quaternion algebra defined over F ramified exactly at $\mathfrak{p}$ and 7 of the real places of F. Let $\mathcal{O}$ be a maximal order in D, and $X_0^D(1)$ the Shimura curve attached to $\mathcal{O}$. Let $C = X_0^D(1)/\langle w_D\rangle$, where w_D is the unique Atkin–Lehner involution on $X_0^D(1)$. Let $\mathrm{Jac}(X_0^D(1))$ and $\mathrm{Jac}(C)$ be the Jacobians of $X_0^D(1)$ and $\mathrm{Jac}(C)$, respectively. In this note, we show that C has several striking properties. First, we prove the following theorem (Theorem 5.13).

Theorem A. *The curve C is a hyperelliptic curve of genus* 16 *defined over* $\mathbb{Q}$.

We first show that C is hyperelliptic over F (Theorem 5.9), then we apply a descent argument from [Sijsling and Voight 2016] to show that both the curve and the hyperelliptic involution are defined over $\mathbb{Q}$. For the first part, we simply count the number of Weierstrass points on C. This count yields that C has 34 Weierstrass points, the maximum number for a hyperelliptic curve of genus 16 by the Weierstrass gap

The author was supported by EPSRC Grants EP/J002658/1 and EP/L025302/1, a Visiting grant from the Max-Planck Institute for Mathematics, and a Simons Collaboration Grant (550029).

MSC2010: primary 11F41; secondary 11F80.

Keywords: abelian varieties, Hilbert modular forms, Shimura curves.

theorem. Half of those Weierstrass points are CM points defined over the Hilbert class field of the unique CM extension E/F of class number 17 contained in $\mathbb{Q}(\zeta_{64})$, the cyclotomic field of 64-th roots of unity.

To show that C is in fact defined over $\mathbb{Q}$, we determine the automorphism group $\mathrm{Aut}(C)$ of C as a curve over F. We do this by exploiting the Čerednik–Drinfel'd 2-adic uniformisation of $X_0^D(1)$ and the fact that the automorphism group of a stable curve injects into an *admissible* subgroup of the automorphism group of its dual graph (see [Deligne and Mumford 1969] and Section 4F for the definition of admissibility). A careful study of the dual graph of the stable model of C over the completion of F at $\mathfrak{p}$ then yields that $\mathrm{Aut}(C) = \mathbb{Z}/2\mathbb{Z}$. As a result, we get that the *only* nontrivial automorphism of C is the hyperelliptic involution, which in this case must be exceptional since the curve C is obtained as the quotient of $X_0^D(1)$ by the unique Atkin–Lehner involution w_D.

Our second result concerns the field of 2-torsion of $\mathrm{Jac}(C)$. It is known that 17 is the smallest odd integer which can occur as the degree of a number field $K/\mathbb{Q}$ for which 2 is the only finite prime which ramifies. That there is no such integer less than 17 follows from [Jones 2010]. On the other hand, Harbater [1994] proved that there is a unique Galois number field $N/\mathbb{Q}$ unramified outside 2 and ∞, with Galois group $\mathrm{Gal}(N/\mathbb{Q}) \simeq F_{17} = \mathbb{Z}/17\mathbb{Z} \rtimes (\mathbb{Z}/17\mathbb{Z})^\times$. So, the fixed field of the Sylow 2-subgroup of F_{17} is a number field of degree 17 in which 2 is the only ramified finite prime. Noam Elkies provided a degree 17 polynomial whose splitting field is N. The computation which led to that polynomial stemmed from a discussion on `mathoverflow.net` [Rouse and Elkies 2014] initiated by Jeremy Rouse. In the context of that discussion, it is natural to ask whether there is a curve defined over $\mathbb{Q}$, with good reduction away from 2, whose field of 2-torsion is the Harbater field N. The following theorem provides an affirmative answer to that question (Theorem 6.1).

Theorem B. *The field of* 2*-torsion of* $\mathrm{Jac}(C)$ *is the Harbater field* N*.*

The fact that the Harbater field can be realised as the field of 2-torsion of a *hyperelliptic* curve of rather large genus, with good reduction outside 2, seems rather remarkable to us. For that reason, we think that it would be very interesting to find a defining equation for C over $\mathbb{Q}$. This is a question of independent interest that we hope to consider in the future.

In fact, we prove a slightly stronger result than Theorem B. Namely, up to isogeny, the Jacobian $\mathrm{Jac}(X_0^D(1))$ decomposes as the product of four abelian varieties of dimension 4 and one of dimension 24. We give two different proofs of the following (Theorem 6.4).

Theorem C. *Let* A *be a simple factor of* $\mathrm{Jac}(X_0^D(1))$*. Then the normal closure of the field of* 2*-torsion of* A *is the Harbater field* N*.*

The second proof of Theorem C uses congruences. Namely, let $S_2^D(1)$ be the space of automorphic forms of level (1) and weight 2 over the quaternion algebra D, and $\mathbb{T}$ be the Hecke algebra acting on $S_2^D(1)$. We show that there are two congruence classes modulo 2 among the newforms in $S_2^D(1)$, whose associated mod 2 residual Galois representations have the same image D_{17}. These two congruence classes are permuted by $\mathrm{Gal}(F/\mathbb{Q})$. As a result, we get that the normal closure of the field of 2-torsion of every simple factor of $\mathrm{Jac}(X_0^D(1))$ is the Harbater field N. Interestingly, the existence of these two *distinct*

congruence classes modulo 2 turns out to have the following amusing consequence: the connectedness of Spec($\mathbb{T}$), which is obtained by an argument à la Mazur [1977, Proposition 10.6], *cannot* arise from a single congruence modulo 2. In other words, the existence of the Harbater field as the normal closure of the field of 2-torsion of $\mathrm{Jac}(X_0^D(1))$ is an obstruction to the connectedness of Spec($\mathbb{T}$) being achieved via a unique congruence modulo 2. This is due to the tautological reason that the semidirect product $F_{17} = D_{17} \rtimes \mathbb{Z}/8\mathbb{Z}$ is *nonsplit*. In fact, we show that the connectedness of Spec($\mathbb{T}$) is given by two different congruences modulo 3 and 5.

Our initial interest in the curve $X_0^D(1)$ stems from a conjecture of Benedict H. Gross which states that, for any prime p, there is a nonsolvable number field $K/\mathbb{Q}$ ramified at p (and possibly at ∞) only. In [Dembélé 2009], we proved that conjecture for $p = 2$ by using Hilbert modular forms of level (1) and weight 2 over F. Theorem C implies that none of the simple factors of $\mathrm{Jac}(X_0^D(1))$ has a 2-torsion field that can be used to provide an affirmative answer to the Gross conjecture for number fields given that N is solvable. Amusingly, it turns out that the simple factors of $\mathrm{Jac}(X_0^D(1))$ are more interesting in relation to other conjectures of Gross [2016] which concern modularity of abelian varieties not of GL_2-type. Indeed, functorially, these simple factors are related to abelian varieties defined over $\mathbb{Q}$ with *small* or even trivial endomorphism rings, but which acquire extra endomorphisms over F, as we explain later (see also [Cunningham and Dembélé 2017]).

The outline of the paper is as follows. In Section 2, we recall the necessary background on Weierstrass points and hyperellipticity. In Section 3, we recall the necessary background on arithmetic groups in quaternion algebras, and compliment this by discussing optimal embeddings into maximal arithmetic Fuchsian groups. In Section 4, we review the theory of Shimura curves, especially their p-adic uniformisation. Finally, in Sections 5 and 6, we discuss our example, its Jacobian and the connection of their 2-torsion fields with the Harbater field.

2. Background on Weierstrass points

Throughout this section, X is a smooth projective curve of genus $g \geq 2$ defined over a field k of characteristic 0, with algebraic closure $\bar{k}$.

2A. ***Definition and properties.*** Let P be a point on X. We say that P is a *Weierstrass point* if there exists a differential form $\omega \in H^0(X, \Omega_X^1)$ such that $\mathrm{ord}_P(\omega) \geq g$. We let $\mathscr{W}$ be the set of all Weierstrass points on $X(\bar{k})$. Alternatively, one can describe $\mathscr{W}$ as follows. Let D be a divisor on X, and $\mathscr{L}(D)$ the Riemann–Roch space associated to D, i.e.,

$$\mathscr{L}(D) := \{f \in k(X)^\times : \mathrm{div}(f) + D \geq 0\} \cup \{0\}.$$

By the Riemann–Roch theorem, $\mathscr{L}(D)$ is finite dimensional, and we let $\ell(D)$ be its dimension.

Proposition 2.1. *Let P be a point on X. Then, $P \in \mathscr{W}$ if and only if $\ell(gP) \geq 2$.*

Proof. This is a consequence of the Riemann–Roch theorem [Hindry and Silverman 2000, §A.4]. □

The *gap sequence* associated to a Weierstrass point P is the set

$$G(P) := \{n \in \mathbb{Z}_{\geq 0} : \ell(nP) = \ell((n-1)P)\}.$$

The *weight* of the Weierstrass point P is defined by

$$w(P) := \left(\sum_{n \in G(P)} n\right) - \frac{g(g+1)}{2}.$$

Theorem 2.2. *Let P be a point on X. Then P is a Weierstrass point if and only if $w(P) \geq 1$, and $\sum w(P)P$ belongs to the complete linear system*

$$\left|\frac{g(g+1)}{2} K_X\right|,$$

where K_X is a canonical divisor on X. In particular, we have that

$$\sum_{P \in \mathscr{W}} w(P) = g(g^2 - 1).$$

Proof. See [Farkas and Kra 1980, §III.5] or [Hindry and Silverman 2000, Exercise A.4.14]. □

2B. *Hyperellipticity.* We recall that X is a *hyperelliptic* curve if there is a degree 2 map $\phi : X \to \mathbb{P}^1$ defined over $\bar{k}$. In that case, ϕ is unique (up to automorphisms of $\mathbb{P}^1$). The map ϕ induces a degree 2 extension $\bar{k}(X)/\bar{k}(\mathbb{P}^1)$, which is Galois since $\mathrm{char}(k) = 0$. So, this gives rise to a map $\iota : X \to X$ called the *hyperelliptic involution*. We say X is hyperelliptic over k if ϕ is defined over k. The following is a well-known classical result.

Proposition 2.3. *Let X be a curve of genus $g \geq 2$ defined over a field k of characteristic 0, and $\mathscr{W}$ the set of Weierstrass points of $X(\bar{k})$. Then, we have*

$$2g + 2 \leq \#\mathscr{W} \leq g^3 - g.$$

Furthermore, X is hyperelliptic if and only if $\#\mathscr{W} = 2g + 2$. In that case, the branch points are the Weierstrass points.

Proof. See [Farkas and Kra 1980, §III.5] or [Hindry and Silverman 2000, Exercise A.4.14]. □

2C. *Galois action.* Let $\mathscr{W}$ be the set of all Weierstrass points over $X(\bar{k})$, then $\mathscr{W}$ is preserved by the action of $\mathrm{Gal}(\bar{k}/k)$. In particular, when X is a hyperelliptic curve, this action factors through the symmetric group S_{2g+2}.

3. Arithmetic Fuchsian groups

From now on, F is a totally real number field of degree g. We denote the real embeddings of F by $v_1, \ldots, v_g$. We let $\mathcal{O}_F$ be the ring of integers of F, and $\mathcal{O}_F^{\times +}$ the group of totally positive units in $\mathcal{O}_F$. We let D be a quaternion algebra defined over F, and fix a maximal order $\mathcal{O}$ in D. Let v be a place of F,

and F_v the completion of F at v. We recall that D is said to be ramified at v if $D_v = D \otimes F_v$ is a division quaternion algebra. We let S_∞ (resp. S_f) be the set of archimedean places (resp. finite places) where D is ramified; and set $S = S_\infty \cup S_f$. We let $r = \#S_f$.

3A. ***Fuchsian groups.*** From now on, we assume that D is ramified at all but one archimedean places; namely, that $S_\infty = \{v_2, \dots, v_g\}$. This means that, we have $D \otimes \mathbb{R} \simeq \mathrm{M}_2(\mathbb{R}) \times \mathbb{H}^{g-1}$, where $\mathbb{H}$ is the Hamilton quaternion algebra over $\mathbb{R}$. We let $j_1 : D \otimes_{v_1} \mathbb{R} \to \mathrm{M}_2(\mathbb{R})$ be the projection onto the factor corresponding to v_1. We will also denote the map induced on the unit groups by $j_1 : (D \otimes_{v_1} \mathbb{R})^\times \to \mathrm{GL}_2(\mathbb{R})$. For the definition of the reduced norm $\mathrm{Nrd} : D \to F$ below, we refer to [Vignéras 1980, Chapitre I, §1] or [Voight 2018, §3.3]. We let

$$\mathcal{O}^1 := \{x \in \mathcal{O} : \mathrm{Nrd}(x) = 1\}; \quad \mathcal{O}^\times := \{x \in \mathcal{O} : \mathrm{Nrd}(x) \in \mathcal{O}_F^\times\}; \quad \mathcal{O}_+^\times := \{x \in \mathcal{O} : \mathrm{Nrd}(x) \in \mathcal{O}_F^{\times +}\}.$$

We recall that the *normaliser* of $\mathcal{O}$ inside D is defined by

$$N_D(\mathcal{O}) := \{x \in D^\times : x\mathcal{O} = \mathcal{O}x\}.$$

We set

$$N_D(\mathcal{O})_+ := \{x \in N_D(\mathcal{O}) : \mathrm{Nrd}(x) \in F_+^\times\}.$$

We let Γ^1 (resp. Γ, $\Gamma_\mathcal{O}$) be the image of $\mathcal{O}^1$ (resp. $\mathcal{O}_+^\times$, $N_D(\mathcal{O})_+$) in $\mathrm{PGL}_2^+(\mathbb{R}) := \mathrm{GL}_2^+(\mathbb{R})/\mathbb{R}^\times$ via j_1, where

$$\mathrm{GL}_2^+(\mathbb{R}) := \{\gamma \in \mathrm{GL}_2(\mathbb{R}) : \det(\gamma) > 0\}.$$

We will also use the same notation to identify these groups with their respective images in $D^\times/F^\times$. We recall that Γ^1 is an *arithmetic Fuchsian group*, i.e., a discrete subgroup of $\mathrm{PSL}_2(\mathbb{R})$. The *commensurability class* of Γ^1, consists of all the subgroups $\Gamma' \subset \mathrm{PGL}_2^+(\mathbb{R})$ that are commensurable with Γ^1, i.e., such that $\Gamma' \cap \Gamma^1$ has finite index in both Γ' and Γ^1. Any Fuchsian group that is commensurable to an arithmetic Fuchsian group is itself arithmetic. So, the commensurability class of Γ^1 is independent of the embedding j_1. We define it simply as the commensurability class of $\mathcal{O}^1$ in $D^\times/F^\times$, and denote it by $\mathscr{C}(D)$. In $\mathscr{C}(D)$, one is particularly interested in those groups Γ' with minimal covolume. Borel [1981] showed that, up to conjugacy, there are finite many such groups, and gave their covolume purely in terms of the number theoretic data used in defining them. These groups are called *maximal arithmetic Fuchsian groups*, and are the main objects of interest to us in this section.

Theorem 3.1 [Borel 1981]. *Every maximal arithmetic Fuchsian group in $\mathscr{C}(D)$ is of the form $\Gamma_\mathcal{O}$, where $\mathcal{O}$ is a maximal order in D. In that case, the covolume of $\Gamma_\mathcal{O}$ is given by*

$$\mathrm{Vol}(\Gamma_\mathcal{O}\backslash\mathfrak{H}) = \frac{8\pi D_F^{3/2}\zeta_F(2)}{(4\pi^2)^g[H : F^{\times 2}]} \prod_{\mathfrak{q} \in S_f} (\mathrm{N}\,\mathfrak{q} - 1),$$

where $H = \{\mathrm{Nrd}(x) : x \in N_D(\mathcal{O})_+\}$. In particular, it depends only on F and S_f.

Proof. See [Borel 1981, §8.4]. □

3B. ***The Atkin–Lehner group.*** We define the *Atkin–Lehner group*

$$W := N_D(\mathcal{O})/F^\times \mathcal{O}^\times.$$

By the Skolem–Noether theorem [Vignéras 1980, Chapitre II, Théorème 2.1], W can be identified with the group of automorphisms of $\mathcal{O}$. It is generated by the classes $[u] \in W$ such that (u) is a principal two-sided ideal whose norm is supported at the prime ideals in S_f. By the Hasse–Schilling–Maass theorem [Vignéras 1980, Chapitre III, Théorème 5.7], W is a *finite* elementary abelian 2-group. So, there is a positive integer r such that

$$W \simeq (\mathbb{Z}/2\mathbb{Z})^r.$$

We define the *positive Atkin–Lehner groups*

$$W_+ := N_D(\mathcal{O})_+/F^\times \mathcal{O}_+^\times, \quad W^1 := N_D(\mathcal{O})/F^\times \mathcal{O}^1.$$

There is a split exact sequence

$$1 \to \mathcal{O}_F^{\times +}/(\mathcal{O}_F^\times)^2 \to W^1 \to W_+ \to 1,$$

which gives an isomorphism

$$W^1 \simeq \mathcal{O}_F^{\times +}/(\mathcal{O}_F^\times)^2 \times W_+ \simeq (\mathbb{Z}/2\mathbb{Z})^s,$$

where $s \leq (n-1)+r$. The rank s of W^1 can be determined from the Dirichlet unit theorem and the fact that the image of W_+ inside W is generated by those principal two-sided ideals whose norms are totally positive and supported at S_f.

3C. ***Optimal embeddings.*** Let E/F be a CM extension, i.e., a totally imaginary quadratic extension. By [Vignéras 1980, Chapitre III, Théorème 3.8], E embeds into D if and only if, every finite place $v \in S_f$ is ramified or inert in E. The following theorem will be very useful for us.

Theorem 3.2. *Let E/F be a* CM *extension, and $\sigma : E \hookrightarrow D$ an embedding. Let $\alpha \in E \setminus F$, and* $\mathrm{disc}(\alpha) = \mathrm{Tr}_{E/F}(\alpha)^2 - 4\,\mathrm{N}_{E/F}(\alpha)$. *Then, up to conjugation, $\sigma(\alpha) \in N_D(\mathcal{O})_+$ if and only if* $\mathrm{disc}(\alpha)/\mathrm{N}_{E/F}(\alpha) \in \mathcal{O}_F$, *and $\mathrm{N}_{E/F}(\alpha) \in F_+^\times$ is supported at S_f modulo squares.*

Proof. This follows from [Chinburg and Friedman 1999, Lemma 4.3] (see also [Maclachlan 2006, Theorem 3.1]). □

Let E/F be a CM extension, and $\mathfrak{O}$ an $\mathcal{O}_F$-order in E. An *optimal embedding* of $\mathfrak{O}$ in $\mathcal{O}$ is a homomorphism $\iota : E \hookrightarrow D$ such that $\iota(\mathfrak{O}) = \iota(E) \cap \mathcal{O}$. We denote the set of optimal embeddings of $\mathfrak{O}$ into $\mathcal{O}$ by $\mathrm{Emb}(\mathfrak{O}, \mathcal{O})$. We fix an embedding $E \hookrightarrow D$. Then, by the Skolem–Noether theorem, every embedding of E into D is of the form $(x \mapsto \alpha x \alpha^{-1})$ for some $\alpha \in D^\times$. So, we can identify $\mathrm{Emb}(\mathfrak{O}, \mathcal{O})$ with the coset space $E^\times \backslash \mathscr{E}(\mathfrak{O}, \mathcal{O})$ where

$$\mathscr{E}(\mathfrak{O}, \mathcal{O}) := \left\{\alpha \in D^\times : \alpha E \alpha^{-1} \cap \mathcal{O} = \mathfrak{O}\right\} = \left\{\alpha \in D^\times : E \cap \alpha^{-1}\mathcal{O}\alpha = \alpha^{-1}\mathfrak{O}\alpha\right\}.$$

Conjugation induces a right action of $N_D(\mathcal{O})/F^\times$ on $\mathrm{Emb}(\mathfrak{O}, \mathcal{O})$. For any subgroup $\Gamma^1 \subset \Gamma \subset N_D(\mathcal{O})/F^\times$, we let $\mathrm{Emb}(\mathfrak{O}, \mathcal{O}; \Gamma)$ be the set of Γ-conjugacy classes of optimal embeddings. Similarly, if $\mathcal{O}^1 \subset G \subset N_D(\mathcal{O})$, we let $\mathrm{Emb}(\mathfrak{O}, \mathcal{O}; G) := \mathrm{Emb}(\mathfrak{O}, \mathcal{O}; \overline{G})$, where $\overline{G}$ is the image of G in $N_D(\mathcal{O})/F^\times$. The set $\mathrm{Emb}(\mathfrak{O}, \mathcal{O}; \Gamma)$ is *finite* since Γ has finite index in $N_D(\mathcal{O})/F^\times$. The cardinality $m(\mathfrak{O}, \mathcal{O}; \Gamma)$ of this set is called the embedding number of $\mathfrak{O}$ into $\mathcal{O}$, with respect to Γ; or simply the embedding number of $\mathfrak{O}$ into $\mathcal{O}$ when $\Gamma = \mathcal{O}^\times$. There are formulae for $m(\mathfrak{O}, \mathcal{O}; \mathcal{O}^\times)$, see for example [Vignéras 1980, Chapitre II, §3 and Chapitre III, §5; Voight 2018, §30]. The following lemma can be used to get $m(\mathfrak{O}, \mathcal{O}; G)$ for any subgroup $\mathcal{O}^1 \subset G \subset \mathcal{O}^\times$.

Lemma 3.3. *Let $\mathcal{O}^1 \subset G \subset \mathcal{O}^\times$ be a subgroup. Then we have*

$$m(\mathfrak{O}, \mathcal{O}; G) = m(\mathfrak{O}, \mathcal{O}; \mathcal{O}^\times)[\mathrm{Nrd}(\mathcal{O}^\times) : \mathrm{Nrd}(G)\,\mathrm{N}_{E/F}(\mathfrak{O}^\times)].$$

Proof. See [Voight 2018, Lemma 30.3.14]. (We note that the statement in [Vignéras 1980, Chapitre III, Corollaire 5.13] is only correct with the inclusion $G \subset N_D(\mathcal{O})$ replaced by $G \subset \mathcal{O}^\times$.) □

Here we are interested in the case when $\mathcal{O}_+^\times \subset G \subset N_D(\mathcal{O})_+$. In particular, we want $\mathrm{Emb}(\mathfrak{O}, \mathcal{O}; N_D(\mathcal{O})_+)$ when $\mathcal{O}$ is a maximal order in D.

Lemma 3.4. *Let $\mathcal{O}_+^\times \subset G \subset N_D(\mathcal{O})_+$ be a subgroup. Then we have*

$$m(\mathfrak{O}, \mathcal{O}; \mathcal{O}_+^\times) = m(\mathfrak{O}, \mathcal{O}; G)\big[\mathrm{Nrd}(G) : \mathrm{Nrd}(G) \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times +}\big].$$

Proof. There is a natural surjection

$$E^\times \backslash \mathscr{E}(\mathfrak{O}, \mathcal{O})/\mathcal{O}_+^\times \to E^\times \backslash \mathscr{E}(\mathfrak{O}, \mathcal{O})/G.$$

To prove the lemma, we need to understand the fibres of this map. For $\alpha \in \mathscr{E}(\mathfrak{O}, \mathcal{O})$, the fibre of $E^\times \alpha G$ is

$$T := E^\times \backslash E^\times \alpha G/\mathcal{O}_+^\times \simeq (\alpha E^\times \alpha^{-1} \cap G)\backslash G/\mathcal{O}_+^\times.$$

It is enough to show that the cardinality of T is independent of α. To see this, we recall that the reduced norm $\mathrm{Nrd} : D_+^\times \to F_+^\times$ induces a map

$$\phi : (\alpha E^\times \alpha^{-1} \cap G)\backslash G/\mathcal{O}_+^\times \to \mathrm{Nrd}(G)/\,\mathrm{Nrd}(G) \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times +},$$

which is a bijection since $\ker(\phi) = \mathcal{O}^1 \subset \mathcal{O}_+^\times \subset G \subset N_D(\mathcal{O})_+$.

Alternatively, we can see that $\mathcal{O}_+^\times$ is a normal subgroup of G. So, we can identify $(\alpha E^\times \alpha^{-1} \cap G)\backslash G/\mathcal{O}_+^\times$ with a subgroup of W_+. This means that $\#T$ divides $\#W_+$, and is always a power of 2. □

Let $\widehat{\mathcal{O}} := \mathcal{O} \otimes \widehat{\mathbb{Z}} = \prod_{v<\infty} \mathcal{O}_v$ and $\widehat{D} := D \otimes \widehat{\mathbb{Q}}$, where $\widehat{\mathbb{Z}}$ and $\widehat{\mathbb{Q}}$ are the finite adèles of $\mathbb{Z}$ and $\mathbb{Q}$, respectively. For every finite place v, let $\mathcal{O}_v^\times \subset G_v \subset N_{D_v}(\mathcal{O}_v)$ be a subgroup, and $\widehat{G} := \prod_{v<\infty} G_v$. We would like to understand the global embedding numbers of the group $\widehat{G}$, or $G := \widehat{G} \cap D_+^\times$. Since D satisfies the Eichler condition, we have $D_+^\times \backslash \widehat{D}^\times / \widehat{\mathcal{O}}^\times \simeq \mathrm{Cl}_F^+$, where Cl_F^+ is the narrow class group of F.

Let $h = \#\mathrm{Cl}_F^+$ be the narrow class number of F, and

$$\widehat{D}^\times = \coprod_{i=1}^{h} D_+^\times g_i \widehat{\mathcal{O}}^\times,$$

where $g_i \in \widehat{D}^\times$, $i = 1, \dots, h$, and $g_1 = 1$. Then, for each i, $\mathcal{O}_i := g_i \widehat{\mathcal{O}} g_i^{-1} \cap D$ is a maximal order, and $N_D(\mathcal{O}_i) = g_i N_{\widehat{D}}(\widehat{\mathcal{O}}) g_i^{-1} \cap D$. Letting $G_i := g_i \widehat{G} g_i^{-1} \cap D_+^\times$, we have $(\mathcal{O}_i)_+^\times \subset G_i \subset N_D(\mathcal{O}_i)_+$.

For $\widehat{G} = \widehat{\mathcal{O}}^\times$, there are formulae for global optimal embeddings numbers (see [Vignéras 1980, Chapitre III, §5; Voight 2018, §30]). For $\widehat{\mathcal{O}}^\times \subset \widehat{G} \subset N_{\widehat{D}}(\widehat{\mathcal{O}})$, we have the following theorem.

Theorem 3.5. *Keeping the above notations, let $G := G_1$ and $h_{\mathfrak{O}}$ be the class number of $\mathfrak{O}$. Then we have*

$$\sum_{i=1}^{h} m(\mathfrak{O}, \mathcal{O}_i; G_i) = \frac{2h_{\mathfrak{O}}}{[H : H \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times +}]} \prod_{v<\infty} m(\mathfrak{O}_v, \mathcal{O}_v; \mathcal{O}_v^\times),$$

where $H := \mathrm{Nrd}(G)$, and $m(\mathfrak{O}_v, \mathcal{O}_v; \mathcal{O}_v^\times)$ is the local embedding number at the place v. (Here v runs over all finite places.)

Proof. By applying Lemma 3.3 with $G = \mathcal{O}_+^\times$, we have

$$\begin{aligned} m(\mathfrak{O}, \mathcal{O}; \mathcal{O}_+^\times) &= m(\mathfrak{O}, \mathcal{O}; \mathcal{O}^\times)[\mathrm{Nrd}(\mathcal{O}^\times) : \mathrm{Nrd}(\mathcal{O}_+^\times)\,\mathrm{N}_{E/F}(\mathfrak{O}^\times)] \\ &= m(\mathfrak{O}, \mathcal{O}; \mathcal{O}^\times)[\mathrm{Nrd}(\mathcal{O}^\times) : \mathcal{O}_F^{\times +}] = 2m(\mathfrak{O}, \mathcal{O}, \mathcal{O}^\times). \end{aligned}$$

The latter equality follows from the fact that D is ramified at all but one archimedean place, the norm theorem [Vignéras 1980, Chapitre III, Théorème 4.1] and the Dirichlet unit theorem.

Now we return to the situation $\mathcal{O}_+^\times \subset G \subset N_D(\mathcal{O})_+$. Combining the above identity with Lemma 3.4, we have

$$2m(\mathfrak{O}, \mathcal{O}; \mathcal{O}^\times) = m(\mathfrak{O}, \mathcal{O}; G)[\mathrm{Nrd}(G) : \mathrm{Nrd}(G) \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times +}].$$

A similar identity holds for the other maximal orders. In other words, for each maximal order $\mathcal{O}_i$, we have

$$2m(\mathfrak{O}, \mathcal{O}_i; \mathcal{O}_i^\times) = m(\mathfrak{O}, \mathcal{O}; G_i)[\mathrm{Nrd}(G_i) : \mathrm{Nrd}(G_i) \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times +}].$$

However, the group $\mathrm{Nrd}(G_i)$ is independent of i again by the norm theorem. Hence setting $H := \mathrm{Nrd}(G)$, we get

$$2m(\mathfrak{O}, \mathcal{O}_i; \mathcal{O}_i^\times) = m(\mathfrak{O}, \mathcal{O}_i; G_i)[H : H \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times +}].$$

So, we have

$$\sum_{i=1}^{h} m(\mathfrak{O}, \mathcal{O}_i; G_i) = \frac{2}{[H : H \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times +}]} \sum_{i=1}^{h} m(\mathfrak{O}, \mathcal{O}_i; \mathcal{O}_i^\times).$$

We then apply [Vignéras 1980, Chapitre III, Théorème 5.11] or [Voight 2018, Theorem 30.7.3] to conclude the proof. □

3D. ***Torsion in maximal arithmetic groups.*** From now on, we will assume that the field F has narrow class number one. However, the results discussed here can be easily adapted to any field by following [Voight 2018, §31 and §39] given that our maximal orders do *not* satisfy the selectivity condition in [Chinburg and Friedman 1999].

Since F has narrow class number one, under the assumptions of Theorem 3.5, we have

$$m(\mathfrak{O}, \mathcal{O}; G) = \frac{2h_{\mathfrak{O}}}{[H : H \cap \mathrm{N}_{E/F}(E^\times)\mathcal{O}_F^{\times+}]} \prod_{v<\infty} m(\mathfrak{O}_v, \mathcal{O}_v; \mathcal{O}_v^\times).$$

Theorem 3.6. *Let $q \geq 2$ be an integer, and e_q the number of elliptic points of order q in G. Suppose that $e_q > 0$. For $q \geq 3$, let $E = F(\zeta_q)$, where ζ_q is a primitive q-th root of unity, and let $\mathscr{S}_q$ be the set of $\mathcal{O}_F$-orders defined by*

$$\mathscr{S}_q := \big\{\mathcal{O}_F[\zeta_q] \subset \mathfrak{O} \subset \mathcal{O}_E : \#\mathfrak{O}^\times_{\mathrm{tors}} = q\big\}.$$

For $q = 2$, let $\mathscr{N}_q$ be a set of representatives for the norms of elements in G in $\mathrm{Nrd}(W_+)$, *and let $\mathscr{S}_q$ be the set of $\mathcal{O}_F$-orders defined by*

$$\mathscr{S}_q := \bigcup_{\substack{n \in \mathscr{N}_q \\ E = F(\sqrt{-n})}} \big\{\mathcal{O}_F[\sqrt{-n}] \subset \mathfrak{O} \subset \mathcal{O}_E\big\}.$$

Then the number of elliptic points of order q in G is given by

$$e_q := \frac{1}{2} \sum_{\mathfrak{O} \in \mathscr{S}_q} m(\mathfrak{O}, \mathcal{O}; G).$$

Proof. The proof is essentially an adaptation of the discussion of [Voight 2018, §39.4] (see also [Vignéras 1980, Chapitre IV, Section 2]); the only difference arises from the elliptic points that are fixed by the Atkin–Lehner group W_+. However, the number of 2-torsion elliptic elements can be computed by combining Theorem 3.2 and Section 3B. □

Remark 3.7. There seems to be very little discussion on the number of elliptic elements (or optimal embeddings) in maximal arithmetic Fuchsian groups. The only literature we could find on this topic is from Michon [1981] and Vignéras [1980, Chapitre IV, §3] for $F = \mathbb{Q}$, and Maclachlan [2006; 2009] for $[F : \mathbb{Q}] > 1$. In the latter case, however, the presentation is very different than ours. Our results are stated in a way as to draw the most parallel with optimal embeddings in Fuchsian groups, which correspond to Shimura curves, given that there is an abundance of literature in this case; see, for example, [Voight 2018, §30].

3E. ***Genus formula.*** Let Γ be a Fuchsian group of signature $(g; e_1, \ldots, e_r)$, then the quotient $\Gamma\backslash\mathfrak{H}$ is a compact Riemann surface, whose volume is given by

$$\mathrm{Vol}(\Gamma\backslash\mathfrak{H}) = 2\pi\left(2g - 2 + \sum_{i=1}^{r}\left(1 - \frac{1}{e_i}\right)\right).$$

When $\Gamma = \Gamma_{\mathcal{O}}$ is maximal in some commensurability class $\mathscr{C}(D)$, the volume depends only on F and S_f according to Theorem 3.1. In fact, it follows from [Maclachlan 2009, Corollary 5.7] that all maximal

arithmetic Fuchsian groups in the commensurability class $\mathscr{C}(D)$ have the same signature, and we can compute their genus by combining the volume formula in Theorem 3.1 with the results of Section 3D (at least when F has narrow class number one).

4. Shimura curves

We keep the notations of Section 3. Here, we summarise the necessary backgrounds on canonical models and $\mathfrak{p}$-adic uniformisation of Shimura curves. Our main references are [Boutot and Carayol 1991; Boutot and Zink 1995; Carayol 1986; Nekovář 2012; Sijsling 2013]. We view F as a subfield of $\mathbb{C}$ via the embedding $v_1 : F \hookrightarrow \mathbb{C}$.

4A. ***Complex uniformisation.*** Let $U = \prod_{\mathfrak{q}} U_{\mathfrak{q}} \subset \widehat{\mathcal{O}}^\times$ be a compact open subgroup, such that $U_{\mathfrak{p}}$ is maximal. We consider the quotient

$$X_U(\mathbb{C}) := D^\times \backslash X \times \widehat{D}^\times / U,$$

where $X := \mathbb{P}^1(\mathbb{C}) - \mathbb{P}^1(\mathbb{R}) = \mathfrak{H}^+ \sqcup \mathfrak{H}^-$, and $\mathfrak{H}^-$ and $\mathfrak{H}^+$ are the lower and upper Poincaré half-planes. Since D is a division algebra, $X_U(\mathbb{C})$ is a Riemann surface.

There is a right action of $\widehat{D}^\times$ on $X \times \widehat{D}^\times$ by conjugation. For each $g \in \widehat{D}^\times$, this induces an isomorphism of complex curves

$$X_U(\mathbb{C}) \xrightarrow{\sim} X_{g^{-1}Ug}(\mathbb{C}).$$

By the strong approximation theorem, we have the following bijections

$$D_+^\times \backslash \widehat{D}^\times / U \simeq D^\times \backslash \{\pm 1\} \times \widehat{D}^\times / U \simeq F_+^\times \backslash \widehat{F}^\times / \operatorname{Nrd}(U).$$

By class field theory, there is a unique abelian extension F_U of F such that the Artin map induces an isomorphism

$$\operatorname{Art}_F : \operatorname{Gal}(F_U/F) \simeq F_+^\times \backslash \widehat{F}^\times / \operatorname{Nrd}(U).$$

So the set $F_+^\times \backslash \widehat{F}^\times / \operatorname{Nrd}(U)$ is a Galois set. Thus there is a finite étale scheme $\mathscr{T}_U$ defined over F such that

$$\mathscr{T}_U(F_U) = \mathscr{T}_U(\bar{F}) = \mathscr{T}_U(\mathbb{C}) = F_+^\times \backslash \widehat{F}^\times / \operatorname{Nrd}(U).$$

Shimura [1970] showed that $X_U(\mathbb{C})$ admits a canonical model defined over F (see also [Deligne 1971]). Namely, we have the following result.

Theorem 4.1. *There is a curve X_U defined over F, called a* ***canonical model****, which satisfies the following properties*:

(i) *The set of complex points of X_U is $X_U(\mathbb{C})$, i.e.,*

$$(X_U \otimes_{F, v_1} \mathbb{C})(\mathbb{C}) = X_U(\mathbb{C}).$$

(ii) *For a compact open $U' \subset U$, the morphism $X_{U'}(\mathbb{C}) \to X_U(\mathbb{C})$ is induced by an F-morphism $X_{U'} \to X_U$.*

(iii) *For each $g \in \widehat{D}^\times$, the morphism $X_U(\mathbb{C}) \to X_{g^{-1}Ug}(\mathbb{C})$ is induced from a F-morphism $X_U \to X_{g^{-1}Ug}$.*

(iv) *The morphism $X_U(\mathbb{C}) \to \mathscr{T}_U(\mathbb{C})$, has connected fibres, and is induced by a morphism of F-schemes $X_U \to \mathscr{T}_U$. In particular, the group of connected component $\pi_0(X_U)$ is a finite étale group scheme over F such that $\pi_0(X_U)(\mathbb{C}) = \pi_0(X_U(\mathbb{C})) = \mathscr{T}_U(\mathbb{C})$, where $\pi_0(X_U(\mathbb{C}))$ is the group of connected components of $X_U(\mathbb{C})$.*

Proof. This is essentially a summary of the properties of canonical models of Shimura curves listed in [Carayol 1986, §1.1 and §1.2]. □

Theorem 4.1(iv) is known as the Shimura reciprocity law. It implies that X_U is an irreducible scheme, which is not geometrically irreducible in general. However, when $\mathrm{Nrd}(U) = \widehat{\mathcal{O}}_F^\times$, then X_U is geometrically irreducible since we assume that F has narrow class number one.

We define the *adelic Atkin–Lehner group* by $\widehat{W} := N_{\widehat{D}}(U)/\widehat{F}^\times U$. By making use of the weak approximation theorem, one can show that

$$\widehat{W} \simeq \prod_{\mathfrak{q} \in S_f \cup S_0} \mathbb{Z}/2\mathbb{Z},$$

where S_0 is the set of primes where $U_\mathfrak{q}$ is nonmaximal.

Corollary 4.2. *The group $\widehat{W}$ acts on $X_U(\mathbb{C})$. This action is induced from an action of $\widehat{W}$ on X_U defined over F. In particular, if $W' \subseteq \widehat{W}$ is a subgroup, then the quotient X_U/W' is defined over F.*

Proof. Every element $g \in \widehat{W}$ defines an automorphism of $X_U(\mathbb{C})$. By Theorem 4.1 (iii), this automorphism descends to F. □

When there is an integral ideal $\mathfrak{N}$ coprime with the discriminant $\mathrm{disc}(D)$ of $\mathcal{O}$, and an Eichler order $\mathcal{O}_0(\mathfrak{N}) \subset \mathcal{O}$ of level $\mathfrak{N}$ such that $U = \widehat{\mathcal{O}_0(\mathfrak{N})}^\times$, we will denote the Shimura curve X_U by $X_0^D(\mathfrak{N})$, or simply write $X_0^D(1)$ when $\mathfrak{N} = (1)$.

4B. ***Bruhat-Tits tree.*** Let $\mathcal{T}_\mathfrak{p}$ be the Bruhat-Tits tree attached to $\mathrm{GL}_2(F_\mathfrak{p})$. Its set of vertices $\mathcal{V}(\mathcal{T}_\mathfrak{p})$ consists of maximal $\mathcal{O}_{F_\mathfrak{p}}$-orders in $\mathrm{M}_2(F_\mathfrak{p})$, two vertices being adjacent if their intersection is an Eichler order of level $\mathfrak{p}$. Let $\vec{\mathcal{E}}(\mathcal{T}_\mathfrak{p})$ denote the set of ordered edges of $\mathcal{T}_\mathfrak{p}$, i.e., the set of ordered pairs (s, t) of adjacent vertices of $\mathcal{T}_\mathfrak{p}$. If $e = (s, t)$, the vertex s is called the *source* of e and the vertex t is called its *target*; they are denoted by $s(e)$ and $t(e)$ respectively.

The Atkin–Lehner involution $\iota : \vec{\mathcal{E}}(\mathcal{T}_\mathfrak{p}) \to \vec{\mathcal{E}}(\mathcal{T}_\mathfrak{p})$ sends the edge $e = (s, t)$ to the opposite edge $\bar{e}$. We let $\mathcal{E}(\mathcal{T}_\mathfrak{p}) = \vec{\mathcal{E}}(\mathcal{T}_\mathfrak{p})/\langle\iota\rangle$ be the set of nonoriented edges.

The tree $\mathcal{T}_\mathfrak{p}$ is endowed with a natural left action of $\mathrm{PGL}_2(F_\mathfrak{p})$ by isometries corresponding to conjugation of maximal orders by elements of $\mathrm{GL}_2(F_\mathfrak{p})$. This action is transitive on both $\mathcal{V}(\mathcal{T}_\mathfrak{p})$ and $\vec{\mathcal{E}}(\mathcal{T}_\mathfrak{p})$.

4C. ***$\mathfrak{p}$-adic uniformisation.*** Let $\overline{F_\mathfrak{p}}$ be an algebraic closure of $F_\mathfrak{p}$, and $\mathbb{C}_\mathfrak{p} := \widehat{\overline{F_\mathfrak{p}}}$ be a fixed completion of $\overline{F_\mathfrak{p}}$. Let $\widehat{\mathscr{H}_\mathfrak{p}}$ be $\mathfrak{p}$-adic upper half plane. This is the formal scheme over $\mathrm{Spf}(\mathcal{O}_{F_\mathfrak{p}})$ defined in [Boutot and Carayol 1991, §1.3] by

$$\widehat{\mathscr{H}_\mathfrak{p}} := \mathbb{P}^1(\mathbb{C}_\mathfrak{p}) - \mathbb{P}^1(F_\mathfrak{p}).$$

The scheme $\widehat{\mathscr{H}}_{\mathfrak{p}}$ admits a natural action by the group $\mathrm{GL}_2(F_{\mathfrak{p}})$, which factors through the adjoint group $\mathrm{PGL}_2(F_{\mathfrak{p}})$. We let

$$\widehat{\mathscr{H}}_{\mathfrak{p}}^{\mathrm{ur}} = \widehat{\mathscr{H}}_{\mathfrak{p}} \times_{\mathrm{Spf}(\mathcal{O}_{F_{\mathfrak{p}}})} \mathrm{Spf}(\mathcal{O}_{F_{\mathfrak{p}}}^{\mathrm{ur}}).$$

Let B be the totally definite quaternion algebra defined over F whose set of ramified *finite* places is $S_f \setminus \{\mathfrak{p}\}$ so that $B_{\mathfrak{p}} \simeq \mathrm{M}_2(F_{\mathfrak{p}})$. (Note that this means that the set of ramified archimedean places of B is $S_\infty \cup \{v_1\}$.) We write $\widehat{B} = B_{\mathfrak{p}} \times B^{\mathfrak{p}}$ and $\widehat{D} = D_{\mathfrak{p}} \times D^{\mathfrak{p}}$, and we fix an isomorphism $\varphi : D^{\mathfrak{p}} \xrightarrow{\sim} B^{\mathfrak{p}}$. We let $K = K_{\mathfrak{p}} \times K^{\mathfrak{p}}$ be a compact open subgroup of $\widehat{B}^\times$ such that $K_{\mathfrak{p}} \simeq \mathrm{GL}_2(\mathcal{O}_{F_{\mathfrak{p}}})$ and $\varphi(U^{\mathfrak{p}}) = K^{\mathfrak{p}}$. We also let

$$K_{\mathfrak{p}}^0 := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in K_{\mathfrak{p}} : c \equiv 0 \bmod \mathfrak{p} \right\},$$

and $K_0(\mathfrak{p}) = K_{\mathfrak{p}}^0 \times K^{\mathfrak{p}}$.

Since $U_{\mathfrak{p}}$ is the maximal compact open subgroup of $D_{\mathfrak{p}}^\times$, the norm map induces an isomorphism $D_{\mathfrak{p}}^\times / U_{\mathfrak{p}} \xrightarrow{\sim} F_{\mathfrak{p}}^\times / \mathcal{O}_{F_{\mathfrak{p}}}^\times$ (see [Vignéras 1980, Chapitre II, Lemme 1.5]). The group $B_{\mathfrak{p}}^\times$ acts on $F_{\mathfrak{p}}^\times / \mathcal{O}_{F_{\mathfrak{p}}}^\times$ through its reduced norm map $\mathrm{Nrd} : B_{\mathfrak{p}}^\times \to F_{\mathfrak{p}}^\times$. We obtain a corresponding action of $B_{\mathfrak{p}}^\times$ on $D_{\mathfrak{p}}^\times / U_{\mathfrak{p}}$. This, together with the isomorphism φ, gives an action of $\widehat{B}^\times$ on $\widehat{D}^\times / U$.

Theorem 4.3 (Čerednik–Drinfel'd). *There exist a model $\mathscr{M}$ of X_U over $\mathcal{O}_{F_{\mathfrak{p}}}$, and an isomorphism of formal schemes*

$$\widehat{\mathscr{M}}^{\mathrm{ur}} = \widehat{\mathscr{M}} \times_{\mathrm{Spf}(\mathcal{O}_{F_{\mathfrak{p}}})} \mathrm{Spf}(\mathcal{O}_{F_{\mathfrak{p}}}^{\mathrm{ur}}) \simeq B^\times \backslash \widehat{\mathscr{H}}_{\mathfrak{p}}^{\mathrm{ur}} \times \widehat{D}^\times / U,$$

where $\widehat{\mathscr{M}}$ is the completion of $\mathscr{M}$ along its special fibre.

Proof. See [Boutot and Zink 1995, Theorem 3.1]. □

4D. ***The dual graph.*** The *dual graph* associated to $B^\times \backslash \widehat{\mathscr{H}}_{\mathfrak{p}}^{\mathrm{ur}} \times \widehat{D}^\times / U$ is the weighted graph

$$\mathcal{G} := B^\times \backslash \mathcal{T}_{\mathfrak{p}} \times \widehat{D}^\times / U.$$

The vertices of $\mathcal{V}(\mathcal{G})$ and oriented edges $\vec{\mathcal{E}}(\mathcal{G})$ of $\mathcal{G}$ are given respectively by

$$\mathcal{V}(\mathcal{G}) := B^\times \backslash \mathcal{V}(\mathcal{T}_{\mathfrak{p}}) \times \widehat{D}^\times / U \quad \text{and} \quad \vec{\mathcal{E}}(\mathcal{G}) := B^\times \backslash \vec{\mathcal{E}}(\mathcal{T}_{\mathfrak{p}}) \times \widehat{D}^\times / U.$$

We define the weight of a vertex $v \in \mathcal{V}(\mathcal{G})$ to be $\#\mathrm{Stab}_{B^\times / F^\times}(v)$, and the weight of an edge $e \in \vec{\mathcal{E}}(\mathcal{G})$ to be $\#\mathrm{Stab}_{B^\times / F^\times}(e)$. For a vertex v, we let $\mathrm{Star}(v)$ denote the set of all edges containing v.

Proposition 4.4. *The maps*

$$\begin{aligned} \vartheta_1 : (B_{\mathfrak{p}}^\times / F_{\mathfrak{p}}^\times K_{\mathfrak{p}}) \times (D_{\mathfrak{p}}^\times / U_{\mathfrak{p}}) \times (D^{\mathfrak{p}\times} / U^{\mathfrak{p}}) &\to (B_{\mathfrak{p}}^\times / K_{\mathfrak{p}}) \times \mathbb{Z} \times (B^{\mathfrak{p}\times} / K^{\mathfrak{p}}) \\ (x_{\mathfrak{p}}, y_{\mathfrak{p}}, y^{\mathfrak{p}}) &\mapsto (x_{\mathfrak{p}}, \mathrm{ord}_{\mathfrak{p}}(\mathrm{Nrd}(y_{\mathfrak{p}})), \varphi(y^{\mathfrak{p}})), \\ \vartheta_2 : (B_{\mathfrak{p}}^\times / F_{\mathfrak{p}}^\times K_{\mathfrak{p}}^0) \times (D_{\mathfrak{p}}^\times / U_{\mathfrak{p}}) \times (D^{\mathfrak{p}\times} / U^{\mathfrak{p}}) &\to (B_{\mathfrak{p}}^\times / K_{\mathfrak{p}}^0) \times \mathbb{Z} \times (B^{\mathfrak{p}\times} / K^{\mathfrak{p}}) \\ (x_{\mathfrak{p}}, y_{\mathfrak{p}}, y^{\mathfrak{p}}) &\mapsto (x_{\mathfrak{p}}, \mathrm{ord}_{\mathfrak{p}}(\mathrm{Nrd}(y_{\mathfrak{p}})), \varphi(y^{\mathfrak{p}})) \end{aligned}$$

induce an isomorphism of bipartite graphs

$$\mathcal{V}(\mathcal{G}) = B^\times\backslash\mathcal{V}(\mathcal{T}_\mathfrak{p}) \times \widehat{D}^\times/U \xrightarrow{\sim} (B^\times\backslash\widehat{B}^\times/K) \times \mathbb{Z}/2\mathbb{Z},$$
$$\vec{\mathcal{E}}(\mathcal{G}) = B^\times\backslash\vec{\mathcal{E}}(\mathcal{T}_\mathfrak{p}) \times \widehat{D}^\times/U \xrightarrow{\sim} (B^\times\backslash\widehat{B}^\times/K_0(\mathfrak{p})) \times \mathbb{Z}/2\mathbb{Z}$$

as follows: *we write* $\mathcal{V}(\mathcal{G}) = \mathcal{V} \sqcup \mathcal{V}' \simeq B^\times\backslash\widehat{B}^\times/K \sqcup B^\times\backslash\widehat{B}^\times/K$, *and we let the adjacency matrix in the basis* $\mathcal{V} \cup \mathcal{V}'$ *be given by the matrix*

$$\begin{bmatrix} 0 & T_\mathfrak{p} \\ T_\mathfrak{p} & 0 \end{bmatrix},$$

where $T_\mathfrak{p}$ *is the Hecke operator at* $\mathfrak{p}$ *acting on the Brandt module* $M := \mathbb{Z}[B^\times\backslash\widehat{B}^\times/K]$. *In that identification, the action of the Atkin–Lehner involution* $w_\mathfrak{p}$ *on* $\mathcal{V}(\mathcal{G})$ *is given by the matrix*

$$\begin{bmatrix} 0 & \mathbf{1}_M \\ \mathbf{1}_M & 0 \end{bmatrix}.$$

Proof. See [Sijsling 2013, Propositions 3.1.8 and 3.1.9], [Nekovář 2012, §1.5] or [Kurihara 1979, §4]. □

Remark 4.5. In the isomorphism of Proposition 4.4, the set of nonoriented edges is given by

$$\mathcal{E}(\mathcal{G}) = B^\times\backslash\mathcal{E}(\mathcal{T}_\mathfrak{p}) \times \widehat{D}^\times/U \simeq B^\times\backslash\widehat{B}^\times/K_0(\mathfrak{p}).$$

The following result is an essential ingredient in the description of the special fibre of the Čerednik–Drinfel'd model described in Theorem 4.3. As we will see later, it is also useful in understanding the automorphism group of the curve X_U.

Theorem 4.6. *Let* $\mathscr{M}$ *be the scheme in Theorem 4.3. Then, we have the following*:

(i) $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} \mathcal{O}_{F_{\mathfrak{p}^2}}$ *is a normal, proper, flat and semistable scheme over* $\mathcal{O}_{F_{\mathfrak{p}^2}}$.

(ii) *The special fibre of* $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} \mathcal{O}_{F_{\mathfrak{p}^2}}$ *is reduced. Its components are rational curves, and all its singular points are ordinary double points.*

(iii) *The weighted dual graph associated to* $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} \mathcal{O}_{F_{\mathfrak{p}^2}}$ *is the graph* $\mathcal{G}$ *described in Proposition 4.4.*

(iv) *Let* $\mathcal{H}$ *be a connected component of* $\mathcal{G}$, *and* $\mathscr{M}_\mathcal{H}$ *the corresponding irreducible component of* $\mathscr{M}$. *Then the arithmetic genus of* $\mathscr{M}_\mathcal{H}$ *is given by the Betti number* $1 + \#\mathcal{E}(\mathcal{H}) - \#\mathcal{V}(\mathcal{H})$.

Proof. See [Nekovář 2012, Proposition 1.5.5] or [Kurihara 1979, Proposition 3.2]. □

4E. ***Special fibre of*** $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} \mathcal{O}_{F_{\mathfrak{p}^2}}$**.** The curve $\mathscr{M}$ is an *admissible* curve over $\mathcal{O}_{F_\mathfrak{p}}$ in the following sense:

(i) $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} \mathcal{O}_{F_{\mathfrak{p}^2}}$ is a normal, proper, flat and semistable scheme over $\mathcal{O}_{F_{\mathfrak{p}^2}}$. Each irreducible component has a smooth generic fibre.

(ii) The completion of the local ring of $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} \mathcal{O}_{F_{\mathfrak{p}^2}}$ at each of its singular points x is isomorphic, as an $\mathcal{O}_{F_\mathfrak{p}}$-algebra, to $\mathcal{O}_{F_\mathfrak{p}}[[X, Y]]/(XY - \varpi_\mathfrak{p}^w)$, where $\varpi_\mathfrak{p}$ is a uniformising element at $\mathfrak{p}$, and $w = w(x) \in \{1, 2, 3, \ldots\}$.

(iii) The special fibre $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} k(\mathfrak{p})$ is reduced; the normalisation of each of its irreducible components is isomorphic to $\mathbb{P}^1(k(\mathfrak{p}))$; its only singular points are ordinary double points, where $k(\mathfrak{p})$ is the residue field of $\mathcal{O}_{F_{\mathfrak{p}^2}}$.

The dual graph encodes the following combinatoric data of the special fibre.

(iv) Each vertex $v \in \mathcal{V}(\mathcal{G})$ corresponds to an irreducible component C_v of the special fibre $\mathscr{M} \otimes_{\mathcal{O}_{F_\mathfrak{p}}} k(\mathfrak{p})$.

(v) Each edge $e = \{v, v'\} \in \mathcal{E}(\mathcal{G})$ corresponds to a singular point in $x_e \in C_v \cap C_{v'}$. The completion of local ring at x_e is of the form $\mathcal{O}_{F_\mathfrak{p}}[[X, Y]]/(XY - \varpi_\mathfrak{p}^w)$, where $w = w(e)$ is the weight of the edge e.

The above description can be found in [Nekovář 2012; Kurihara 1979].

4F. ***Automorphism groups.*** An *automorphism of weighted graph* $\mathcal{G}$ is an automorphism of graphs which preserves the weights of the edges. We will denote the group of such automorphisms by $\mathrm{Aut}(\mathcal{G})$. We note that there is a natural inclusion $\mathrm{Aut}(\mathcal{G}) \subset \mathrm{Aut}^s(\mathcal{G})$, where $\mathrm{Aut}^s(\mathcal{G})$ is the automorphism group of the underlying simple graph to $\mathcal{G}$.

For the next statement, we recall the notion of admissibility from [Kontogeorgis and Rotger 2008]. We say that an element $\omega \in \mathrm{Aut}(\mathcal{G})$ is *admissible* if there is no vertex $v \in \mathcal{V}(\mathcal{G})$ fixed by ω such that $\mathrm{Star}(v)$ has at least 3 edges also fixed by ω. We say that a subgroup $H \subset \mathrm{Aut}(\mathcal{G})$ is *admissible* if every nontrivial element $\omega \in H$ is admissible.

Proposition 4.7. *Let $W' \subset \widehat{W}$ be a subgroup. Then, we have the following*:

(1) *The dual graph of $(\mathscr{M}/W') \otimes_{\mathcal{O}_{F_\mathfrak{p}}} \mathcal{O}_{F_{\mathfrak{p}^2}}$ is the graph $\mathcal{G}' = (\mathcal{G}/W')^*$, where * means we remove all loops from the quotient graph $\mathcal{G}/W'$.*

(2) *Assume that the genus of X_U/W' is at least 2, and let $\mathcal{G}_{st}$ be the dual graph of the stable model $(\mathscr{M}/W')_{st}$ of $\mathscr{M}/W'$. Then there is a natural injection $\varrho : \mathrm{Aut}(X_U/W') \hookrightarrow \mathrm{Aut}(\mathcal{G}_{st})$ whose image $\mathrm{im}(\varrho)$ lies in an admissible subgroup.*

Proof. Part (1) follows from general properties of Mumford curves. From [Deligne and Mumford 1969, Lemmas 1.12 and 1.16], and universal properties of stable models, there is an injection $\varrho : \mathrm{Aut}(X_U/W') \hookrightarrow \mathrm{Aut}(\mathcal{G}_{st})$. To prove Part (2), we only need to show that every nontrivial element in the image of ϱ is admissible. To this end, let $\omega \in \mathrm{Aut}(X_U/W')$ be such that $\varrho(\omega)$ fixes a vertex v, and at least 3 edges in $\mathrm{Star}(v)$. Then, since every automorphism of the projective line, which fixes at least 3 points is the identity, the restriction $\omega|_{C_v}$ is the identity, where C_v is the irreducible component associated to v. This would imply that, as an automorphism of the Riemann surface $(X_U/W')(\mathbb{C})$, ω fixes more than $2g(X_U/W') + 2$ points, where $g(X_U/W')$ is the genus of X_U/W'. Hence ω must be the identity. Therefore, if ω is nontrivial, then $\varrho(\omega)$ must be admissible. □

5. The hyperelliptic Shimura quotient curve

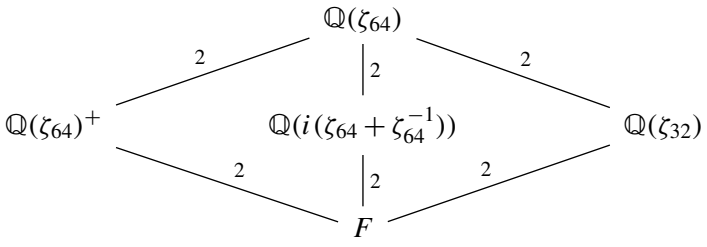
5A. ***The quaternion algebra.*** Let $F = \mathbb{Q}(\alpha) = \mathbb{Q}(\zeta_{32} + \zeta_{32}^{-1})$ be the maximal totally real subfield of the cyclotomic field of the 32-nd roots of unity. This field is defined by the polynomial $x^8 - 8x^6 + 20x^4 - 16x^2 + 2$.

Let σ be a generator of $\mathrm{Gal}(F/\mathbb{Q})$. Let $\mathcal{O}_F$ be the ring of integers of F. Let $v_1, \ldots, v_8$ be the real places of F. We consider the quaternion algebra D/F ramified at $v_2, \ldots, v_8$ and the unique prime $\mathfrak{p}$ above 2. More concretely, we have $D = \left(\frac{u,-1}{F}\right)$, where $u = -\alpha^2 + \alpha$ has signature $(+, -, \ldots, -)$. Let $\mathcal{O}$ be the maximal order in D given by

$$\mathcal{O} := \mathcal{O}_F\left[1, i, \frac{(\alpha^7+\alpha^6+\alpha^4+1)+\alpha^7 i + j}{2}, \frac{(\alpha^7+\alpha^6+\alpha^4+1)i + k}{2}\right].$$

We also let B/F be the totally definite quaternion algebra ramified exactly at all the real places $v_1, \ldots, v_8$, and fix a maximal order $\mathcal{O}_B$ in B. Both these orders were computed using the quaternion algebras package in `Magma` [1997] implemented by Voight [2005].

5B. ***The CM field and its embedding.*** We recall the following diagram:



The subfield $K := \mathbb{Q}(\beta) = \mathbb{Q}(i(\zeta_{64} + \zeta_{64}^{-1}))$ is the unique CM extension of F with class number 17. For later, we observe that $\beta^2 = -2 - \alpha$, where $\mathfrak{p} = (2 + \alpha)$. Since $\mathfrak{p}$ is the unique prime of F that ramifies in both K and D, we see that K is a splitting field of D by [Vignéras 1980, Chapitre III, Théorème 3.8]. It is possible to compute an explicit embedding $K \hookrightarrow D$ using the quaternion algebras package in `Magma` (see [Voight 2005]), but we will not need such a map here.

5C. ***The spaces of forms.*** Let $S_2(\mathfrak{p})^{\mathrm{new}}$ be the new subspace of Hilbert cusp forms of level $\mathfrak{p}$ and weight 2. This is a 40-dimensional space. Let $S_2^D(1)$ be the space of automorphic forms of level (1) and weight 2 on D, and let $S_2^B(\mathfrak{p})^{\mathrm{new}}$ be the new subspace of automorphic forms of level $\mathfrak{p}$ and weight 2 on B. By the Jacquet–Langlands correspondence, we have isomorphisms of Hecke modules

$$S_2(\mathfrak{p})^{\mathrm{new}} \simeq S_2^D(1) \simeq S_2^B(\mathfrak{p})^{\mathrm{new}}.$$

The space $S_2(\mathfrak{p})^{\mathrm{new}}$ decomposes into 5 Hecke constituents of dimensions 4, 4, 4, 4 and 24 respectively. (We note that all the computations have been performed using the Hilbert modular forms package in `Magma`, the algorithms are described in [Dembélé and Donnelly 2008; Dembélé and Voight 2013; Greenberg and Voight 2011].) There are choices of newforms f, f', g, g' and h in those constituents such that we have:

(i) The forms f and f' have the same coefficient field $L_f = L_{f'}$, which is the real quartic field $\mathbb{Q}(\zeta_{15})^+$ given by $x^4 + x^3 - 4x^2 - 4x + 1$. They satisfy the relations ${}^{\sigma}f = f'$ and ${}^{\sigma^2}f = f^{\tau}$, where τ is a generator of $\mathrm{Gal}(L_f/\mathbb{Q})$.

(ii) The forms g and g' have the same coefficient field $L_g = L_{g'}$, which is the real quartic subfield of $\mathbb{Q}(\zeta_{95})^+$ given by $x^4 + 19x^3 - 59x^2 + 19x + 1$. They satisfy the relations ${}^{\sigma}g = g'$ and ${}^{\sigma^2}g = g^{\tau}$, where τ is a generator of $\mathrm{Gal}(L_g/\mathbb{Q})$.

(iii) The coefficient field of the form h is a field L_h of degree 24, which is cyclic over the field $K_h = \mathbb{Q}(c)$ defined by $c^3 + c^2 - 229c + 167 = 0$. More precisely, it is the ray class field of conductor $\mathfrak{c} = \left(\frac{1}{2}(c^2 - 16c + 25)\right)$. The form h satisfies the relation ${}^{\sigma}h = h^{\tau}$, where τ is a generator of $\mathrm{Gal}(L_h/K_h)$.

(We summarise that data in Table 1, and the relations among the forms.) Let w and w_D be the Atkin–Lehner involutions acting on $S_2(\mathfrak{p})^{\mathrm{new}}$ and $S_2^D(1)$, respectively. The Atkin–Lehner involution w acts as follows:

$$wf = -f, \quad wf' = -f', \quad wg = -g, \quad wg' = -g', \quad wh = h.$$

We recall that $w_D = -w$.

5D. ***The Shimura curve and its quotient.*** Let $X_0^D(1)$ be the Shimura curve attached to $\mathcal{O}$. Let w_D be the Atkin–Lehner involution at $\mathfrak{p}$, and $C := X_0^D(1)/\langle w_D \rangle$. We can canonically identify $S_2^D(1)$ with the space of 1-differential forms on $X_0^D(1)$. From the discussion in Section 5C, it follows that $X_0^D(1)$ is a curve of genus 40; and that C is a curve of genus 16.

Theorem 5.1. *The curves $X_0^D(1)$ and C have the respective signatures* $(40; 3^{18}, 16^1)$ *and* $(16; 2^{17}, 3^9, 32^1)$.

Proof. The complex points of the curve $X_0^D(1)$ are determined by the quotient $\Gamma^1 \backslash \mathfrak{H}$, where Γ^1 is the image of $\mathcal{O}^1$ inside $\mathrm{PSL}_2(\mathbb{R})$. So it is a Shimura curve. So, we can compute the signature of $X_0^D(1)$ using the Shimura curves package in `Magma`, which was implemented by Voight [2009]. This gives that $X_0^D(1)$ has signature $(40; 3^{18}, 16^1)$.

The curve $C = X_0^D(1)/\langle w_D \rangle$ is given by the maximal arithmetic Fuchsian group $\Gamma_{\mathcal{O}}$. It is not a Shimura curve. Although Voight has implemented algorithms for computing with maximal arithmetic Fuchsian groups, they are not publicly available yet. So, we compute the signature of C by using the results of Section 3.

Let $q > 2$ be an integer. Then, by Theorem 3.2, $\Gamma_{\mathcal{O}}$ contains an elliptic element of order q if and only if the following three conditions are satisfied:

(i) $2\cos(2\pi/q) \in F$;

(ii) No prime $\mathfrak{q} \in S_f$ splits in $E = F(\zeta_q)$;

(iii) The ideal generated by $2 + 2\cos(2\pi/q)$ is supported at S_f modulo squares.

It is enough to test all integers q between 3 and 64. The only $q \geq 3$ which satisfy these three conditions are 3, 4, 6, 8, 16 and 32.

For $q = 4$, 8, 16 or 32, we have $E = F(\zeta_q) = \mathbb{Q}(\zeta_{32})$. In that case, the only $\mathcal{O}_F$-order which contains $\mathcal{O}_F[\zeta_{32}]$ and optimally embeds into D is the maximal order $\mathcal{O}_E$. By Theorem 3.6, we get that $e_{32} = 1$.

Newform	Coefficient field L_f	Fixed field $K_f = L_f^{\Delta}$	$\mathrm{Gal}(L_f/K_f)$
$f,\ f'$	$\mathbb{Q}(\zeta_{15})^+$	$\mathbb{Q}$	$\mathbb{Z}/4\mathbb{Z}$
$g,\ g'$	Quartic subfield of $\mathbb{Q}(\zeta_{95})^+$	$\mathbb{Q}$	$\mathbb{Z}/4\mathbb{Z}$
h	Ray class field of modulus $\mathfrak{c} = \left(\frac{1}{2}(c^2 - 16c + 25)\right)$	$\mathbb{Q}(c) := \mathbb{Q}[x]/(r(x))$, $r = x^3 + x^2 - 229x + 167$	$\mathbb{Z}/8\mathbb{Z}$
Relations	${}^{\sigma}f = f'$ and ${}^{\sigma^2}f = f^{\tau}$	${}^{\sigma}g = g'$ and ${}^{\sigma^2}g = g^{\tau}$	${}^{\sigma}h = h^{\tau}$

Table 1. Newforms of level $\mathfrak{p}$ and weight 2 on $F = \mathbb{Q}(\zeta_{32})^+$

For $q = 3$, we have $E = F\left(\frac{1}{2}(1+\sqrt{-3})\right)$. In that case, the only $\mathcal{O}_F$-order which contains $\mathcal{O}_F\left[\frac{1}{2}(1+\sqrt{-3})\right]$ and optimally embeds into D is also the maximal order $\mathfrak{O} := \mathcal{O}_E$. We have $h_{\mathfrak{O}} = 9$. Now since the prime $\mathfrak{p}$ is inert in the relative extension E/F, we have $[H : H \cap \mathrm{N}_{E/F}(E^{\times})\mathcal{O}_F^{\times +}] = 2$. So, by Theorem 3.6, we get that $e_3 = 9$.

Finally, for $q = 2$, we have $W^1 = W_+ = \mathbb{Z}/2\mathbb{Z}$ since F has narrow class number one and there is a unique prime in S_f; namely, the prime $\mathfrak{p}$ above 2. So the unique CM extension E/F which satisfies the condition of Theorem 3.2 is the extension K discussed in Section 5B. Recall that the ideal $\mathfrak{p}$ is generated by the totally positive element $n = 2 + \alpha$. The only $\mathcal{O}_F$-order which contains $\mathcal{O}_F[\sqrt{-n}]$ and optimally embeds into D is also the maximal order $\mathfrak{O} := \mathcal{O}_K$. We have $h_{\mathfrak{O}} = 17$. Now since the prime $\mathfrak{p}$ is ramified in the relative extension K/F, we have $[H : H \cap \mathrm{N}_{E/F}(E^{\times})\mathcal{O}_F^{\times +}] = 1$. So, by Theorem 3.6, we get that $e_2 = 17$.

So we conclude that there are 3 classes of elliptic elements in $\Gamma_{\mathcal{O}}$ of orders 2, 3 and 32, with respective multiplicities 17, 9 and 1.

By the volume formula in Theorem 3.1 and the genus formula in Section 3E, the genus g of the curve C must satisfy the equality

$$\frac{\mathrm{Vol}(\Gamma_{\mathcal{O}} \backslash \mathfrak{H})}{2\pi} = \frac{1455}{32} = 2(g-1) + 17\left(1 - \frac{1}{2}\right) + 9\left(1 - \frac{1}{3}\right) + \left(1 - \frac{1}{32}\right).$$

Solving this, we get that $g = 16$. Hence the curve C has signature $(16; 2^{17}, 3^9, 32^1)$. □

Lemma 5.2. *The curve $X_0^D(1)$ and the Atkin–Lehner involution w_D are both defined over $\mathbb{Q}$. In particular, the curve C descends to $\mathbb{Q}$.*

Proof. Since $\sigma(\mathfrak{p}) = \mathfrak{p}$ and the ray class group of modulus $\mathfrak{p}v_2 \cdots v_8$ is trivial, the curve $X_0^D(1)$ is defined over F by [Doi and Naganuma 1967, Corollary], and the field of moduli is $\mathbb{Q}$. The field $\mathbb{Q}(\zeta_{32})$ is a splitting field for D whose class number is one. So, there is a unique CM point attached to the extension $\mathbb{Q}(\zeta_{32})/F$, and it is defined over F. Therefore, by [Sijsling and Voight 2016, Corollary 1.9], the curve $X_0^D(1)$ descends to $\mathbb{Q}$.

Alternatively, by using the moduli interpretation in [Carayol 1986], or the more recent work [Tian and Xiao 2016], one can show that both $X_0^D(1)$ and w_D are defined over $\mathbb{Q}$. □

5E. ***The dual graph of the quotient curve.*** The dual graph $\mathcal{G}'$ of the curve $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$ is displayed in Figure 1. It was computed by using Propositions 4.4 and 4.7. The computations combine both `Magma` [1997] and `Sage` [2019].

Lemma 5.3. *The automorphism group of* $\mathcal{G}'$ *is* $\mathrm{Aut}(\mathcal{G}') \simeq \mathbb{Z}/4\mathbb{Z} \ltimes (\mathbb{Z}/2\mathbb{Z})^4$.

Proof. We compute the dual graph $\mathcal{G}'$ of the quotient $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$ using Propositions 4.4 and 4.7. Let B be the definite quaternion algebra defined in Section 5A. Then, by Propositions 4.4 and 4.7, the dual graphs $\mathcal{G}$ and $\mathcal{G}'$ are determined by the Brandt module $M_B := \mathbb{Z}[B^\times \backslash \widehat{B}^\times / \widehat{\mathcal{O}}_B^\times]$. In this case, the class number of the maximal order $\mathcal{O}_B$ is 58, and a basis of this module is given by equivalence classes of $\mathcal{O}_B$-right ideals. We let $v_1, v_2, \ldots, v_{58}$ be such a basis, which we order so that the weights of the elements are in decreasing order. We get the following sequence of weights: 32, 24, 16, 8^2, 4^3, 3^4, 2^6 and 1^{40}, where the exponent indicates the number of times each weight is repeated. Similarly, we compute the set of edges, and we obtain the following sequence for their weights: 32, 16^2, 8^6, 4^6, 2^{12} and 1^{128}. By combining this with the Hecke operator $T_{\mathfrak{p}}$, we obtain the graph $\mathcal{G}'$ in Figure 1.

We compute the automorphism group $\mathrm{Aut}^s(\mathcal{G}')$ of the underlying simple graph using `Magma`, and check that every element in $\mathrm{Aut}^s(\mathcal{G}')$ preserves the weights of the edges, i.e., that $\mathrm{Aut}(\mathcal{G}') = \mathrm{Aut}^s(\mathcal{G}')$.

To determine the group structure of $\mathrm{Aut}(\mathcal{G}')$, we first check that there is a unique normal subgroup of $\mathrm{Aut}(\mathcal{G}')$ which is isomorphic to $(\mathbb{Z}/2\mathbb{Z})^4$. Finally, we show that there is a unique cyclic subgroup of order 4 whose intersection with $(\mathbb{Z}/2\mathbb{Z})^4$ is the neutral element. □

Remark 5.4. As a byproduct of the computation of $\mathcal{G}'$, we check that

$$1 + \#\mathcal{E}(\mathcal{G}') - \#\mathcal{V}(\mathcal{G}') = 1 + 73 - 58 = 16,$$

which is the genus of $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$, or equivalently C.

Let $(\mathscr{M}/\langle w_{\mathfrak{p}}\rangle)_{st}$ be the stable model of $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$, and $\mathcal{G}_{st}$ its dual graph. By definition, $(\mathscr{M}/\langle w_{\mathfrak{p}}\rangle)_{st}$ is stable if for all $v \in \mathcal{V}(\mathcal{G}_{st})$, we have $\#\,\mathrm{Star}(v) \geq 3$. So, we obtain $(\mathscr{M}/\langle w_{\mathfrak{p}}\rangle)_{st}$ by blowing down all components C_v associated to a vertex v such that $\#\,\mathrm{Star}(v) < 3$. On graphs, this corresponds to doing the following:

(a) For all $v \in \mathcal{V}(\mathcal{G}')$, with $\#\,\mathrm{Star}(v) = 1$, remove v and all edges in $\mathrm{Star}(v)$.

(b) For each $v \in \mathcal{V}(\mathcal{G}')$, with $\#\,\mathrm{Star}(v) = 2$, contract the chain $v' \overset{e'}{-} v \overset{e''}{-} v''$ to $v' \overset{e}{-} v''$ with $w(e) = w(e') + w(e'')$.

By applying this process to the curve $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$, and then relabelling the resulting graph, we obtain the stable model whose dual graph $\mathcal{G}_{st}$ is given by Figure 2. The graph $\mathcal{G}_{st}$ has 30 vertices and 45 edges so that

$$1 + \#\mathcal{E}(\mathcal{G}_{st}) - \#\mathcal{V}(\mathcal{G}_{st}) = 1 + 45 - 30 = 16.$$

Lemma 5.5. *Let* $(\mathscr{M}/\langle w_{\mathfrak{p}}\rangle)_{st}$ *be the stable model of* $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$, *and* $\mathcal{G}_{st}$ *its dual graph. Then,* $\mathcal{G}_{st}$ *is a connected graph such that* $\mathrm{Aut}(\mathcal{G}_{st}) = \mathrm{Aut}(\mathcal{G}')$.

Proof. This follows from a direct calculation. □

Lemma 5.6. *Every admissible subgroup of* $\mathrm{Aut}(\mathcal{G}_{st})$ *of exponent* 2 *has order* 2.

Proof. First, we note that, since the degree of the Hecke operator $T_{\mathfrak{p}}$ is 3, and $(\mathscr{M}/\langle w_{\mathfrak{p}}\rangle)_{st}$ is stable, $\#\,\mathrm{Star}(v) = 3$ for each $v \in \mathcal{V}(\mathcal{G}_{st})$.

In the notations of Figure 2, we label the vertices $1, 2, \ldots, 30$. There are 19 permutations of order 2 in $\mathrm{Aut}(\mathcal{G}_{st}) \subset S_{30}$. Of those 19 permutations, there are exactly 4 with the same support of length 28. Each of the remaining 17 has a support whose length belongs to $\{2, 4, 6, 8\}$. The permutations of length 28 fix the vertices $v = 1$ and $v' = 2$. So, they must be admissible since a nonadmissible element must fix at least 4 different vertices. For each of remaining 17 permutations, one easily sees that the complement of its support contains a vertex v and its $\mathrm{Star}(v)$, meaning that it cannot be admissible.

To conclude the proof of the lemma, we let σ_i, $i = 1, 2, 3, 4$ be the 4 admissible permutations obtained above, and we check that $\sigma_i\sigma_j$ is not admissible for $i \neq j$. □

Lemma 5.7. *There is an injection* $\mathrm{Aut}(C) \hookrightarrow H$ *into an admissible subgroup of* $\mathrm{Aut}(\mathcal{G}_{st})$ *of exponent* 2. *In particular* $\mathrm{Aut}(C)$ *has order at most* 2.

Proof. In Section 5G, we will show that the endomorphism ring of each of the simple factor of $\mathrm{Jac}(C)$ is a totally real field. (This follows from the decomposition (2).) Using this, we see that $\mathrm{Aut}(C) \subset (\mathbb{Z}/2\mathbb{Z})^4$. So, by Proposition 4.7, $\mathrm{Aut}(C)$ injects into an admissible subgroup H of $\mathrm{Aut}(\mathcal{G}_{st})$ of exponent 2. By Lemma 5.6, H has order at most 2. □

Remark 5.8. The graph $\mathcal{G}'$ of the integral model $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$ (see Figure 1) is an example of a graph whose automorphism group does *not* have an element that is admissible. Indeed, it is easy to see that every element of $\mathrm{Aut}(\mathcal{G}')$ must fix the vertex v_4 and the 3 edges of weight 8 contained in $\mathrm{Star}(v_4)$. However, the vertex v_4 and $\mathrm{Star}(v_4)$ are removed when we blow down $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$ to obtain the stable model $(\mathscr{M}/\langle w_{\mathfrak{p}}\rangle)_{st}$ (see Figure 2). This example shows that [Kontogeorgis and Rotger 2008, Proposition 3.4] is incorrect as stated and needs to be modified slightly.

5F. ***Hyperellipticity of the curve C.*** We are now ready to prove one of our main results.

Theorem 5.9. *The curve C is hyperelliptic over F.*

Proof. Let $\gamma \in \Gamma_{\mathcal{O}}$ be an elliptic element of order 2, and P a fixed point by γ. Then P is a CM point by construction, and γ acts on the local ring $\mathcal{O}_{C,P}$ as an involution. More specifically, letting t be a uniformiser at P, we see that γ acts on t as

$$t \pmod{t^2} \mapsto -t \pmod{t^2}.$$

This forces any global differential form in $H^0(C, \Omega^1_C)$, which vanishes at P, to vanish to *even* order. We claim that this implies that P is a Weierstrass point. To prove this, we use Riemann–Roch.

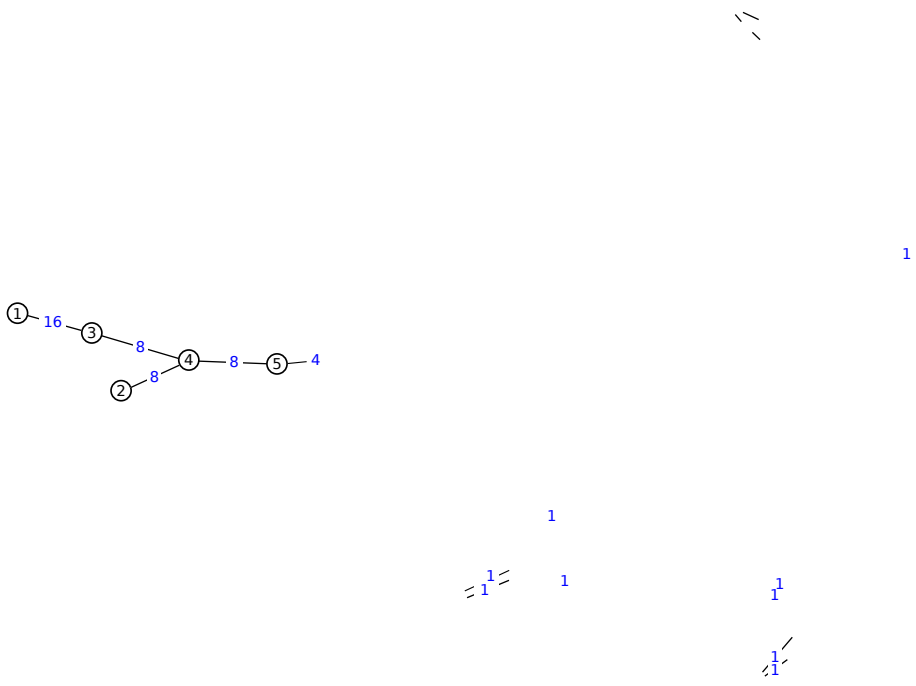


Figure 1. The dual graph $\mathcal{G}'$ of the quotient curve $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$.

Let K_C the canonical divisor. Then, we have $\ell(K_C - 2P) = \ell(K_C - P)$, i.e., every differential that vanishes at P vanishes to order 2. By Riemann–Roch, we have

$$\ell(K_C - 2P) - \ell(2P) = \deg(K_C - 2P) - g + 1 = (2g-4) - g + 1 = g - 3;$$

and

$$\ell(K_C - P) - \ell(P) = \deg(K_C - P) - g + 1 = (2g-3) - g + 1 = g - 2.$$

So, if $\ell(K_C - 2P) = \ell(K_C - P)$, then $\ell(2P) - \ell(P) = 1$. Now, since $\mathscr{L}(P)$ is the space of constant functions, we see that $\mathscr{L}(2P)$ must be nontrivial, and thus P is a (hyperelliptic) Weierstrass point.

From the above argument, it follows that C has 17 hyperelliptic Weierstrass points that are all CM. By Shimura reciprocity law, these CM points are all defined over the Hilbert class field H_K of K, where K is the CM-field defined in Section 5B. Let M be the normal closure of H_K over F. Then $[M : F] = 34$ and $\mathrm{Gal}(M/F) \simeq D_{17}$; and the action of $\mathrm{Gal}(\bar{F}/F)$ on the set of Weierstrass points $\mathscr{W}$ must factor through it (see Section 2C). Therefore, we must have $\#\mathscr{W} = 34$. In other words, C has 34 hyperelliptic Weierstrass points. Since C has genus 16, it must therefore be hyperelliptic by Proposition 2.3. □

Remark 5.10. It follows from the proof of Theorem 5.9 that half of the Weierstrass points on C are CM, while the remaining half are non CM. This means that the hyperelliptic involution must necessarily be exceptional. However, this should be expected since $C = X_0^D(1)/\langle w_D \rangle$, where w_D is the *unique* Atkin–Lehner involution acting on $X_0^D(1)$.

Theorem 5.11. *The automorphism group of the curve C is* $\mathrm{Aut}(C) = \mathbb{Z}/2\mathbb{Z}$.

Proof. By Theorem 5.9, the group $\mathrm{Aut}(C)$ is nontrivial since it contains the hyperelliptic involution. By Lemma 5.7, it injects into an admissible subgroup H of $\mathrm{Aut}(\mathcal{G}_{st})$ of order 2. □

Remark 5.12. By Theorem 5.11, $\mathrm{Aut}(C) = \mathbb{Z}/2\mathbb{Z}$, so that $\mathrm{Aut}(X_0^D(1)) = (\mathbb{Z}/2\mathbb{Z})^s$, with $1 \le s \le 2$. We note that $s = 2$ if and only if the hyperelliptic involution on C comes from an exceptional automorphism on $X_0^D(1)$. We also note that it is conjectured that there are only finitely many Shimura curves X defined over $\mathbb{Q}$ such that $\mathrm{Aut}(X)$ contains an exceptional automorphism (see [Kontogeorgis and Rotger 2008]). This conjecture would imply that there are very few Shimura curve quotients defined over $\mathbb{Q}$ which have automorphisms arising from exceptional automorphisms. However, analogues of this conjecture have barely been explored over totally real fields.

Theorem 5.13. *The curve C is hyperelliptic over* $\mathbb{Q}$.

Proof. Since C descends to $\mathbb{Q}$, it is enough to show that the hyperelliptic involution $\iota : C \to C$ also descends to $\mathbb{Q}$. By Theorem 5.11, $\mathrm{Aut}(C)/\langle \iota \rangle$ is trivial. Furthermore, the field $\mathbb{Q}(\zeta_{32})$ is a splitting field for D whose class number is one. So the CM point attached to the extension $\mathbb{Q}(\zeta_{32})/F$ is defined over F. So C descends to $\mathbb{Q}$ as a hyperelliptic curve by [Sijsling and Voight 2016, Proposition 4.8]. □

Remark 5.14. One should be able to compute an equation for C by using [Voight and Willis 2014]. However, currently, the strategy for doing so is not fully implemented. It should also be possible to use a generalisation of the p-adic approach discussed in [Franc and Masdeu 2014], which was inspired by [Kurihara 1979; 1994]. Given that the determination of equations for Shimura curves defined over totally real fields is one question that is of independent interest in its own right, we hope to return to this in the future.

Remark 5.15. We note that Michon [1981] (and also unpublished work of Ogg) provided a complete list of all hyperelliptic Shimura curves with square-free level defined over $\mathbb{Q}$. Shimura curves defined over $\mathbb{Q}$ which admit hyperelliptic quotients have also been investigated quite a bit; see, for example, [Molina 2012; González and Molina 2016; Guo and Yang 2017]. In contrast, there has been very little work on these types of questions for Shimura curves defined over totally real fields F larger than $\mathbb{Q}$. This makes Theorem 5.9 of the more striking. Indeed, not only does it give one of the few examples of Shimura curves with a hyperelliptic quotient over a totally real field, but also one whose genus is larger than most known examples over $\mathbb{Q}$.

5G. *The Jacobian varieties* $\mathbf{Jac}(X_0^D(1))$ *and* $\mathbf{Jac}(C)$. In this section, we explain the connection between the simple factors of $\mathrm{Jac}(X_0^D(1))$ and the conjectures in [Gross 2016]. There is more in [Cunningham and Dembélé 2017], where this connection is established via lifts of Hilbert modular forms.

Figure 2. The dual graph $\mathcal{G}_{st}$ of the stable model for the quotient $\mathscr{M}/\langle w_{\mathfrak{p}}\rangle$.

From the discussion in Sections 5C and 5D, we have the decomposition for $\mathrm{Jac}(X_0^D(1))$ over F (up to isogeny):

$$\mathrm{Jac}(X_0^D(1)) \sim A_f \times A_{f'} \times A_g \times A_{g'} \times A_h. \tag{1}$$

From (1), and the fact that $w_D = -w$, we see that

$$\mathrm{Jac}(C) \sim A_f \times A_{f'} \times A_g \times A_{g'}. \tag{2}$$

The fourfolds A_f and $A_{f'}$ (resp. A_g and $A_{g'}$) are Galois conjugate. We will see later that one of consequences of the compatibility between the base change action and Hecke orbits is that the decompositions (1) and (2) descend to subfields of F.

Theorem 5.16. *The abelian variety A_h descends to a* 24*-dimensional variety B_h defined over $\mathbb{Q}$, with good reduction outside* 2*, such that* $\mathrm{End}_{\mathbb{Q}}(B_h) \otimes \mathbb{Q} = K_h$ *and*

$$L(B_h, s) = \prod_{\Pi' \in [\Pi_h]} L(\Pi', s),$$

where π_h is the automorphic representation of $\mathrm{GL}_2(\mathbb{A}_F)$ *attached to h, Π_h it lifts to* $\mathrm{GSpin}_{17}(\mathbb{A}_{\mathbb{Q}})$*, and $[\Pi_h]$ the Hecke orbit of Π_h.*

Proof. By Table 1, there exists a generator $\tau \in \mathrm{Gal}(L_h/K_h)$ such that ${}^{\sigma}h = h^{\tau}$. So, by [Cunningham and Dembélé 2017, Theorem 5.4], π_h lifts to an automorphic representation Π_h on a split form of $\mathrm{GSpin}_{17}(\mathbb{A}_{\mathbb{Q}})$, with field of rationality the cubic field K_h. The Hecke orbit $[\Pi_h]$ of Π_h has 3 elements, and by functoriality

$$L(B_h, s) = \prod_{\Pi' \in [\Pi_h]} L(\Pi', s).$$

It follows that $\mathrm{End}_{\mathbb{Q}}(B_h) \otimes \mathbb{Q} = K_h$. Since the level of the form h is the unique prime $\mathfrak{p}$ above 2, B_h has good reduction outside 2. □

Now, we turn to the quotient $C := X_0^D(1)/\langle w_D \rangle$.

Theorem 5.17. *The abelian varieties A_f and $A_{f'}$ (resp. A_g and $A_{g'}$) descend to pairwise conjugate fourfolds B_f and $B_{f'}$ (resp. B_g and $B_{g'}$) over $\mathbb{Q}(\sqrt{2})$, with trivial endomorphism rings, such that*

$$L(B_f, s) = L(\Pi_f, s) \quad \text{and} \quad L(B_{f'}, s) = L(\Pi_{f'}, s),$$
$$L(B_g, s) = L(\Pi_g, s) \quad \text{and} \quad L(B_{g'}, s) = L(\Pi_{g'}, s),$$

where $\pi_f, \pi_{f'}, \pi_g$ and $\pi_{g'}$ are the automorphic representations of $\mathrm{GL}_2(\mathbb{A}_F)$ *attached to f, f', g and g', respectively; and Π_f, $\Pi_{f'}$, Π_g and $\Pi_{g'}$ their respective lifts to* $\mathrm{GSpin}_9/\mathbb{Q}(\sqrt{2})$*. They have good reduction outside* $(\sqrt{2})$.

Proof. The identities in Table 1, combined with [Cunningham and Dembélé 2017, Theorem 5.4], implies that $\pi_f, \pi_{f'}, \pi_g$ and $\pi_{g'}$ indeed lift to automorphic representations $\Pi_f, \Pi_{f'}, \Pi_g$ and $\Pi_{g'}$ on $\mathrm{GSpin}_9/\mathbb{Q}(\sqrt{2})$ with field of rationality $\mathbb{Q}$. Consequently, the fourfolds A_f, $A_{f'}$, A_g and $A_{g'}$ descend to pairwise conjugate fourfolds B_f and $B_{f'}$ (resp. B_g and $B_{g'}$) such that

$$\mathrm{End}_{\mathbb{Q}(\sqrt{2})}(B_f) = \mathrm{End}_{\mathbb{Q}(\sqrt{2})}(B_{f'}) = \mathrm{End}_{\mathbb{Q}(\sqrt{2})}(B_g) = \mathrm{End}_{\mathbb{Q}(\sqrt{2})}(B_{g'}) = \mathbb{Z}.$$

The equalities of L-series follow by functoriality. For the same reason as above, the fourfolds have good reduction outside $(\sqrt{2})$. □

Remark 5.18. The decomposition (1) is only true *a priori* over F. However, Theorem 5.16 and Theorem 5.17 imply that it descends to $\mathbb{Q}(\sqrt{2})$. In fact, the products $A_f \times A_{f'}$ (resp. $A_g \times A_{g'}$) further descend to $\mathbb{Q}$. And so, the decomposition (1) will descend to $\mathbb{Q}$ if we group them accordingly.

5H. ***The connectedness of*** **Spec($\mathbb{T}$).** Let $\mathbb{T}$ be the $\mathbb{Z}$-subalgebra of $\mathrm{End}_{\mathbb{C}}(S_2^D(1))$ acting on $S_2^D(1)$. We recall that $S_2^D(1)$ is isomorphic to $S_2(\mathfrak{p})^{\mathrm{new}}$ as a Hecke module.

Proposition 5.19. Spec($\mathbb{T}$) *is connected.*

Proof. The curve $X_0^D(1)$ is a Shimura curve of prime level, and each Hecke constituent appears with multiplicity one. So, the proof in [Mazur 1977, Proposition 10.6] applies readily. □

The following two propositions determine the congruences which realise the connectedness of Spec($\mathbb{T}$).

Proposition 5.20. *The forms f, f', g and g' are congruent modulo* 5.

Proof. The prime 5 is totally ramified in $L_f = L_{f'}$. Let $\mathfrak{P}_5$ be the unique prime above it, and $\rho_{f,5}, \rho_{f',5} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathcal{O}_{L_f,\mathfrak{P}_5})$ the $\mathfrak{P}_5$-adic representations attached to f and f', respectively. By reduction modulo $\mathfrak{P}_5$, we get two representations $\bar{\rho}_{f,5}, \bar{\rho}_{f',5} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathbb{F}_5)$. From Table 1, we have and ${}^{\sigma}f = f'$, ${}^{\sigma^2}f = f^{\tau}$. Also, since $\mathfrak{P}_5$ is totally ramified in L_f, we have $\tau(\mathfrak{P}_5) = \mathfrak{P}_5$. It follows that $\bar{\rho}_{f,5} = \bar{\rho}_{f',5}$ is a base change from $\mathbb{Q}(\sqrt{2})$. The prime 5 is also totally ramified in $L_g = L_{g'}$. With obvious notations, the same argument as above shows that $\bar{\rho}_{g,5} = \bar{\rho}_{g',5}$ is also a base change from $\mathbb{Q}(\sqrt{2})$.

By using the multiplicity one argument in [Billerey et al. 2018, §6], we show that $\bar{\rho}_{f,5} \simeq \bar{\rho}_{g,5}$. This implies that f, f', g and g' are congruent modulo 5. □

Proposition 5.21. *The forms f, f' and h are congruent modulo* 3.

Proof. There is a unique prime $\mathfrak{P}_3$ above 3 in $L_f = L_{f'}$; it has inertia degree 2 and ramification degree 2. Let $\rho_{f,3}, \rho_{f',3} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathcal{O}_{L_f,\mathfrak{P}_3})$ the $\mathfrak{P}_3$-adic representations attached to f and f', respectively. By reduction modulo $\mathfrak{P}_3$, we get two representations $\bar{\rho}_{f,3}, \bar{\rho}_{f',3} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathbb{F}_9)$. From Table 1, we have and ${}^{\sigma}f = f'$, ${}^{\sigma^2}f = f^{\tau}$. Also, since $\mathfrak{P}_3$ is the unique prime above 3 in L_f, we have $\tau(\mathfrak{P}_3) = \mathfrak{P}_3$. It follows that $\bar{\rho}_{f,3} = \bar{\rho}_{f',3}$ is a base change from $\mathbb{Q}(\zeta_{16})^+$.

In the cubic subfield K_h of L_h, the prime 3 factors as $(3) = \mathfrak{p}_3\mathfrak{p}_3'$, where $\mathfrak{p}_3$ has inertia degree 1, and $\mathfrak{p}_3'$ inertia degree 2. The prime $\mathfrak{p}_3'$ is totally ramified in L_h. We let $\mathfrak{P}_3'$ be the unique prime above it, and $\rho_{h,3} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathcal{O}_{L_h,\mathfrak{P}_3'})$ the $\mathfrak{P}_3'$-adic representation attached to h. By reduction modulo $\mathfrak{P}_3'$, we get a representation $\bar{\rho}_{h,3} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathbb{F}_9)$. From Table 1, we have and ${}^{\sigma}h = h^{\tau}$. Also, since $\mathfrak{P}_3'$ is the unique prime above $\mathfrak{p}_3'$ in L_h, we have $\tau(\mathfrak{P}_3') = \mathfrak{P}_3'$. It follows that $\bar{\rho}_{h,3}$ is also a base change from $\mathbb{Q}(\zeta_{16})^+$.

By using the multiplicity one argument in [Billerey et al. 2018, §6], we show that $\bar{\rho}_{f,3} \simeq \bar{\rho}_{h,3}$. This implies that f, f' and h are congruent modulo 3. □

6. The 2-torsion field of Jac($X_0^D(1)$) and the Harbater field

The main result of this section establishes that every simple factor of $\mathrm{Jac}(X_0^D(1))$ has a 2-torsion field whose normal closure is the Harbater field. We start with the following theorem.

Theorem 6.1. *Let N the field of* 2*-torsion of* Jac(C) *over* $\mathbb{Q}$*. Then N is the Harbater field.*

Proof. Keeping the notation in the proof of Theorem 5.9, N is the normal closure of M. It follows from this, and direct calculations, that $\mathrm{Gal}(N/\mathbb{Q}) \simeq F_{17}$. By construction, N is unramified outside 2 and ∞.

However, by [Harbater 1994, Theorem 2.25], there is a unique Galois number field unramified outside 2 and ∞, with Galois group F_{17}. So N must be the Harbater field. □

Remark 6.2. The field N is the splitting field of the polynomial

$$H := x^{17}-2x^{16}+8x^{13}+16x^{12}-16x^{11}+64x^9-32x^8-80x^7+32x^6+40x^5+80x^4+16x^3-128x^2-2x+68.$$

This polynomial was computed by Noam Elkies following a `mathoverflow.net` discussion [Rouse and Elkies 2014] initiated by Jeremy Rouse. We thank David P. Roberts for bringing this discussion to our attention.

6A. ***The mod 2 Hecke eigensystems.*** Let $\mathbb{T}_f$, $\mathbb{T}_{f'}$ $\mathbb{T}_g$, $\mathbb{T}_{g'}$ and $\mathbb{T}_h$ be the $\mathbb{Z}$-subalgebras acting on the Hecke constituents of f, f', g, g' and h respectively. From the discussion in Section 5C, we have

$$\begin{aligned}\mathbb{T}\otimes\mathbb{Q} &= (\mathbb{T}_f\otimes\mathbb{Q})\times(\mathbb{T}_{f'}\otimes\mathbb{Q})\times(\mathbb{T}_g\otimes\mathbb{Q})\times(\mathbb{T}_{g'}\otimes\mathbb{Q})\times(\mathbb{T}_h\otimes\mathbb{Q})\\ &= L_f\times L_{f'}\times L_g\times L_{g'}\times L_h.\end{aligned}$$

By direct calculations, we get the following:

- $[\mathcal{O}_{L_f} : \mathbb{T}_f] = [\mathcal{O}_{L_{f'}} : \mathbb{T}_{f'}]$ divides 3,
- $[\mathcal{O}_{L_g} : \mathbb{T}_g] = [\mathcal{O}_{L_{g'}} : \mathbb{T}_{g'}] = 1$,
- $[\mathcal{O}_{L_h} : \mathbb{T}_h]$ divides $3\cdot 5^6$.

Therefore $\mathbb{T}\otimes\mathbb{Z}_2$ decomposes into $\mathbb{Z}_2$-algebras as

$$\mathbb{T}\otimes\mathbb{Z}_2 = (\mathbb{T}_f\otimes\mathbb{Z}_2)\times(\mathbb{T}_{f'}\otimes\mathbb{Z}_2)\times(\mathbb{T}_g\otimes\mathbb{Z}_2)\times(\mathbb{T}_{g'}\otimes\mathbb{Z}_2)\times(\mathbb{T}_h\otimes\mathbb{Z}_2).$$

The prime 2 is inert in $L_f = L_{f'}$, and $L_g = L_{g'}$, so the first four factors are local $\mathbb{Z}_2$-algebras. Let $\mathfrak{m}_f$, $\mathfrak{m}_{f'}$, $\mathfrak{m}_g$ and $\mathfrak{m}_{g'}$ be the corresponding maximal ideals. Then, by the identities in Table 1, we have $\sigma(\mathfrak{m}_f) = \mathfrak{m}_{f'}$ and $\sigma^2(\mathfrak{m}_f) = \tau_f(\mathfrak{m}_f)$ for some $\tau_f \in \mathrm{Gal}(\mathbb{F}_{16}/\mathbb{F}_2)$; and $\sigma(\mathfrak{m}_g) = \mathfrak{m}_{g'}$ and $\sigma^2(\mathfrak{m}_g) = \tau_g(\mathfrak{m}_g)$ for some $\tau_g \in \mathrm{Gal}(\mathbb{F}_{16}/\mathbb{F}_2)$. We let $\theta_f, \theta_{f'}, \theta_g, \theta_{g'} : \mathbb{T}\otimes\mathbb{Z}_2 \to \mathbb{F}_{16}$ be the corresponding mod 2 Hecke eigensystems.

Next, we recall that L_h is the ray class field of conductor $\mathfrak{c} = \left(\frac{1}{2}(c^2 - 16c + 25)\right)$ over the field $K_h = \mathbb{Q}(c)$, with $c^3 + c^2 - 229c + 167 = 0$. The prime 2 is totally ramified in K_h. Letting $\mathfrak{p}_2$ be the unique prime above it, we get that $\mathfrak{p}_2 = \mathfrak{P}\mathfrak{P}'$, where $\mathfrak{P}$ and $\mathfrak{P}'$ are inert primes, and $\tau(\mathfrak{P}) = \mathfrak{P}'$. Therefore, there are two maximal ideals $\mathfrak{m}_h$ and $\mathfrak{m}'_h$ in $\mathbb{T}_h\otimes\mathbb{Z}_2$ such that $\sigma(\mathfrak{m}_h) = \mathfrak{m}'_h$ and $\sigma^2(\mathfrak{m}_h) = \tau_h(\mathfrak{m}_h)$. We let $\theta_h, \theta'_h : \mathbb{T}\otimes\mathbb{Z}_2 \to \mathbb{F}_{16}$ be the resulting two mod 2 Hecke eigensystems.

Proposition 6.3. *The forms f, f', g, g' and h give rise to two mod 2 Hecke eigensystems θ and θ' that $\theta' = \theta\circ\sigma$ and $\theta\circ\sigma^2 = \bar{\tau}\circ\theta$, where $\mathrm{Gal}(\mathbb{F}_{16}/\mathbb{F}_2) = \langle\bar{\tau}\rangle$. Up to relabelling, we have $\theta = \theta_f = \theta_g = \theta_h$, and $\theta' = \theta_{f'} = \theta_{g'} = \theta'_h$.*

Proof. We will apply the multiplicity one argument in [Billerey et al. 2018, §6] to deduce that, up to relabelling, $\theta_f = \theta_g = \theta_h$, and $\theta_{f'} = \theta_{g'} = \theta'_h$. Let M be the underlying $\mathbb{F}_2$-module to $\mathbb{T}\otimes\mathbb{F}_2$. Then, the pair $(\theta_f, \theta_{f'})$ comes from two simple Hecke constituents of dimension 4 over $\mathbb{F}_2$ that are conjugate by the

action of $\mathrm{Gal}(F/\mathbb{Q})$. These Hecke constituents belong to the socle S of M, i.e., the largest semisimple $\mathbb{T}\otimes\mathbb{F}_2$-submodule of M. Likewise for the pairs $(\theta_g,\theta_{g'})$ and (θ_h,θ_h'). Let $\mathbb{T}'$ be the $\mathbb{Z}$-subalgebra of $\mathbb{T}$ generated by the Hecke operators $T_{\mathfrak{p}}$, with $\mathrm{N}\,\mathfrak{p}\le 1000$. We view M as a $\mathbb{T}'\otimes\mathbb{F}_2$-module, and let S' be its socle. By direct calculations in `Magma`, we show that S' has two irreducible constituents, and each constituent has dimension 4 and multiplicity one. Furthermore, each of those constituents decomposes into 4 one-dimensional Hecke constituents over $\mathbb{T}'\otimes\mathbb{F}_{16}$. This means that S' must necessarily be the socle of M viewed as a $\mathbb{T}\otimes\mathbb{F}_2$-module, and that its $\mathbb{T}'\otimes\mathbb{F}_{16}$-decomposition is also the $\mathbb{T}\otimes\mathbb{F}_{16}$-decomposition of S. By comparing these one-dimensional $\mathbb{F}_{16}$-valued Hecke eigensystems with the reduction modulo 2 of the Hecke eigenvalues of the newforms in $S_2^D(1)$, we see that $\theta_f=\theta_g=\theta_h$, and $\theta_{f'}=\theta_{g'}=\theta_h'$, up to relabelling. The identities $\theta'=\theta\circ\sigma$ and $\theta\circ\sigma^2=\bar{\tau}\circ\theta$ follow from the relations between the forms. □

6B. *The fields of* 2*-torsion of the simple factors of* $\mathbf{Jac}(X_0^D(1))$.

Theorem 6.4. *Let A be a simple factor of* $\mathrm{Jac}(X_0^D(1))$, *and $M=F(A[2])$ the field of* 2*-torsion of A. Then, the normal closure of M is the Harbater field N.*

We will give two proofs of this result, starting with the simplest one.

First proof of Theorem 6.4. In light of Proposition 6.3, it is enough to prove this for the simple factors of $\mathrm{Jac}(C)$. To this end, recall that

$$\mathrm{Jac}(C)\sim A_f\times A_{f'}\times A_g\times A_{g'}.$$

By Theorem 6.1, we know that the field of 2-torsion of $\mathrm{Jac}(C)$ is the Harbater field. So it is enough to show that the compositum of the fields of 2-torsion of its simple factors is also the Harbater field. But, again by Proposition 6.3, A_f and A_g have the same field of 2-torsion. It is the field M_θ cut out by the Galois representation attached to the Hecke eigensystem θ. Similarly, $A_{f'}$ and $A_{g'}$ have the same field of 2-torsion, the field $M_{\theta'}$ cut out by the Galois representation attached to θ'. Since θ and θ' are interchanged by $\mathrm{Gal}(F/\mathbb{Q})$, we must have $M_{\theta'}=M_\theta^\sigma$. Therefore M_θ and $M_{\theta'}$ have the same normal closure $N_\theta=N_{\theta'}$. By replacing the isogeny $\phi_f:\mathrm{Jac}(C)\to A_f$ if necessary, we can assume that M_θ, and hence N_θ, is a subfield of N. From the Frobenius data attached to θ, we see that the order of $\mathrm{Gal}(N_\theta/\mathbb{Q})$ is divisible by 17, hence $[N:N_\theta]\mid 16$. Since $\mathrm{Gal}(N/\mathbb{Q})\simeq F_{17}$ has no nontrivial normal subgroup whose order divides 16, we conclude that $N_\theta=N$. □

For the second proof, we need the following result.

Proposition 6.5. *Let θ and θ' be the Hecke eigensystems in Proposition 6.3, and let $\bar{\rho},\bar{\rho}':\mathrm{Gal}(\overline{\mathbb{Q}}/F)\to\mathrm{GL}_2(\mathbb{F}_{16})$ the mod* 2 *Galois representations attached to them. Then, there are characters $\chi,\chi':\mathrm{Gal}(\overline{\mathbb{Q}}/K)\to\mathbb{F}_{2^8}^\times$, with trivial conductor such that $\bar{\rho}=\mathrm{Ind}_K^F\,\chi$, and $\bar{\rho}'=\mathrm{Ind}_K^F\,\chi'$.*

Proof. We already computed the Hecke constituents of the space $S_2(1)$ in [Dembélé 2009]. The mod 2 Hecke eigensystems in that case have coefficient fields $\mathbb{F}_{2^s}$ where $s=1,2,8$. Therefore, since θ has coefficient field $\mathbb{F}_{16}$, it cannot arise from an eigenform of level 1. By the Serre conjecture for totally real fields (the totally ramified case) [Gee and Savitt 2011], it must appear on the quaternion algebra D' with

level (1) and nontrivial weight. The same is true for θ'. In fact, the analysis conducted in the proof of Proposition 6.3 also shows that they are the only eigensystems that can appear at that weight. (We note that there are only two Serre weights in this case.)

Let $\chi : \mathrm{Gal}(\overline{\mathbb{Q}}/K) \to \overline{\mathbb{F}}_2^\times$ be a character with trivial conductor such that $\chi^s \neq \chi$, where $\mathrm{Gal}(K/F) = \langle s \rangle$. By class field theory, we can identity χ with its image under the Artin map. Since χ is unramified, it must factor as $\chi : K^\times \backslash \mathbb{A}_K^\times \twoheadrightarrow \mathrm{Cl}_K \to \overline{\mathbb{F}}_2^\times$. Furthermore, since $\mathrm{Cl}_K \simeq \mathbb{Z}/17\mathbb{Z}$, we must have $\chi : K^\times \backslash \mathbb{A}_K^\times \to \mathbb{F}_{2^8}^\times$, and the representation $\bar{\rho}_\chi := \mathrm{Ind}_K^F \chi : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathbb{F}_{16})$ has coefficients in $\mathbb{F}_{16}$. So, $\bar{\rho}_\chi$ has level (1) and nontrivial weight by the argument above. Therefore, it must be isomorphic to a Galois conjugate of $\bar{\rho}$. Up to relabelling, we can assume that $\bar{\rho} \simeq \bar{\rho}_\chi$. Since θ and θ' are $\mathrm{Gal}(F/\mathbb{Q})$-conjugate, there is also a character $\chi' : K^\times \backslash \mathbb{A}_K^\times \to \mathbb{F}_{2^8}^\times$ such that $\bar{\rho}' \simeq \bar{\rho}_{\chi'}$.

Alternatively, we can show that θ appears on D' with the nontrivial weight without using the fact that it has coefficients in $\mathbb{F}_{16}$. Indeed, we have

$$\bar{\rho}_\chi|_{I_\mathfrak{p}} \simeq \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix}.$$

Let $K_\mathfrak{P}$ be the completion of K at $\mathfrak{P}$, the unique prime above $\mathfrak{p}$. Since $K = F[\beta]$, and $\beta^2 = -2 - \alpha$ is a generator of $\mathfrak{p}$, then we have $K_\mathfrak{p} = F_\mathfrak{p}[\sqrt{\varpi}]$, where ϖ is a uniformiser of $F_\mathfrak{p}$. Therefore, $\bar{\rho}_\chi|_{D_\mathfrak{p}}$ doesn't arise from a finite flat group scheme. Hence, $\bar{\rho}_\chi$ must have nontrivial weight. □

We are now ready for the second proof of Theorem 6.4.

Second proof of Theorem 6.4. Let $\bar{\rho}_\theta, \bar{\rho}_{\theta'} : \mathrm{Gal}(\overline{\mathbb{Q}}/F) \to \mathrm{GL}_2(\mathbb{F}_{16})$ be the mod 2 Galois representations attached to the eigensystems θ and θ'. By Proposition 6.5, $\bar{\rho}_\theta$ and $\bar{\rho}_{\theta'}$ are dihedral and we have that $\mathrm{im}(\bar{\rho}_\theta) = \mathrm{im}(\bar{\rho}_{\theta'}) = D_{17}$. Let M_θ, $M_{\theta'}$ be the fields cut out by $\bar{\rho}_\theta$ and $\bar{\rho}_{\theta'}$; and N_θ and $N_{\theta'}$ the normal closure of M_θ and $M_{\theta'}$, respectively. By Proposition 6.3, we have $M_{\theta'} = M_\theta^\sigma$, hence $N_\theta = N_{\theta'}$. Also, by construction M_θ and M_θ^σ are unramified extension of K. So, by uniqueness of the Hilbert class field, we must have $M_\theta = M_\theta^\sigma = M_\theta M_\theta^\sigma = H_K$, where $M_\theta M_\theta^\sigma$ is the compositum of M_θ and M_θ^σ; and H_K is the Hilbert class field of K. Since $\theta \circ \sigma^2 = \bar{\tau} \circ \theta$, we have

$$\mathrm{Gal}(N_\theta/\mathbb{Q}) = D_{17} \rtimes \mathbb{Z}/8\mathbb{Z} = F_{17}.$$

Again by [Harbater 1994, Theorem 2.25], we must have $N = N_\theta = N_{\theta'}$. □

Remark 6.6. From Theorem 6.4, we see that none of the fourfolds A_f, $A_{f'}$, A_g or $A_{g'}$ can be the Jacobian of a hyperelliptic curve since the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/F)$ on the points of 2-torsion cannot factor through S_{10} (see Section 2C). However, as we explained earlier, A_f, $A_{f'}$, A_g and $A_{g'}$ descend, separately, into pairwise conjugate abelian varieties over $\mathbb{Q}(\sqrt{2})$. And the products $A_f \times A_{f'}$ and $A_g \times A_{g'}$ are 8-dimensional abelian varieties which further descend to $\mathbb{Q}$. So, we conclude with the following questions. Do there exist hyperelliptic curves C_f and C_g defined over F such that

$$\mathrm{Jac}(C_f) \sim A_f \times A_{f'} \quad \text{and} \quad \mathrm{Jac}(C_g) \sim A_g \times A_{g'}?$$

If so, do these two curves descend to $\mathbb{Q}$ as well? We were asked these two questions by Noam Elkies in an email. An affirmative answer to them would mean that the Harbater field is given by hyperelliptic curves of genus 8, which is much smaller. A priori, the hyperelliptic polynomials of these curves should have degree 18. However, the structure of the Galois group $\mathrm{Gal}(N/\mathbb{Q}) = F_{17}$ indicates that one of their roots would be rational and could be moved to ∞. This means that the hyperelliptic polynomials of the curves C_f and C_g would in fact have degree 17, the same as that of the Elkies polynomial displayed earlier.

Acknowledgements

I would like to thank Frank Calegari for several helpful email exchanges; Vladimir Dokchitser and Céline Maistret for some useful discussion; and Jeroen Sijsling for carefully reading an earlier draft of this work. I would also like to give a special thanks to John Voight as this note owes a lot to the lengthy discussions I had with him on this topic. I learned of the discussion about the Harbater field between Jeremy Rouse and Noam D. Elkies from David P. Roberts who pointed us to the `mathoverflow.net` post related to this. So, I would like to thank him for this. During the course of this project, I stayed at the following institutions: Dartmouth College, King's College London, the Max Planck Institute for Mathematics in Bonn, and the University of Barcelona; I would like to thank them for their generous hospitality. I also thank the referees for many helpful suggestions. Finally, as alluded to earlier, this note originated with a question of Benedict Gross. So I would like to thank him for this, and for his constant encouragement.

References

[Billerey et al. 2018] N. Billerey, I. Chen, L. Dembélé, L. Dieulefait, and N. Freitas, "Some extensions of the modular method and Fermat equations of signature $(13, 13, n)$", preprint, 2018. arXiv

[Borel 1981] A. Borel, "Commensurability classes and volumes of hyperbolic 3-manifolds", *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **8**:1 (1981), 1–33. MR Zbl

[Boutot and Carayol 1991] J.-F. Boutot and H. Carayol, "Uniformisation p-adique des courbes de Shimura: les théorèmes de Čerednik et de Drinfel'd", pp. 45–158 in *Courbes modulaires et courbes de Shimura* (Orsay, France, 1987/1988), Astérisque **196-197**, Soc. Math. France, Paris, 1991. MR Zbl

[Boutot and Zink 1995] J.-F. Boutot and T. Zink, "The p-adic uniformization of Shimura curves", preprint, 1995.

[Carayol 1986] H. Carayol, "Sur la mauvaise réduction des courbes de Shimura", *Compos. Math.* **59**:2 (1986), 151–230. MR Zbl

[Chinburg and Friedman 1999] T. Chinburg and E. Friedman, "An embedding theorem for quaternion algebras", *J. Lond. Math. Soc.* (2) **60**:1 (1999), 33–44. MR Zbl

[Cunningham and Dembélé 2017] C. Cunningham and L. Dembélé, "Lifts of Hilbert modular forms and application to modularity of abelian varieties", preprint, 2017. arXiv

[Deligne 1971] P. Deligne, "Travaux de Shimura", exposé 389, pp. 123–165 in *Séminaire Bourbaki,* 1970/1971, Lecture Notes in Math. **244**, Springer, 1971. 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

[Dembélé 2009] L. Dembélé, "A non-solvable Galois extension of $\mathbb{Q}$ ramified at 2 only", *C. R. Math. Acad. Sci. Paris* **347**:3-4 (2009), 111–116. MR Zbl

[Dembélé and Donnelly 2008] L. Dembélé and S. Donnelly, "Computing Hilbert modular forms over fields with nontrivial class group", pp. 371–386 in *Algorithmic number theory*, edited by A. J. van der Poorten and A. Stein, Lecture Notes in Comput. Sci. **5011**, Springer, 2008. MR Zbl

[Dembélé and Voight 2013] L. Dembélé and J. Voight, "Explicit methods for Hilbert modular forms", pp. 135–198 in *Elliptic curves, Hilbert modular forms and Galois deformations*, edited by H. Darmon et al., Birkhäuser, Basel, 2013. MR Zbl

[Doi and Naganuma 1967] K. Doi and H. Naganuma, "On the algebraic curves uniformized by arithmetical automorphic functions", *Ann. of Math.* (2) **86** (1967), 449–460. MR Zbl

[Farkas and Kra 1980] H. M. Farkas and I. Kra, *Riemann surfaces*, Grad. Texts in Math. **71**, Springer, 1980. MR Zbl

[Franc and Masdeu 2014] C. Franc and M. Masdeu, "Computing fundamental domains for the Bruhat–Tits tree for $\mathrm{GL}_2(\mathbb{Q}_p)$, p-adic automorphic forms, and the canonical embedding of Shimura curves", *LMS J. Comput. Math.* **17**:1 (2014), 1–23. MR Zbl

[Gee and Savitt 2011] T. Gee and D. Savitt, "Serre weights for mod p Hilbert modular forms: the totally ramified case", *J. Reine Angew. Math.* **660** (2011), 1–26. MR Zbl

[González and Molina 2016] J. González and S. Molina, "The kernel of Ribet's isogeny for genus three Shimura curves", *J. Math. Soc. Japan* **68**:2 (2016), 609–635. MR Zbl

[Greenberg and Voight 2011] M. Greenberg and J. Voight, "Computing systems of Hecke eigenvalues associated to Hilbert modular forms", *Math. Comp.* **80**:274 (2011), 1071–1092. MR Zbl

[Gross 2016] B. H. Gross, "On the Langlands correspondence for symplectic motives", *Izv. Ross. Akad. Nauk Ser. Mat.* **80**:4 (2016), 49–64. In Russian; translated in *Izv. Math.* **80**:4 (2016), 678–692. MR Zbl

[Guo and Yang 2017] J.-W. Guo and Y. Yang, "Equations of hyperelliptic Shimura curves", *Compos. Math.* **153**:1 (2017), 1–40. MR Zbl

[Harbater 1994] D. Harbater, "Galois groups with prescribed ramification", pp. 35–60 in *Arithmetic geometry* (Tempe, AZ, 1993), edited by N. Childress and J. W. Jones, Contemp. Math. **174**, Amer. Math. Soc., Providence, RI, 1994. MR Zbl

[Hindry and Silverman 2000] M. Hindry and J. H. Silverman, *Diophantine geometry: an introduction*, Grad. Texts in Math. **201**, Springer, 2000. MR Zbl

[Jones 2010] J. W. Jones, "Number fields unramified away from 2", *J. Number Theory* **130**:6 (2010), 1282–1291. MR Zbl

[Kontogeorgis and Rotger 2008] A. Kontogeorgis and V. Rotger, "On the non-existence of exceptional automorphisms on Shimura curves", *Bull. Lond. Math. Soc.* **40**:3 (2008), 363–374. MR Zbl

[Kurihara 1979] A. Kurihara, "On some examples of equations defining Shimura curves and the Mumford uniformization", *J. Fac. Sci. Univ. Tokyo Sect. IA Math.* **25**:3 (1979), 277–300. MR Zbl

[Kurihara 1994] A. Kurihara, "On p-adic Poincaré series and Shimura curves", *Int. J. Math.* **5**:5 (1994), 747–763. MR Zbl

[Maclachlan 2006] C. Maclachlan, "Torsion in arithmetic Fuchsian groups", *J. Lond. Math. Soc.* (2) **73**:1 (2006), 14–30. MR Zbl

[Maclachlan 2009] C. Maclachlan, "Existence and non-existence of torsion in maximal arithmetic Fuchsian groups", *Groups Complex. Cryptol.* **1**:2 (2009), 287–295. MR Zbl

[Magma 1997] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system, I: The user language", *J. Symbolic Comput.* **24**:3-4 (1997), 235–265. MR Zbl

[Mazur 1977] B. Mazur, "Modular curves and the Eisenstein ideal", *Inst. Hautes Études Sci. Publ. Math.* **47** (1977), 33–186. MR Zbl

[Michon 1981] J.-F. Michon, "Courbes de Shimura hyperelliptiques", *Bull. Soc. Math. France* **109**:2 (1981), 217–225. MR Zbl

[Molina 2012] S. Molina, "Equations of hyperelliptic Shimura curves", *Proc. Lond. Math. Soc.* (3) **105**:5 (2012), 891–920. MR Zbl

[Nekovář 2012] J. Nekovář, "Level raising and anticyclotomic Selmer groups for Hilbert modular forms of weight two", *Canad. J. Math.* **64**:3 (2012), 588–668. MR Zbl

[Rouse and Elkies 2014] J. Rouse and N. Elkies, "Degree 17 number fields ramified only at 2", MathOverflow thread, 2014, available at https://mathoverflow.net/q/172148.

[Sage 2019] W. A. Stein et al., "Sage mathematics software", 2019, available at http://www.sagemath.org. Version 8.7.

[Shimura 1970] G. Shimura, "On canonical models of arithmetic quotients of bounded symmetric domains", *Ann. of Math.* (2) **91** (1970), 144–222. MR Zbl

[Sijsling 2013] J. Sijsling, "Canonical models of arithmetic (1; e)-curves", *Math. Z.* **273**:1-2 (2013), 173–210. MR Zbl

[Sijsling and Voight 2016] J. Sijsling and J. Voight, "On explicit descent of marked curves and maps", *Res. Number Theory* **2** (2016), art. id. 27. MR Zbl

[Tian and Xiao 2016] Y. Tian and L. Xiao, "On Goren–Oort stratification for quaternionic Shimura varieties", *Compos. Math.* **152**:10 (2016), 2134–2220. MR Zbl

[Vignéras 1980] M.-F. Vignéras, *Arithmétique des algèbres de quaternions*, Lecture Notes in Math. **800**, Springer, 1980. MR Zbl

[Voight 2005] J. M. Voight, *Quadratic forms and quaternion algebras: algorithms and arithmetic*, Ph.D. thesis, University of California, Berkeley, 2005, available at https://search.proquest.com/docview/305032530.

[Voight 2009] J. Voight, "Computing fundamental domains for Fuchsian groups", *J. Théor. Nombres Bordeaux* **21**:2 (2009), 469–491. MR Zbl

[Voight 2018] J. Voight, "Quaternion algebras", preprint, 2018, available at https://tinyurl.com/voightquat.

[Voight and Willis 2014] J. Voight and J. Willis, "Computing power series expansions of modular forms", pp. 331–361 in *Computations with modular forms* (Heidelberg, 2011), edited by G. Böckle and G. Wiese, Contrib. Math. Comput. Sci. **6**, Springer, 2014. MR Zbl

Communicated by Bjorn Poonen
Received 2019-07-24 Revised 2020-02-27 Accepted 2020-03-28

lassina.dembele@gmail.com *Department of Mathematics, University of Luxembourg, Luxembourg*

msp

ALGEBRA AND NUMBER THEORY 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2743

Generating series of a new class of orthogonal Shimura varieties

Eugenia Rosu and Dylan Yott

For a new class of Shimura varieties of orthogonal type over a totally real number field, we construct special cycles and show the modularity of Kudla's generating series in the cohomology group.

1. Introduction

For Hilbert modular surfaces, Hirzebruch and Zagier [1976] showed that certain generating series that have as coefficients the Hirzebruch–Zagier divisors are modular forms of weight 1. Further inspired by this work, Gross, Kohnen and Zagier [Gross et al. 1987] showed that a generating series that has Heegner divisors as coefficients is modular of weight $\frac{3}{2}$. This approach is unified by Borcherds [1999], who showed more generally the modularity of generating series with Heegner divisor classes as coefficients in the Picard group over $\mathbb{Q}$.

Kudla and Millson extended the results to Shimura varieties of orthogonal type over a totally real number field and showed the modularity in the cohomology group in [Kudla 1997a], based on work from [Kudla and Millson 1986; 1987; 1990]. This is further extended by Yuan, Zhang and Zhang [Yuan et al. 2009], who showed the modularity of the generating series in the Chow group.

In the current paper, inspired by the above work of Kudla and Millson, we construct special cycles on a different Shimura variety of orthogonal type over a totally real number field F and show the modularity of Kudla's generating series in the cohomology group.

We consider the Shimura variety corresponding to the reductive group $\mathrm{Res}_{F/\mathbb{Q}} G$, where $G = \mathrm{GSpin}(V)$ is the GSpin group for V a quadratic space over a totally real number field F, $[F : \mathbb{Q}] = d$. We choose V of signature $(n, 2)$ at e real places and signature $(n+2, 0)$ at the remaining $d-e$ places. Kudla, Millson and Yuan, Zhang, Zhang have treated the case of $e = 1$, while we allow $e \in \{1, \dots, d\}$.

If $e > 1$, there is no simpler divisor case, which makes the analysis much harder. In particular, there is a very technical convergence issue that does not appear in the work of Kudla and Millson.

We present now the setting of the paper. For F be a totally real field with real embeddings $\sigma_1, \dots \sigma_d$, let $\mathbb{A} = \mathbb{A}_F$ be the ring of adeles of F and let V be a quadratic space over F of signature $(n, 2)$ at the infinite

MSC2010: primary 11G18; secondary 19E15.

Keywords: Shimura varieties, special cycles, modular, generating series, automorphic in cohomology group, Green functions, Kudla–Millson form, theta functions.

places $\sigma_1, \ldots, \sigma_e$ and of signature $(n+2, 0)$ elsewhere. Let G denote the reductive group $\mathrm{GSpin}(V)$ over F. We define the hermitian symmetric domain D corresponding to G to be

$$D = D_1 \times D_2 \times \ldots \times D_e,$$

where D_i is the Hermitian symmetric domain of oriented negative definite 2-planes in $V_{\sigma_i} = V \otimes_{\sigma_i} \mathbb{R}$.

Then $(\mathrm{Res}_{F/\mathbb{Q}} G, D)$ is a Shimura datum and for any open compact subgroup K of $G(\mathbb{A}_f)$, this gives us the complex Shimura variety

$$M_K(\mathbb{C}) \simeq G(F)\backslash D \times G(\mathbb{A}_f)/K.$$

For $i = 1, \ldots, e$ we let L_{D_i} be the complex line bundle corresponding to the points of D_i. We also define the projections maps $p_i : D \to D_i$ and then the line bundles $p_i^* L_{D_i} \in \mathrm{Pic}(D)$ descend to line bundles $L_{K,i} \in \mathrm{Pic}(M_K) \otimes \mathbb{Q}$.

Let W be a totally positive subspace of V, meaning that $W_{\sigma_i} = W \otimes_{\sigma_i} \mathbb{R}$ is a positive subspace of $V_{\sigma_i} = V \otimes_{\sigma_i} \mathbb{R}$ for all places $1 \le i \le d$. We define $V_W = W^\perp$ to be the space of vectors in V that are orthogonal to W, $G_W = \mathrm{GSpin}(V_W)$ and $D_W = D_{W,1} \times \cdots \times D_{W,e}$ the Hermitian symmetric domain associated to G_W, where $D_{W,i}$ consists of the lines in D_i perpendicular to W. We actually have the natural identifications

$$G_W = \{g \in G : gw = w, \forall w \in V_W\}, \quad D_W = \{(\tau_1, \ldots, \tau_e) \in D : \langle w, \tau_i \rangle = 0,\ \forall w \in W, \forall 1 \le i \le e\},$$

where $\langle \cdot, \cdot \rangle$ is the inner product corresponding to q_i, the quadratic form on V_{σ_i}, that extends to $V_{\sigma_i}(\mathbb{C})$ by $\mathbb{C}$-linearity.

Then $(\mathrm{Res}_{F/\mathbb{Q}} G_W, D_W)$ is a Shimura datum and we have a morphism

$$(\mathrm{Res}_{F/\mathbb{Q}} G_W, D_W) \to (\mathrm{Res}_{F/\mathbb{Q}} G, D)$$

of Shimura data. For $K \subset G(\mathbb{A}_f)$ an open compact subgroup and $g \in G(\mathbb{A}_f)$, we can define the complex Shimura variety,

$$M_{gKg^{-1},W} = G_W(F)\backslash D_W \times G_W(\mathbb{A}_f)/(gKg^{-1} \cap G_W(\mathbb{A}_f)).$$

Moreover, we have an injection of $M_{gKg^{-1},W}$ into M_K given by

$$M_{gKg^{-1},W} \to M_K, \quad [\tau, h] \to [\tau, hg].$$

We define the cycle $Z(W, g)_K$ to be the image of the morphism above. Note that $Z(W, g)_K$ is represented by the subset $D_W \times G_W(\mathbb{A}_f)gK$ of $D \times G(\mathbb{A}_f)$.

Now let $x = (x_1, \ldots, x_r) \in V(F)^r$ and let $U(x) := \mathrm{Span}_F\{x_1, \ldots, x_r\}$ be a subspace of V. Then we define *Kudla's special cycles*:

$$Z(x, g)_K = \begin{cases} Z(U(x), g)_K((-1)^e c_1(L_{K,1}^\vee) \cdots c_1(L_{K,e}^\vee))^{r - \dim U} & \text{if } U(x) \text{ is totally positive,} \\ 0 & \text{otherwise.} \end{cases}$$

Here c_1 denotes the Chern class of a line bundle. We will also use the notation $G_x := G_{U(x)}$, $D_x := D_{U(x)}$, $V_x := V_{U(x)}$. Note that if $x = (x_1, \dots, x_r) \in V(F)^r$, we have $G_x = G_{x_1} \cap \cdots \cap G_{x_r}$, as well as $D_x = D_{x_1} \cap \cdots \cap D_{x_r}$.

Now we will define *Kudla's generating function*. For any Schwartz–Bruhat functions $\phi_f \in S(V^r(\mathbb{A}_f))^K$ and g' in $\widetilde{\mathrm{Sp}}_{2r}(\mathbb{A})$, where $\widetilde{\mathrm{Sp}}_{2r}(\mathbb{A})$ is the metaplectic cover of the symplectic group $\mathrm{Sp}_{2r}(\mathbb{A})$, we define the generating series

$$Z(g', \phi_f) = \sum_{x \in G(F)\backslash V^r} \sum_{g \in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K} r(g'_f)\phi_f(g^{-1}x) W_{T(x)}(g'_\infty) Z(x, g)_K.$$

Here r is the Weil representation of $\widetilde{\mathrm{Sp}}_{2r}(\mathbb{A}) \times O(V^r_{\mathbb{A}})$, where $T(x) = \frac{1}{2}(\langle x_i, x_j \rangle)_{1 \le i,j \le r} \in M_r(F)$ is the intersection matrix of x, and $W_{T(x)}$ is the standard Whittaker function for $T(x)$. Note that when $e = 1$, for $g_f = \mathrm{Id}$ and a careful choice of g'_∞ we recover the generating series presented in [Yuan et al. 2009].

The following is the main theorem of the paper:

Theorem 1.1. *Let $\phi_f \in S(V^r(\mathbb{A}_f))^K$ be any Schwartz–Bruhat function invariant under K. Then the series $[Z(g', \phi_f)]$ is an automorphic form, discrete of parallel weight $1 + \frac{n}{2}$ for $g' \in \widetilde{\mathrm{Sp}}_{2r}(\mathbb{A})$ and valued in $H^{2er}(M_K, \mathbb{C})$.*

By modularity here we mean that, for any linear function $l : H^{2er}(M_K, \mathbb{C}) \to \mathbb{C}$, the generating series obtained by acting via l on the cohomology classes of the special cycles

$$l(Z(g', \phi_f)) = \sum_{x \in G(F)\backslash V^r} \sum_{g \in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K} r(g'_f)\phi_f(g^{-1}x) W_{T(x)}(g'_\infty) l(Z(x, g)_K).$$

is absolutely convergent and an automorphic form with coefficients in $\mathbb{C}$ in the usual sense.

The case $e = 1$ was proved by Kudla and Millson in [Kudla 1997a], based on [Kudla and Millson 1986; 1987; 1990]. Yuan, Zhang and Zhang [Yuan et al. 2009] proved further the modularity of $Z(g', \phi_f)$ in the Chow group. One can further conjecture that for $e > 1$ the series $Z(g', \phi_f)$ is an automorphic form, discrete of weight $1 + \frac{n}{2}$ for $g' \in \widetilde{\mathrm{Sp}}_{2r}(\mathbb{A})$ valued in $\mathrm{CH}^{er}(M_K)_{\mathbb{C}}$. This is out of reach at the moment, but one can expect to extend the methods of Borcherds [1999] to show the modularity in the Chow group.

We will present now the ideas of the proof. We prove the cases $e > 1$ by extending the ideas of Kudla and Millson. For each cycle $Z(x, g)$ we want to construct a Green current $\eta(x, g)$ of $Z(x, g)$ in $M_K(\mathbb{C})$. Via the isomorphism $H^{2er}_{dR}(X_K, \mathbb{C}) \simeq H^{BM}_{2er(n-1)}(X_K, \mathbb{C})$, where the former is de Rham cohomology while the latter is Borel–Moore cohomology, we have the identification of cohomology classes

$$[Z(x, g)] = [\omega(\eta(x, g))],$$

where $\omega(\eta(x, g))$ is the Chern form corresponding to the Green current $\eta(x, g)$.

Let $x \in V(F)^r$ such that $U(x) := \mathrm{Span}_F\{x_1, \dots, x_r\}$ is a totally positive definite k-subspace of V. Define

$$x' := (x'_1, \dots, x'_k)$$

such that $x'_1 = x_{i_1}, \dots, x'_k = x_{i_k}$ with $1 \le i_1 < \cdots < i_k \le r$ the smallest indices for which $U(x') = U(x)$.

We take the currents defined by Kudla and Millson $\eta_0(x_j', \tau_i)$ of $D_{x_j,i}$ in D_i, where $1 \le j \le k,\ 1 \le i \le e$. Taking further the $*$-product of the currents $\eta_0(x_j', \tau_i)$ for $1 \le i \le e$, we get a Green current of $D_{x,i}$ in D_i:

$$\eta_1(x', \tau_i) = \eta_0(x_1', \tau_i) * \eta_0(x_2', \tau_i) * \cdots * \eta_0(x_k', \tau_i).$$

Taking the pullbacks via the projections $p_i : D \to D_i$ and taking the $*$-product, we obtain a Green current of D_x in D:

$$\eta_2(x', g) = p_1^*\eta_1(x', \tau_1) * p_2^*\eta_1(x', \tau_2) * \cdots * p_e^*\eta_1(x', \tau_e).$$

Furthermore, we average the current $\eta_2(x', g)$ on a lattice to get

$$\eta_3(x', \tau; g, h) = \sum_{\gamma \in G_x(F)\backslash G(F)} \eta_2(x', \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h),$$

which is a Green current for $G(F)(D_x \times G_x(\mathbb{A}_f) g K/K)$ in $D \times G(\mathbb{A}_f)/K$. Showing the convergence of the sum in the definition $\eta_3(x', \tau; g, h)$ represents the most technical part of the proof and it is treated in Section 3G.

As $\eta_3(x', \tau; g, h)$ is invariant under the left action of $G(F)$, $\eta_3(x', \tau; g, h)$ descends to a Green current $\eta_4(x', \tau; g, h)$ of $Z(U(x), g)_K$ in M_K. Here $G(F)(\tau, h)K \in M_K$.

Taking the Chern forms, the $*$-product turns into wedge product and the averages, as well as the pullbacks, are preserved. $\omega_0(x_j', \tau_i)$ is the Chern form of $\eta_0(x_j', \tau_i)$ that is defined by Kudla and Millson in [Kudla 1997a], based on work from [Kudla and Millson 1986; 1987; 1990]. Furthermore, we have that

$$\omega_1(x', \tau_i) = \omega_0(x_1, \tau_i) \wedge \cdots \wedge \omega_0(x_k', \tau_i),$$
$$\omega_2(x', \tau) = p_1^*\omega_1(x', \tau_1)) \wedge p_2^*\omega_1(x', \tau_2) \wedge \cdots \wedge p_e^*\omega_1(x', \tau_e))$$

are the Chern forms of $\eta_1(x', \tau_i)$ and $\eta_2(x', \tau)$ respectively, and that

$$\omega_3(x', \tau; g, h) = \sum_{\gamma \in G_x(F)\backslash G(F)} \omega_2(x', \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h)$$

is the Chern form of the Green current $\eta_3(x', \tau; g, h)$. Finally, $\omega_3(x', \tau; g, h)$ descends to $\omega_4(x', \tau; g, h)$ corresponding to the divisor $Z(U(x), g)_K$ in M_K and is the Chern form of $\eta_4(x', \tau; g, h)$.

We defined above ω_2, ω_3 and ω_4 for $x' \in V(F)^k$ with $\dim U(x') = k$. We actually can extend the definitions of ω_2, ω_3 and ω_4 for $x \in V(F)^r$ when $\dim U(x) < r$ as well. For $x \in V(F)^r$, if $\dim U(x) = k$, we have the equality of cohomology classes $[Z(U(x), g)] = [\omega_4(x', \tau; g, h)]$ in $H^{2ek}(M_K, \mathbb{C})$ and we can actually show further that we also have

$$[Z(x, g)] = [\omega_4(x, \tau; g, h)]$$

in $H^{2er}(M_K, \mathbb{C})$. Plugging in $[\omega_4(x, \tau; g, h)]$ for the cohomology class of $[Z(x, g)]$, we take the pullback p^* of the natural projection map $p : D \times G(\mathbb{A}_f)/K \to M_K$ and unwind the sums. Then we get

$$p^*[Z(g', \phi)] = \sum_{x \in V(F)^r} r(g_f')\phi_f(x) W_{T(x)}(g_\infty')\omega_1(x, \tau). \tag{1}$$

It is enough to show that (1) is an automorphic form with values in $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$. We show this using the properties of the Kudla–Millson form on the weight of each individual $\omega_0(x, \tau_i)$, as we can rewrite (1) as

$$p^*[Z(g', \phi)] = \sum_{x \in V(F)^r} r(g'_f)\phi_f(x) r(g'_\infty)(e^{-2\pi \operatorname{tr} T(x)} \omega_1(x, \tau)),$$

and the right-hand side is a theta function of weight $(n+2)/2$ with values in $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$; thus it is automorphic.

2. Background

2A. ***Complex geometry.*** We will recall now some background from complex geometry (see for example [Chriss and Ginzburg 1997; Griffiths and Harris 1978]).

Let X be a connected compact complex manifold of dimension m. Suppose Y is a closed compact complex submanifold of codimension d. Then Y has no boundary and is thus a $2(m-d)$ chain in X. We can take the class of Y to be $[Y] \in H_{2(m-d)}(X, \mathbb{C})$. Note that we have the perfect pairing

$$H_{2(m-d)}(X, \mathbb{C}) \times H_{dR}^{2(m-d)}(X, \mathbb{C}) \to \mathbb{C},$$

given by $(Y, \eta) \to \int_Y \eta$. Thus $H_{2(m-d)}(X, \mathbb{C}) \simeq H_{dR}^{2(m-d)}(X, \mathbb{C})^\vee$. We also have the perfect pairing

$$H_{dR}^{2(m-d)}(X, \mathbb{C}) \times H_{dR}^{2d}(X, \mathbb{C}) \to \mathbb{C},$$

given by $(\eta, \omega) \to \int_X \eta \wedge \omega$. Thus $H_{dR}^{2(m-d)}(X, \mathbb{C})^\vee \simeq H_{dR}^{2d}(X, \mathbb{C})$. We can compose these isomorphisms to get

$$H_{2(m-d)}(X, \mathbb{C}) \simeq H_{dR}^{2d}(X, \mathbb{C}). \tag{2}$$

For X noncompact, we similarly can take the isomorphism

$$H_{2(m-d)}^{BM}(X, \mathbb{C}) \simeq H_{dR,c}^{2(m-d)}(X, \mathbb{C})^\vee \simeq H_{dR}^{2d}(X, \mathbb{C}), \tag{3}$$

where the first group is the Borel–Moore homology, which allows infinite linear combinations of simplexes, while the second group is the de Rham cohomology with compact support, which uses closed differential forms with compact support.

Now for Y a closed submanifold of X, in light of the above isomorphisms, a closed $2d$-form ω on X in $H_{dR}^{2d}(X, \mathbb{C})$ represents the class $[Y]$ in $H_{2(m-d)}(X, \mathbb{C})$ (respectively $H_{2(m-d)}^{BM}(X, \mathbb{C})$ when X noncompact), if and only if

$$\int_Y \eta = \int_X \omega \wedge \eta$$

for any closed $2(m-d)$ form η on X.

If X is not connected, we restrict the above to each of the connected components.

2B. ***Green currents and Chern forms.*** We recall some background on Green currents, following mainly [Gillet and Soulé 1990].

Let X be a quasiprojective complex manifold of dimension m. We define $A^{p,q}(X)$ and $A_c^{p,q}(X)$ to be the spaces of (p,q)-differential forms and (p,q)-differential forms with compact support, respectively. Let $D_{p,q}(X) = A_c^{p,q}(X)^*$ be the space of functionals that are continuous in the sense of de Rham [1955]. That is, for a sequence $\{\omega_r\} \in A^{p,q}(X)$ with support contained in a compact set $K \subset X$ and for $T \in D_{p,q}(X)$, we must have $T(\omega_r) \to 0$ if $\omega_r \to 0$, meaning that the coefficients of ω_r and finitely many of their derivatives tend uniformly to 0.

We also recall the differential operators

$$d = \partial + \bar{\partial}, \quad d^c = \frac{1}{4\pi i}(\partial - \bar{\partial}), \quad dd^c = \frac{i}{2\pi}\partial\bar{\partial}.$$

2B1. *Currents.* We define $D^{p,q} := D_{m-p,m-q}$ the space of (p,q)-currents. Then we have an inclusion $A^{p,q}(X) \to D^{p,q}(X)$ given by $\omega \to [\omega]$, where we define the current

$$[\omega](\alpha) = \int_X \omega \wedge \alpha, \tag{4}$$

for any $\alpha \in A_c^{m-p,m-q}(X)$.

For $Y \subset X$ a closed complex submanifold of dimension p, let $\iota : Y \hookrightarrow X$ be the natural inclusion and we also define a current $\delta_Y \in D^{p,p}(X)$ by

$$\delta_Y(\alpha) = \int_Y \iota^*\alpha,$$

for any $\alpha \in A_c^{m-p,m-p}$.

Definition 2.1. A *Green current* for a codimension p analytic subvariety $Y \subset X$ is a current $g \in D^{p-1,p-1}(X)$ such that

$$dd^c g + \delta_Y = [\omega_Y] \tag{5}$$

for some smooth form $\omega_Y \in A^{p,p}(X)$.

This means for $\eta \in A_c^{m-p,m-p}$, we have

$$\int_X g dd^c \eta = \int_X \omega_Y \wedge \eta - \int_Y \eta.$$

It implies that for a closed form with compact support η the left-hand side equals 0, and thus $\int_X \omega_Y \wedge \eta = \int_Y \eta$. Thus for g a Green current of Y in X, we have as cohomology classes in the isomorphism (3):

$$[Y] = [\omega_Y].$$

2B2. *Green functions and Green forms.* One natural way to obtain Green currents is from Green functions. For $Y \subset X$ a closed compact submanifold of codimension 1, a *Green function of* Y is a smooth function

$$g : X \backslash Y \to \mathbb{R}$$

which has a logarithmic singularity along Y. This means that, for any pair (U, f_U) with $U \subset X$ open and $f_U : U \to \mathbb{C}$ a holomorphic function such that $Y \cap U$ is defined by $f_U = 0$, the function

$$g + \log |f_U|^2 : U \backslash (Y \cap U) \to \mathbb{R}$$

extends uniquely to a smooth function on U.

This definition can be extended for $Y \subset X$ a closed complex submanifold of codimension p of X. We can define smooth forms $g_Y \in A^{p-1,p-1}(X)$ of logarithmic type along Y such that the current $[g_Y] \in D^{p-1,p-1}$ given as in (4) by

$$[g_Y](\eta) = \int_X \eta \wedge g_Y,$$

is a Green current. We call such smooth forms *Green forms of* Y *in* X. We will occasionally abuse notation and use g_Y for both the Green form and the Green current corresponding to g_Y.

2B3. *Chern forms.* Now let g be a Green function of $Y \subset X$, for Y a divisor on X. For $U \subset X$ let $f_U = 0$ be the local defining equation of $U \cap Y$. We define locally

$$\omega_U = dd^c(g + \log |f_U|^2).$$

By gluing together all ω_U we get a differentiable form ω_Y over X. We call this the *Chern form* associated to the Green function g. In general for Y of codimension p in X, for a Green form g_Y of Y in X we call ω_Y the Chern form corresponding to g_Y.

2B4. *Star product.* Another natural way to get Green currents is by taking their $*$-product. For Y, Z closed irreducible subvarieties of a smooth variety X such that Y and Z intersect properly, let g_Y, g_Z Green forms of Y and Z, respectively. Then the $*$-product $[g_Y] * [g_Z]$ is defined by Gillet and Soulé [1990] to be

$$[g_Y] * [g_Z] = [g_Y] \wedge \delta_Z + [\omega_Y] \wedge g_Z, \tag{6}$$

where $[\omega_Y] \wedge g_Z(\eta) = \int_X \eta \wedge \omega_Y \wedge g_Z$ and $[g_Y] \wedge \delta_Z = \pi_*[\pi^* g_Y]$, where $\pi : Z \to X$ is the embedding map. For the definition of pushforwards of currents see [Gillet and Soulé 1990]. We can also define similarly the $*$-product $[g_Y] * G_Z$ for g_Y a Green form of Y and G_Z a Green current for Z (see [Gillet and Soulé 1990]).

Moreover, from [Soulé 1992, Theorem 4, p. 50], when Y and Z have the Serre intersection multiplicity 1, we have that $[g_Y] * [g_Z]$ is a Green current for $Y \cap Z$ and

$$dd^c([g_Y] * [g_Z]) = [\omega_Y \wedge \omega_Z] - \delta_{Y \cap Z}. \tag{7}$$

2B5. *Pullback.* Also from [Soulé 1992, (3.2, p. 50)] for Z an irreducible smooth projective complex variety such that $f : Z \to X$ is a map with $f^{-1}(Y) \neq Z$, if g_Y is a Green form of logarithmic type along Y, $f^* g_Y$ is a Green form of logarithmic type along $f^{-1}(Y)$. We define the pullback of currents $f^*[g_Y] := [f^* g_Y]$ and, when the components of $f^{-1}(Y)$ have Serre intersection multiplicity 1, the current $f^*[g_Y]$ satisfies

$$dd^c f^*[g_Y] + \delta_{f^{-1}(Y)} = [f^* \omega_Y]. \tag{8}$$

3. Construction of Green currents and Chern forms

In this section we construct a Green current of $Z(U, g)_K$ in M_K for U a totally positive subspace of $V(F)$.

3A. ***The Shimura variety.*** Recall $\sigma_1, \ldots, \sigma_d$ are the embeddings of F into $\mathbb{R}$ and let (V, q) be a quadratic space such that $V_{\sigma_i} = V \otimes_{\sigma_i} \mathbb{R}$, has signature $(n, 2)$ for $1 \leq i \leq e$ and signature $(n+2, 0)$ otherwise. V has the inner product given by $\langle x, y \rangle = q(x+y) - q(x) - q(y)$. This can be naturally extended to V_{σ_i} at each place σ_i for $1 \leq i \leq d$, and we denote by q_i the quadratic form corresponding to this inner product.

We defined in the introduction the Hermitian symmetric domain

$$D = D_1 \times \cdots \times D_e,$$

where D_i consists of all the oriented negative definite planes in V_{σ_i}. We can actually write explicitly the definition of D_i as

$$D_i = \left\{ v \in V_{\sigma_i}(\mathbb{C}) : \langle v, v \rangle = 0,\ \langle v, \bar{v} \rangle < 0 \right\} / \mathbb{C}^\times \subset \mathbb{P}(V_{\sigma_i}(\mathbb{C})),$$

where $\langle \cdot, \cdot \rangle$ is the inner product corresponding to q_i that extends to $V_{\sigma_i}(\mathbb{C})$ by $\mathbb{C}$-linearity, and $v \mapsto \bar{v}$ is the involution on $V_{\sigma_i}(\mathbb{C}) = V_{\sigma_i} \otimes_{\mathbb{R}} \mathbb{C}$ induced by complex conjugation on $\mathbb{C}$.

We now recall the definition of $\mathrm{GSpin}(V)$. Let (V, q) be a quadratic space over F and $C(V, q) = \left(\bigoplus_k V^{\otimes k}\right)/I$ be the Clifford algebra of (V, q), where we are taking the quotient by the ideal $I = \{q(v) - v \otimes v \mid v \in V\}$.

Then $C(V, q)$ has dimension $2^{\dim(V)}$ and we have a $\mathbb{Z}$-grading on $T(V) = \bigoplus_k V^{\otimes k}$. The map $V \to V$, $v \to -v$ naturally extends to an algebra automorphism $\alpha : C(V, q) \to C(V, q)$. Then there is a natural $\mathbb{Z}/2\mathbb{Z}$-grading on $C(V, q)$ given by $C(V, q) = C_0(V, q) \oplus C_1(V, q)$, where

$$C_i(V, q) = \{x \in C(V, q) : \alpha(x) = (-1)^i x\}, \quad i = 0, 1.$$

We naturally have $V \subset C_1(V, q)$. Then we can define the GSpin group of V:

$$\mathrm{GSpin}(V) = \{g \in C_0(V, q)^\times \mid gVg^{-1} = V\}.$$

We denote $G = \mathrm{GSpin}(V)$ and note that G acts on V by conjugation. The group $\mathrm{Res}_{F/\mathbb{Q}} G$ is reductive over $\mathbb{Q}$ and the pair $(\mathrm{Res}_{F/\mathbb{Q}} G, D)$ is a Shimura datum. For $K \subset G(\mathbb{A}_F)$ an open compact subgroup, this gives us the complex Shimura variety

$$M_K(\mathbb{C}) \simeq G(F) \backslash D \times G(\mathbb{A}_f) / K.$$

For more details on the Shimura variety M_K see [Shih 1978].

We also define the complex line bundle L_{D_i} to be the restriction to D_i of the tautological complex line bundle on $\mathbb{P}(V_{\sigma_i}(\mathbb{C}))$. Then for the projection maps $p_i : D \to D_i$, we get the line bundles $p_i^* L_{D_i} \in \mathrm{Pic}(D)$, which further descend to the line bundles $L_{K,i} \in \mathrm{Pic}(M_K) \otimes \mathbb{Q}$ over M_K, defined to be

$$L_{K,i} = G(F) \setminus (p_i^* L_{D_i} \times G(\mathbb{A}_f)/K).$$

3B. ***Green functions of $D_{x,i}$ in D_i.*** We first recall how to construct a Green function of $D_{x,i}$ in D_i, where

$$D_{x,i} = \{\tau_i \in D, \langle \tau, x \rangle = 0\}.$$

Let $\tau \in D_i$. It corresponds to a negative definite 2-plane W in V_{σ_i} and we can write any $x \in V_{\sigma_i}$ as $x = x_\tau + x_{\tau^\perp}$, where $x_\tau \in W$ and $x_{\tau^\perp} \in W^\perp$. We define

$$R(x, \tau) = -q_i(x_\tau), \quad q_\tau(x) = q_i(x) + 2R(x, \tau).$$

Note that this implies $R(x, \tau) = 0$ if and only if $\tau \in D_{x,i}$. For $x \neq 0$ and $q_i(x) < 0$, then $D_{x,i}$ is empty, and the statement that $R(x, \tau) = 0$ if and only if $\tau \in D_{x,i}$ is void, thus still true.

In terms of an orthogonal basis we can write $\tau = \alpha + \beta\sqrt{-1}$ with $\alpha, \beta \in V_{\sigma_i}$ such that $\langle \alpha, \beta \rangle = 0$ and $\langle \alpha, \alpha \rangle = \langle \beta, \beta \rangle < 0$. Then τ corresponds to the negative oriented plane $W_\tau = \mathbb{R}\alpha + \mathbb{R}\beta \subset V(\mathbb{R})$, and we have

$$R(x, \tau) = -\frac{\langle x, \alpha \rangle^2}{\langle \alpha, \alpha \rangle} - \frac{\langle x, \beta \rangle^2}{\langle \beta, \beta \rangle}.$$

Another important property that we use is $R(gx, g\tau) = R(x, \tau)$. This is easily seen in the definition above as the inner product is invariant under the action of g.

Moreover, we show below that $-\log(R(x, \tau))$ is a Green function for $D_{x,i}$ in D_i:

Lemma 3.1. *For fixed $x \in V$, $x \neq 0$, and $\tau \in D_i \setminus D_{x,i}$, the function $-\log(R(x, \tau))$ is a Green function for $D_{x,i}$ in D_i.*

Proof. Recall the line bundle L_{D_i} is the restriction to D_i of the tautological complex line bundle on $\mathbb{P}(V_{\sigma_i}(\mathbb{C}))$. It has the fiber $L_\tau = \tau\mathbb{C} \subset V_{\sigma_i}(\mathbb{C})$ and we have a map

$$s_x(\tau) : L_\tau \to \mathbb{C}, \quad v \mapsto \langle x, v \rangle.$$

This defines an element $s_x(\tau) \in L_\tau^\vee$. As τ varies, we get a map

$$s_x : D_i \to L_{D_i}^\vee, \quad \tau \mapsto s_x(\tau).$$

Then s_x is a holomorphic section of the line bundle $L_{D_i}^\vee$. This section has a hermitian metric

$$\|s_x(\tau)\|^2 = \frac{|\langle x, v \rangle|^2}{|\langle v, \bar{v} \rangle|},$$

where $v \in L_\tau$ is any nonzero vector. In terms of an orthogonal basis we can write $v = \alpha + \beta\sqrt{-1}$ such that $\langle \alpha, \beta \rangle = 0$ and $\langle \alpha, \alpha \rangle = \langle \beta, \beta \rangle < 0$. Then

$$\|s_x(\tau)\|^2 = \frac{\langle x, \alpha \rangle^2 + \langle x, \beta \rangle^2}{|\langle \alpha, \alpha \rangle + \langle \beta, \beta \rangle|} = -\frac{\langle x, \alpha \rangle^2}{2\langle \alpha, \alpha \rangle} - \frac{\langle x, \beta \rangle^2}{2\langle \beta, \beta \rangle} \quad \text{and} \quad x_\tau = \frac{\langle x, \alpha \rangle}{\langle \alpha, \alpha \rangle}\alpha + \frac{\langle x, \beta \rangle}{\langle \beta, \beta \rangle}\beta.$$

Computing directly gives us $R(x,\tau) = 2\|s_x(\tau)\|^2$. It follows by a theorem of Poincaré–Lelong (see [Soulé 1992, Theorem 2, p. 41]) that $-\log(R(x,\tau))$ is a Green function for $D_{x,i}$ in D_i. □

For $x \in V(F)$, $\tau \in D_i$, we have the Green function defined by Kudla and Millson (see [Kudla 1997a]),

$$\eta(x,\tau) = f(2\pi R(x,\tau)), \tag{9}$$

where $f(t) = -\operatorname{Ei}(-t) = \int_t^\infty (e^{-x}/x)\,dx$ is the exponential integral. Note that

$$f(t) = -\log(t) - \gamma - \int_0^t \frac{e^{-x}-1}{x}\,dx,$$

where γ is the Euler–Mascheroni constant. The function $f(t)$ is smooth on $(0,\infty)$, $f(t)+\log(t)$ is infinitely differentiable on $[0,\infty)$, and $f(t)$ decays rapidly as $t \to \infty$. Thus using Lemma 3.1 we easily see that $\eta(x,\tau)$ is a Green function of $D_{x,i}$ in D_i.

Furthermore, Kudla and Millson have constructed explicitly the Chern form $\varphi_{KM}^{(1)}(x,\tau)$ of $\eta(x,\tau)$. We recall its definition and properties in the following section.

Note that we can consider $\eta(x,\tau)$ as a restriction to D_i of the Green function $f(2\pi\|s_x(v)\|^2) = f(2\pi|\langle x,v\rangle|^2/|\langle v,\bar{v}\rangle|)$ of $\mathbb{P}_x(V_{\sigma_i(\mathbb{C})}) := \{v \in \mathbb{P}_x(V_{\sigma_i(\mathbb{C})}) : \langle v,x\rangle = 0\}$ inside $\mathbb{P}(V_{\sigma_i}(\mathbb{C}))$. Then the theory of Section 2B, in particular the definition of the $*$-product, hold by restricting to D_i.

3C. ***The Kudla–Millson form φ_{KM}.*** We will now recall some results from [Kudla 1997a], based on previous work of Kudla and Millson [1986; 1987; 1990]. Our goal is to present explicitly the construction of the form $\varphi_{KM}^{(1)}$.

For this section we will use the notation $V_{\mathbb{R}}$ for a quadratic space over $\mathbb{R}$ with signature $(n,2)$, $G = \mathrm{GSpin}(V_{\mathbb{R}})$ and D the space of oriented negative 2-planes in $V_{\mathbb{R}}$. We fix a point $z_0 \in D$ and let $K = \mathrm{Stab}(z_0)$ be its stabilizer in $\mathrm{GSpin}(V_{\mathbb{R}})$. Then

$$D \simeq G/K \simeq \mathrm{SO}(n,2)/(\mathrm{SO}(n)\times\mathrm{SO}(2)).$$

Let $\mathfrak{g}_0 = \mathrm{Lie}(G)$ be the Lie algebra of G and $\mathfrak{k}_0 = \mathrm{Lie}(K)$ be the Lie algebra of K. We denote the complexifications of these lie algebras by $\mathfrak{g}$ and $\mathfrak{k}$, respectively. We also can identify the Lie subalgebra $\mathfrak{p}_0 \subset \mathfrak{g}_0$ given by

$$\mathfrak{p}_0 = \left\{ \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} : B \in M_{n\times 2}(\mathbb{R}) \right\} \simeq M_{n\times 2}(\mathbb{R}).$$

Moreover, we can give $\mathfrak{p}_0$ a complex structure using $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{R})$ acting as multiplication on the right. We denote by $\mathfrak{p}_+$ and $\mathfrak{p}_-$ the $\pm i$ eigenspaces of $\mathfrak{p}$. Then we have a Harish-Chandra decomposition

$$\mathfrak{g} = \mathfrak{k} + \mathfrak{p}_+ + \mathfrak{p}_-.$$

Moreover for the space of differential forms of type (a,b) on D we have an isomorphism

$$\Omega^{a,b}(D) \simeq \left[C^\infty(G) \otimes \textstyle\bigwedge^{a,b}(\mathfrak{p}^*)\right]^K,$$

where on the right-hand side we have the wedge product $\bigwedge^{a,b}(\mathfrak{p}^*) = \bigwedge^a \mathfrak{p}_+^* \wedge \bigwedge^b \mathfrak{p}_-^*$ for $\mathfrak{p}_+^*, \mathfrak{p}_-^*$ the dual spaces of $\mathfrak{p}_+$ and $\mathfrak{p}_-$, respectively.

Recall that $\widetilde{\mathrm{Sp}}_{2m}(\mathbb{R})$ is the metaplectic cover of $\mathrm{Sp}_{2m}(\mathbb{R})$, and let K' be the preimage under the projection map $\widetilde{\mathrm{Sp}}_{2m}(\mathbb{R}) \to \mathrm{Sp}_{2m}(\mathbb{R})$ of the compact subgroup

$$\left\{\left(\begin{smallmatrix} A & B \\ -B & A \end{smallmatrix}\right),\ A + iB \in U(m)\right\},$$

where $U(m)$ is the unitary group. The group K' has a character $\det^{1/2}$ whose square descends to the determinant character of $U(m)$.

Then Kudla and Millson constructed a Schwartz form

$$\varphi_{KM}^{\circ,(m)}(x, \tau) \in (\mathcal{S}(V_{\mathbb{R}}^m) \otimes \Omega^{m,m}(D))^G,$$

where $\mathcal{S}(V_{\mathbb{R}}^m)$ is the Schwartz space over $V_{\mathbb{R}}^m$, and by invariance under G we mean

$$\varphi_{KM}^{\circ,(m)}(gx, g\tau) = \varphi_{KM}^{\circ,(m)}(x, \tau).$$

We present their result below:

Theorem. *There exists an element $\varphi_{KM}^{\circ,(m)}(x, \tau) \in (\mathcal{S}(V_{\mathbb{R}}^m) \otimes \Omega^{m,m}(D))^G$ with the following properties*:

(1) *For $k' \in K'$ such that $\iota(k') = \left(\begin{smallmatrix} A & B \\ -B & A \end{smallmatrix}\right)$ under the natural map $\iota : \widetilde{\mathrm{Sp}}_{2m}(\mathbb{R}) \to \mathrm{Sp}_{2m}(\mathbb{R})$, we have*

$$r(k')\varphi_{KM}^{\circ,(m)} = (\det(k'))^{(n+2)/2}\varphi_{KM}^{\circ,(m)}.$$

(2) $d\varphi_{KM}^{\circ,(m)} = 0$, *i.e., for any $x \in V_{\mathbb{R}}^m$, the form $\varphi_{KM}^{\circ,(m)}(x, \cdot)$ is a closed (m, m)-form on D which is G_x-invariant.*

We define below $\varphi_{KM}^{\circ,(m)}$ explicitly following [Kudla 1997a]. The form $\varphi_{KM}^{(m),\circ}$ is denoted by $\varphi^{(m)}$ in [Kudla 1997a]. First we will construct $\varphi_{KM}^{\circ,(1)}$.

Note that we have an isomorphism

$$[S(V_{\mathbb{R}}) \otimes \Omega^{1,1}(D)]^G \simeq \left[S(V_{\mathbb{R}}) \otimes \textstyle\bigwedge^{1,1}\mathfrak{p}^*\right]^K$$

given by evaluating at z_0. Recall that we identified the Lie algebra $\mathfrak{p}_0 = \left\{\left(\begin{smallmatrix} 0 & B \\ B^T & 0 \end{smallmatrix}\right) : B \in M_{n\times 2}(\mathbb{R})\right\} \simeq M_{n\times 2}(\mathbb{R})$. Then we have the differential forms $\omega_{i,j} \in \Omega^1(D) = \Omega^{1,0}(D) \oplus \Omega^{0,1}(D)$, $1 \le i \le n$, $1 \le j \le 2$, defined by the function $\omega_{i,j} \in \mathfrak{p}_0^*$, $\omega_{i,j} : \mathfrak{p}_0 \simeq M_{n\times 2}(\mathbb{R}) \to \mathbb{R}$ given by the map $u = (u_{st})_{1\le s\le n, 1\le t\le 2} \to u_{ij}$.

We first define for $x = (x^{(1)}, \ldots, x^{(n+2)}) \in V_{\mathbb{R}}$ the form $\varphi_{KM}^{(1)}(x)$ that is also G-invariant:

$$\varphi_{KM}^{(1)}(x) = e^{-2\pi R(x,z_0)}\left(\sum_{i,j=1}^{n} 2x^{(i)}x^{(j)}\omega_{i,1} \wedge \omega_{j,2} - \frac{1}{2\pi}\sum_{i=1}^{n}\omega_{i,1} \wedge \omega_{i,2}\right). \tag{10}$$

We further define $\varphi_{KM}^{\circ,(1)}(x)$ to be $\varphi_{KM}^{\circ,(1)}(x) = e^{-2\pi q_{z_0}(x)}\varphi_{KM}^{(1)}(x)$, and finally, for $x = (x_1, \ldots, x_m) \in V^m$ we define

$$\varphi_{KM}^{(m)}(x) = \varphi_{KM}^{(1)}(x_1) \wedge \cdots \wedge \varphi_{KM}^{(1)}(x_m), \tag{11}$$

as well as

$$\varphi_{KM}^{\circ,(m)}(x) = e^{-2\pi\sum_{i=1}^{m} q_{z_0}(x_i)}\varphi_{KM}^{(m)}(x).$$

Recall the Green function $\eta(x,\tau) = f(2\pi R(x,\tau))$, where $x \in V(F)$ and $\tau \in D_i$. It has the important property [Kudla 1997b, Proposition 4.10]

$$dd^c[\eta(x,\cdot)] + \delta_{D_{x,i}} = [\varphi_{KM}^{(1)}(x,\cdot)], \tag{12}$$

where $\varphi_{KM}^{(1)} \in (S(V) \otimes \Omega^{1,1}(D_i))^K$ is the Schwartz form defined above. This implies that $\varphi_{KM}^{(1)}(x,\tau)$ is the Chern form corresponding to the Green function $\eta(x,\tau)$. Note that (12) is mentioned in [Kudla 2003, Theorem 4.10] for $F = \mathbb{Q}$, but holds in general for F with a fixed real place σ_i for which V_{σ_i} has signature $(n,2)$.

3D. ***Averaging of Green currents and their Chern forms.*** Now let $x = (x_1,\dots,x_r) \in V(F)^r$ such that $U(x) = \mathrm{Span}_F\{x_1,\dots,x_r\}$ is a totally positive k-subspace of $V(F)$, $k \le r$. Our goal is to construct a Green current of $Z(U(x), g)$ in M_K and its corresponding Chern form.

We define $x' = (x_1',\dots,x_k')$ such that $x_1' = x_{i_1},\dots,x_k' = x_{i_k}$ and $U(x') = U(x)$. To make this uniquely defined, we pick the smallest indices $(i_1,\dots,i_k)$ for which this happens. Note further that as $U(x) = U(x')$, we also have $D_x = D_{x'}$, $V_x = V_{x'}$ and $G_x = G_{x'}$.

For $\tau_i \in D_i$ and $x_j' \in V(F)$ for $1 \le j \le r$, $1 \le i \le e$, we define as in (9):

$$f_i(x_j',\tau_i) := f(2\pi R(x_j',\tau_i))$$

that is a Green function of $D_{x_j',i}$ in D_i.

We can further fix $z_{0,i} \in D_i$ for $1 \le i \le e$ and we define the Kudla–Millson forms $\varphi_{KM}^{(1)}(x_j',\tau_i) \in (\mathcal{S}(V) \otimes \Omega(D_i)^{(1,1)})^G$ for $\tau_i \in D_i$, $x_j' \in \mathcal{S}(V)$, as in Section 3C, that satisfy the equation

$$dd^c[f_i(x_j',\cdot)] + \delta_{D_{x_j',i}} = [\varphi_{KM}^{(1)}(x_j',\cdot)]. \tag{13}$$

As $x_1',\dots,x_k'$ are linearly independent, the submanifolds $D_{x_j',i}$ intersect properly inside D_i and thus we can take the $*$-product of the Green functions $f_i(x_j',\tau_i)$ for $1 \le j \le k$. Denote

$$\eta_1(x',\tau_i) = f_i(x_1',\tau_i) * \cdots * f_i(x_k',\tau_i).$$

Then, from (7), this is a Green current for $D_{x,i} = D_{x',i} = D_{U(x'),i} = \bigcap_{j=1}^k D_{x_j',i}$ in D_i for $1 \le i \le e$.

As the star product turns into wedge product when we take the Chern forms (see (7)), the Chern form associated to $\eta_1(x_j,\tau)$ is going to be

$$\omega_1(x',\tau_i) = \varphi_{KM}^{(1)}(x_1',\tau_i) \wedge \cdots \wedge \varphi_{KM}^{(1)}(x_k',\tau_i).$$

Note that $\omega_1(x',\tau_i) = \varphi_{KM}^{(k)}(x',\tau_i)$ and thus from the definition (6) of the star product, η_1 satisfies

$$dd^c[\eta_1(x',\cdot)] + \delta_{D_{x,i}} = [\varphi_{KM}^{(k)}(x',\cdot)]. \tag{14}$$

Let $p_i : D \to D_i$ be the natural projections as before. Then, from (8), $p_i^*\eta_1(x,\tau_i)$ is a Green function of $p_i^* D_{x,i}$ in D and the form $p_i^*\varphi_{KM,i}^{(k)}(x',\tau_i)$ satisfies

$$dd^c[p_i^*\eta_1(x',\cdot)] + \delta_{D_{x,i}} = [p_i^*\varphi_{KM}^{(k)}(x',\cdot)]. \tag{15}$$

By taking the $*$-product, for $\tau = (\tau_1, \ldots, \tau_e) \in D \setminus D_x$ we define

$$\eta_2(x', \tau) = p_1^* \eta_1(x', \tau_1) * \cdots * p_e^* \eta_1(x', \tau_e).$$

This is a Green current of D_x in D. This follows from (8), as the divisors $p_i^* D_{x,i}$ have Serre's intersection multiplicity 1 in D. The Chern form of $\eta_2(x', \tau)$ is going to be

$$\omega_2(x', \tau) = p_1^* \omega_1(x', \tau_1) \wedge \cdots \wedge p_e^* \omega_1(x', \tau_e),$$

satisfying

$$dd^c[\eta_2(x', \cdot)] + \delta_{D_x} = [\omega_2(x', \cdot)]. \tag{16}$$

We further take for $(\tau, h) \in D \times G(\mathbb{A}_f)$ the average of Green currents:

$$\eta_3(x', \tau; g, h) = \sum_{\gamma \in G_x(F) \setminus G(F)} \eta_2(x', \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h).$$

Note that this can be rewritten as

$$\eta_3(x', \tau; g, h) = \sum_{\gamma \in \Gamma_h} \eta_2(\gamma^{-1} x', \tau),$$

where $\Gamma_h = G_x(F) \setminus G(F) \cap G_x(\mathbb{A}_f) g K h^{-1}$ is a lattice in $G(F)$. It is clear from the average that η_3 has a singularity along $G(F)(D_x \times G_x(\mathbb{A}_f) g K/K)$ in $D \times G(\mathbb{A}_f)/K$. However, note that it is not obvious that this function converges. We are actually going to prove in Section 3G the following proposition:

Proposition 3.2. *Let $x \in V(F)^k$ such that $U(x)$ is a totally positive k-subspace of $V(F)$. Then we have that the defining sum of $\eta_3(x, \tau; g, h)$ is absolutely convergent and $\eta_3(x, \tau; g, h)$ is a Green current of $G(F)(D_x \times G_x(\mathbb{A}_f) g K/K)$ in $D \times G(\mathbb{A}_f)/K$.*

This implies that $\eta_3(x', \tau; g, h)$ is a Green current of $G(F)(D_x \times G_x(\mathbb{A}_f) g K/K)$ in $D \times G(\mathbb{A}_f)/K$. To get the Chern form we apply dd^c locally and glue all the local forms using again [Soulé 1992, Theorem 4, p. 50]. This is possible due to the discussion at the end of the proof of Proposition 3.2 in Section 3G.

Then η_3 has the Chern form

$$\omega_3(x', g; \tau, h) = \sum_{\gamma \in \Gamma_h} \omega_2(\gamma^{-1} x', \tau),$$

where $\Gamma_h = G_{x'}(F) \setminus G(F) \cap G_{x'}(\mathbb{A}_f) g K h^{-1}$ as before.

As η_3 is invariant under the action of $G(F)$, it descends to a Green current via the projection map $p : D \times G(\mathbb{A}_f)/K \to M_K$ to

$$\eta_4(x', \tau; g, h),$$

where (τ, h) represent the class $G(F)(\tau, h)K$ in M_K. The Green current condition (5) is also preserved under the projection map, and the singularity is given by exactly the cycle $Z(U(x), g)_K$ inside the Shimura variety M_K. Thus we get:

Proposition 3.3. *For x' defined as above, $\eta_4(x', \tau; g, h)$ is a Green current of $Z(U(x), g)_K$ in M_K.*

Note that $\omega_3(x', \tau; g, h)$ descends as well to the Chern form $\omega_4(x', \tau; g, h)$ of $\eta_4(x', \tau; g, h)$. Moreover, the Chern form $\omega_3(x', \tau; g, h)$ is the pullback under the projection map $p : D \times G(\mathbb{A}_f)/K \to M_K$ of $\omega_4(x', \tau)$:

$$\omega_3(x', \tau; g, h) = p^* \omega_4(x', \tau; g, h).$$

3E. ***Extending notation.*** In the previous section we have defined the Chern forms $\omega_2, \omega_3, \omega_4$ for $x' = (x'_1, \ldots, x'_k)$ with the coordinates $x'_1, \ldots, x'_k$ linearly independent. We want to extend the definition to $x = (x_1, \ldots, x_k)$ in $V(F)^k$ when the coordinates $x_1, \ldots, x_k$ are linearly dependent over F. In order to do that, we take $\omega_1(x, \tau_i) = \varphi^{(k)}_{KM}(x, \tau_i)$, $\omega_2(x, \tau) = p_1^* \omega_1(x, \tau_1) \wedge \cdots \wedge p_e^* \omega_1(x, \tau_e)$, and

$$\omega_3(x, \tau; g, h) = \sum_{\gamma \in G_x(F) \backslash G(F)} \omega_2(x, \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h).$$

We will show in Section 3G in Proposition 3.9 that ω_3 is well-defined.

Also note that for U a totally positive k-dimensional subspace of $V(F)$ we can pick any $y = (y_1, \ldots, y_k)$ such that $U(y) = U$ and $\eta_4(y, \tau; g, h)$ is going to be a Green current of $Z(U, g)$ in M_K with its corresponding Chern form $\omega_4(y, \tau; g, h)$.

We can actually extend the definition of $\eta_2, \eta_3, \omega_2, \omega_3$ for $v \in \mathrm{GL}_k(F_\infty)$ when $x = (x_1, \ldots, x_k) \in V(F)^k$ such that $U(x)$ is a totally positive k-plane inside of V. We define

$$\eta_2(vx, \tau) = p_1^* \eta_1(v_1 x, \tau_1) * \cdots * p_e^* \eta_1(v_e x, \tau_e),$$

where $v_i = \sigma_i(v) \in \mathrm{GL}_k(\mathbb{R})$ for $1 \le i \le e$. Note that $G_{v_i x} = G_x$ and $D_{v_i x, i} = D_{x,i}$ for all $1 \le i \le e$ and $\eta_2(vx, \tau)$ is a Green form of D_x in D.

We define further

$$\eta_3(vx, \tau; g, h) = \sum_{\gamma \in G_x(F) \backslash G(F)} \eta_2(vx, \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h),$$

where $\eta_3(vx, \tau; g, h)$ is a Green form of $G(F)(D_x \times G_x(\mathbb{A}_f) g K/K)$ in $D \times G(\mathbb{A}_f)/K$. The proof of convergence is similar to the one for $\eta_3(x, \tau; g, h)$.

The Chern forms of $\eta_2(vx, \tau)$ and $\eta_3(vx, \tau)$ are going to be, respectively,

$$\omega_2(vx, \tau) = p_1^* \omega_1(v_1 x, \tau_1) \wedge \cdots \wedge p_e^* \omega_1(v_e x, \tau_e),$$
$$\omega_3(vx, \tau; g, h) = \sum_{\gamma \in G_x(F) \backslash G(F)} \omega_2(vx, \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h).$$

Propositions 3.2 and 3.9 extend as well for $\eta_3(vx, \tau; g, h)$ and $\omega_3(vx, \tau; g, h)$, thus they are well defined. As they are invariant under the action of $G(F)$, η_3 and ω_3 further descend to the Green current $\eta_4(vx, \tau; g, h)$ of $Z(U(x), g)$ in M_K that has the corresponding Chern form $\omega_4(vx, \tau; g, h)$.

Moreover, we extend the notation of ω_2, ω_3 for $x = (x_1, \dots, x_k)$ with $\dim U(x) \le k$ by taking

$$\omega_2(vx, \tau) = p_1^*\omega_1(v_1x, \tau_1) \wedge \cdots \wedge p_e^*\omega_1(v_ex, \tau_e),$$
$$\omega_3(vx, \tau; g, h) = \sum_{\gamma \in G_x(F)\backslash G(F)} \omega_2(vx, \gamma\tau) 1_{G_x(\mathbb{A}_f)gK}(\gamma h).$$

Proposition 3.9 extends as well, making ω_3 well-defined in general.

3F. ***Chern forms for $x = 0$.*** Recall that we defined in Section 3A the line bundles $L_{K,i} \in \mathrm{Pic}(M_{K,i}) \otimes \mathbb{Q}$. For $x = 0$, we claim that we can still define ω_i for $1 \le i \le 4$ and the same relationships hold as in Section 3D. Moreover, we are going to have

$$Z(0, g) = \omega_4(0, \tau).$$

We define the Chern form $\omega_1(0, \tau_i) = (-1)^r \varphi_{KM}^{(r)}(0, \tau_i)$. Here recall

$$\varphi_{KM}^{(1)}(0, \tau_i) = -\frac{1}{2\pi} \sum_{j=1}^{n} \omega_{j,1} \wedge \omega_{j,2}(\tau_i)$$

and $\varphi_{KM}^{(r)}(0, \tau_i) = \bigwedge^r \varphi_{KM}^{(1)}(0, \tau_i)$ as defined in Section 3C.

Lemma 3.4. *$\varphi_{KM}^{(1)}(0, \tau_i) = -c_1(L_{D_i}^\vee)$, for $1 \le i \le e$.*

This is Corollary 4.12 in [Kudla 2003]. Kudla considers $F = \mathbb{Q}$, but the result is unchanged for a totally real number field F with a fixed embedding σ_i into $\mathbb{R}$ such that V_{σ_i} has signature $(n, 2)$.

Thus from Lemma 3.4 we have $\omega_1(0, \tau_i) = (-1)^r c_1(L_{D_i}^\vee)^r$. Then as before we define $\omega_2(0, \tau) = p_1^*\omega_1(0, \tau_1) \wedge \cdots \wedge p_e^*\omega_1(0, \tau_e)$. Note that $\omega_2(0, \tau) = (-1)^{re} p_1^* c_1(L_{D_1}^\vee)^r \wedge \cdots \wedge p_e^* c_1(L_{D_e}^\vee)^r$. Furthermore, as $G_0 = G$, when we average over $\Gamma_h = G_0(F) \setminus (G(F) \cap G_0(\mathbb{A}_f) g K h^{-1})$ we get

$$\omega_3(0, \tau; g, h) = \omega_2(0, \tau).$$

Moreover, we have as before $\omega_3(0, \tau) = p^*\omega_4(0, \tau)$, and thus

$$\omega_4(0, \tau) = (-1)^{re} p^* p_1^* c_1(L_{D_1}^\vee)^r \cdots p^* p_e^* c_1(L_{D_e}^\vee)^r = (-1)^{re} c_1(L_K^\vee)^r,$$

where $c_1(L_K^\vee) := c_1(L_{K,1}^\vee) \cdots c_1(L_{K,e}^\vee)$. Finally, note that $\omega_4(0, \tau)$ is exactly the cycle $Z(0, g)_K$ in M_K.

3G. ***Convergence of $\eta_3(x, \tau; g, h)$ and $\omega_3(x, \tau; g, h)$.*** Now we are ready to show the convergence of $\eta_3(x, \tau; g, h)$. More precisely, we are going to prove Proposition 3.2.

Before we continue, we mention two short lemmas that tell us about the behavior of $R(x, \tau)$ when τ varies in a compact set in D_i and x varies in a lattice. The first lemma tells us that the quadratic forms q_τ bound each other:

Lemma 3.5. *Let $K_i \subset D_i$ be a compact set. Fix $\tau_0 \in K_i$. Then there exist $c, d > 0$ such that*

$$c q_{\tau_0}(x) \le q_\tau(x) \le d q_{\tau_0}(x)$$

for all $\tau \in K_i$.

Proof. Let $\tau \in K_i$ and $x \in V$, $x \neq 0$. Consider the function $\psi : K_i \times \{x \in V \mid q_{\tau_0}(x) = 1\} \to \mathbb{R}$, $\psi(\tau, x) = q_\tau(x)$. Since q_{τ_0} is positive definite, the set of vectors of norm 1 is a sphere and thus compact. Hence the domain is compact and thus the image is compact, and thus bounded. Since $x \neq 0$, it must also be bounded away from 0. Thus we can find constants c, d such that

$$c \leq q_\tau\left(\frac{x}{\sqrt{q_{\tau_0}(x)}}\right) \leq d$$

and $cq_{\tau_0}(x) \leq q_\tau(x) \leq dq_{\tau_0}(x)$ as desired. □

The second lemma tells us how $R(x, \tau)$ increases when x varies in a lattice:

Lemma 3.6. *For a compact set $K_0 \subset D$ and a lattice $\Gamma \subset G(F)$, there are only finitely many $\gamma \in \Gamma$ such that $R(\gamma^{-1}x, \tau_i) \leq N$ for any $\tau = (\tau_1, \ldots, \tau_e) \in K_0$. More precisely, if $\dim V = n + 2$, we have at most $O(N^{n/2+1})$ such $\gamma \in \Gamma$.*

Proof. Fix some $\tau_0 \in K_0 \cap D_i$. If for $y \in \Gamma x$ we have $R(y, \tau_i) = (q_{\tau_i}(y) - a)/2 < N$, then from the previous lemma this implies that there exists $c > 0$ such that $q_{\tau_0}(y) < (a + 2N)/c$. Thus y lies in a $(n+2)$-dimensional sphere in V of radius $\sqrt{(a+2N)/c}$. The result follows. □

Now we want to compute the summands of

$$\eta_3(x, g; \tau, h) = \sum_{\gamma \in \Gamma_h} p_1^* \eta_1(\gamma^{-1}x, \tau_1) * p_2^* \eta_1(\gamma^{-1}x, \tau_2) * \cdots * p_e^* \eta_e(\gamma^{-1}x, \tau_e), \tag{17}$$

where $\Gamma_h = G_x(F) \setminus G(F) \cap G_x(\mathbb{A}_f) g K h^{-1}$. Recall $\eta_1(x, \tau_i) = \eta_0(x_1, \tau_i) * \cdots * \eta_0(x_k, \tau_i)$, where $\eta_0(x, \tau_i) = f(2\pi R(x, \tau_i))$.

We compute first the general formula for the $*$-product of N Green currents:

Lemma 3.7. *Let $f_1, \ldots, f_N$ Green forms for the cycles $Y_1, \ldots, Y_N$ inside X, chosen such that the star product $[f_1] * \cdots * [f_N]$ is well-defined. Let $\varphi_1, \ldots, \varphi_N$ be their corresponding Chern forms. Then we have the $*$-product of N-terms*:

$$[f_1] * [f_2] * \cdots * [f_N] = \sum_{j=1}^{N} \varphi_1 \wedge \cdots \varphi_{j-1} \wedge [f_j] \wedge \delta_{Y_{j+1}} \wedge \cdots \wedge \delta_{Y_N}.$$

Proof. We denote $\delta_{i,j} = \delta_i \wedge \delta_{i+1} \cdots \wedge \delta_j$, $\varphi_{i,j} = \varphi_i \wedge \cdots \wedge \varphi_j$ for $i \leq j$ and we take $\delta_{i,j} = \varphi_{i,j} = 1$ for $i > j$. We show the result by induction. For $n = 2$, we have $[f_1] * [f_2] = f_1 \wedge \delta_2 + \varphi_1 \wedge f_2$. Assume the result is true for n. Then we have

$$[f_2] * [f_3] * \cdots * f_{n+1} = \sum_{k=2}^{n+1} \varphi_{2,k-1} \wedge [f_k] \wedge \delta_{k+1,n+1}.$$

By definition, we have

$$[f_1]*([f_2]*[f_3]*\cdots*[f_{n+1}]) = [f_1]\wedge\delta_{2,n+1} + \varphi_1\wedge([f_2]*[f_3]*\cdots*[f_{n+1}])$$
$$= [f_1]\wedge(\delta_{2,n+1}) + \sum_{k=2}^{n+1}\varphi_1\wedge\varphi_{2,k-1}\wedge[f_k]\wedge\delta_{k+1,n+1}.$$

This is exactly $\sum_{k=1}^{n+1}\varphi_{1,k-1}\wedge[f_k]\wedge\delta_{k+1,n+1}$ which finishes the proof.

We want to apply the above lemma to each of the $*$-products summands in (17) that define η_3:

$$p_1^*\eta_0(\gamma^{-1}x_1,\tau_1)*\cdots p_1^*\eta_0(\gamma^{-1}x_k,\tau_1)*\cdots* p_e^*\eta_e(\gamma^{-1}x_1,\tau_e)*\cdots* p_e^*\eta_e(\gamma^{-1}x_k,\tau_e).$$

Denote $f_i = p_i^*\eta_0$ and $\varphi_i = p_i^*\omega_0$. Then we get the terms

$$\sum_{i=1}^{e}\sum_{j=1}^{k}\varphi_1(\gamma^{-1}x_1,\tau_1)\wedge\cdots\wedge f_i(\gamma^{-1}x_j,\tau_1)\wedge\cdots\wedge\delta_{p_e^*D_{x_k}}, \tag{18}$$

where all the terms before f_i are the smooth forms φ and all the terms following f_i are the operators δ. □

Proof of Proposition 3.2. To show the convergence of η_3, we need to show that for μ a smooth form with compact support, the integral $\int_X \eta_3\wedge\mu$ converges, where $X = D\times G(\mathbb{A}_f)/K$. Note that we can cover the compact support $\mathrm{supp}(\mu)$ of μ by finitely many open sets and in each of them we can write μ in local coordinates as a linear combination of smooth functions that are bounded inside $\mathrm{supp}(\mu)$. Thus it is enough to show that the form η_3 converges to a smooth form on compacts.

We are interested in averaging the terms (18):

$$\sum_{i=1}^{e}\sum_{j=1}^{k}\varphi_1(y_1,\tau_1)\wedge\cdots\wedge f_i(y_j,\tau_1)\wedge\cdots\wedge\delta_{p_e^*D_{x_k}},$$

for τ inside a compact set $K_0\subset D$, where the average is taken over $y=(y_1,\dots,y_k)\in\Gamma_h x$. For the terms containing at least one δ, the terms

$$\varphi_1(\gamma^{-1}x_1,\tau_1)\wedge\cdots\wedge f_i(\gamma^{-1}x_j,\tau_1)\wedge\cdots\wedge\delta_{p_e^*D_{x_k}}$$

are nonzero only for $\tau_e\in D_{\gamma^{-1}x_k,e}$. However, this implies $R(\gamma^{-1}x_k,\tau_e)=0$ and this only happens for finitely many $\gamma\in\Gamma$ when $\tau_e\in K_0$ inside a compact from Lemma 3.6. Thus the sum

$$F_1(x,\tau) = \sum_{j=1}^{k}\sum_{\substack{i=1\\(i,j\neq(e,k))}}^{e}\sum_{\gamma\in\Gamma_h}\varphi_1(\gamma^{-1}x_1,\tau_1)\wedge\cdots\wedge f_i(\gamma^{-1}x_j,\tau_1)\wedge\cdots\wedge\delta_{p_e^*D_{x_k}}$$

is finite. This leaves the last term,

$$F_2(x,\tau) = \sum_{\gamma\in\Gamma_h}\varphi_1(\gamma^{-1}x_1,\tau_1)\wedge\cdots\wedge\varphi_e(\gamma^{-1}x_{k-1},\tau_e)\wedge f_e(\gamma^{-1}x_k,\tau_e),$$

which we treat below in Lemma 3.8. We show that the sum $F_2(x,\tau)$ converges uniformly on compacts to a smooth form. This finishes the proof of the convergence in Proposition 3.2.

Note that $F_1(x,\tau)$ is a finite sum of forms, while $F_2(x,\tau)$ is the average of wedge products of smooth forms which converges to a smooth form.

To check the Green current condition (5) is met by $\eta_3(x,\tau;g,h)$, again it is enough to check the condition on compact sets. Note first that $\tau_i \in D_{y_i}$ only for finitely many $y \in \Gamma_h x$ when τ is inside a compact set K_0. For $\tau_i \in D_{y_i}$ then we have a finite sum of terms η_2 that satisfy the Green current condition (5): $dd^c\eta_2(y,\tau)+\delta_{D_{y,\tau}} = [\omega_2(y,\tau)]$. For all the other terms, we do not have singularities, and as $\sum_{\gamma\in\Gamma_h}\eta_2(\gamma^{-1}x,\tau)$ and all its derivatives converge to a smooth form, we can just take dd^c to get

$$dd^c\sum_{\gamma\in\Gamma_h}\eta_2(\gamma^{-1}x,\tau)=\sum_{\gamma\in\Gamma_h}dd^c\eta_2(\gamma^{-1}x,\tau)=\sum_{\gamma\in\Gamma_h}\omega_2(\gamma^{-1}x,\tau),$$

giving us the condition (5) for η_3. Moreover, note that its Chern form is

$$\omega_3(x,\tau;g,h)=\sum_{\gamma\in\Gamma_h}\omega_2(\gamma^{-1}x,\tau).$$

This finishes the proof of Proposition 3.2. □

As promised, we show the convergence of $F_2(x,\tau)$ below:

Lemma 3.8. *The average*

$$F_2(x,\tau;g,h)=\sum_{y\in\Gamma_h x}\varphi_1(y_1,\tau_1)\wedge\cdots\wedge\varphi_1(y_k,\tau_1)\wedge\cdots\wedge\varphi_e(y_1,\tau_e)\wedge\cdots\wedge\varphi_e(y_{k-1},\tau_e)\wedge f_e(y_k,\tau_e)$$

converges uniformly on compacts to a smooth form.

Proof. Let K_0 be a compact. We are free to discard finitely many terms from our average of the star product without affecting the convergence, so we discard the terms for which $f_e(y_k,\tau_e)=0$ on K_0. For $y=(y^{(1),i},\dots,y^{(n+2),i})$ coordinates determined by the point $z_{0,i}$ in $D_{y,i}$, we recall the explicit definition of $\varphi_i(y,\tau_i)=p_i^*\varphi_{KM}(y,\tau_i)$ that we presented in Section 3C:

$$\varphi_i(y,\tau_i)=e^{-2\pi R(y,z_{0,i})}\left(\sum_{1\le s,t\le n}y^{(s),i}y^{(t),i}p_i^*(\omega_{s,1i}\wedge\omega_{t,2i})-\frac{1}{\pi}\sum_{1\le s\le n}p_i^*(\omega_{s,1i}\wedge\omega_{s,2i})\right).$$

Thus, in the average, all the terms are of the form

$$e^{-2\pi\sum_{j=1}^{k}\sum_{i=1}^{e}R(y_j,z_{0,i})}e^{2\pi R(y_k,z_{0,e})}f_e(y_k,\tau_e)\bigwedge_{i=1}^{e}\bigwedge_{\substack{j=1\\(i,j)\neq(e,k)}}^{k}(y_j^{(s),i}y_j^{(t),i})^f\,p_i^*\omega_{s,1i}\wedge p_i^*\omega_{t,2i}(\tau_i).$$

The forms $p_i^*\omega_{s,1i}$, $p_i^*\omega_{s,2i}$ are smooth on K_0 and the values of the smooth functions representing them in local coordinates are bounded inside a compact. As they are independent of y, the convergence of $F_2(x,\tau)$ reduces to the convergence of

$$\sum_{y\in\Gamma_h x}e^{-2\pi\sum_{i=1}^{e-1}\sum_{j=1}^{k}R(y_j,z_{0,i})}e^{-2\pi\sum_{j=1}^{k-1}R(y_j,z_{0,e})}f_e(y_k,\tau_e)P(y).$$

Here

$$P(y)=\prod_{i=1}^{e}\prod_{\substack{j=1\\(i,j)\neq(e,k)}}^{k}\sum_{1,\leq s,t\leq n}\sum_{f=0}^{1}(y_j^{(s),i}y_j^{(t),i})^f$$

is a polynomial of degree $2k(e-1)$.

Similarly, for computing the derivatives of $F_2(x,z)$ we are reduced to computing averages of the wedge products

$$\frac{\partial}{\partial^{R_{1,1}}\tau_1\partial^{S_{1,1}}\overline{\tau_1}}\varphi_1(y_1,\tau_1)\wedge\cdots\wedge\frac{\partial}{\partial^{R_{1,k}}\tau_1\partial^{S_{1,k}}\overline{\tau_1}}\varphi_1(y_k,\tau_1)\wedge\cdots\wedge\frac{\partial}{\partial^{R_{e,1}}\tau_e\partial^{S_{e,1}}\overline{\tau_i}}\varphi_e(y_1,\tau_e)\wedge\cdots$$
$$\wedge\frac{\partial}{\partial^{R_{e,k-1}}\tau_e\partial^{S_{e,k-1}}\overline{\tau_e}}\varphi_e(y_{k-1},\tau_e)\wedge\frac{\partial}{\partial^{R_{e,k}}\tau_e\partial^{S_{e,k}}\overline{\tau_e}}f_e(y_k,\tau_e).$$

We will break the proof in two main steps below:

Step 1: We claim that it is enough to show that the sums

$$\sum_{y\in\Gamma_h x}\frac{\partial}{\partial^{R_{e,k}}\tau_e\partial^{S_{e,k}}\overline{\tau_e}}f_e(y_k,\tau_e)\tag{19}$$

converge for any integers $R_{e,k}, S_{e,k}\geq 0$.

In order to show this, let us compute first the partial derivatives in τ_i of the terms $\varphi(y_j,\tau_i)$ with $(j,i)\neq(k,e)$. We get

$$\frac{\partial}{\partial^R\tau_i\partial^S\overline{\tau_i}}\varphi(y_j,\tau_i)=e^{-2\pi R(y_j,z_{0,i})}\sum(y^{(s),i}y^{(t),i})^f\frac{\partial}{\partial^R\tau_i\partial^S\overline{\tau_i}}p_i^*\omega_{s,1i}\wedge p_i^*\omega_{t,2i}(\tau_i),$$

where $f\in\{0,1\}$ and $1\leq s,t\leq n$. Since $p_i^*\omega_{s,2i}\wedge p_i^*\omega_{t,2i}$ are smooth forms on compacts, the terms $\partial/(\partial^R\tau_i\partial^S\overline{\tau_i})p_i^*\omega_{s,1i}\wedge p_i^*\omega_{t,2i}(\tau_i)$ are smooth as well. Then the problem reduces to showing that the coefficients

$$\sum_{y\in\Gamma_h x}e^{-2\pi\sum_{i=1}^{e-1}\sum_{j=1}^{k}R(y_j,z_{0,i})}e^{-2\pi\sum_{j=1}^{k-1}R(y_j,z_{0,e})}f_e(y_k,\tau_e)P(y)\frac{\partial}{\partial^{R_{e,k}}\tau_e\partial^{S_{e,k}}\overline{\tau_e}}f_e(y_k,\tau_e)$$

converge on compacts.

We can discard finitely many terms for which we have $R(y_j,\tau_i)\leq 1$ for any pair (i,j) with $1\leq i\leq e$ and $1\leq j\leq k$. Then we can bound

$$\sum_{s,t=1}^{n}\sum_{f=0}^{1}(y_j^{(s),i}y_j^{(t),i})^f\leq(q_i(x_j)+R(y_j,\tau))^{n^2}.$$

Thus we can further bound

$$|P(y)|\leq C\prod_{i=1}^{e}\prod_{\substack{j=1\\(i,j)\neq(e,k)}}^{k}(q_i(x_j)+R(y_j,z_{0,i}))^{n^2}.$$

By discarding finitely many terms from the lattice, we can bound $e^{-2\pi R(y_j,\tau_i)}R(y_j, z_{0,i})^m \leq 1$, for any $1 \leq m \leq n^2$ and then

$$e^{-R(y_j,z_{0,i})}(q_i(x_j) + R(y_j, z_{0,i}))^{n^2} \leq (q_i(x_j) + 1)^{n^2},$$

which is a constant. Thus we need to show that the sums

$$C' \sum_{y\in\Gamma_h x} \frac{\partial}{\partial^{R_{e,k}}\tau_e \partial^{S_{e,k}}\overline{\tau_e}} f_e(y_k, \tau_e)$$

converge for any integers $R_{e,k}, S_{e,k} \geq 0$, as claimed in (19).

Step 2: Now we show the convergence of (19), in two parts.

(1) First we show the case of $\sum_{y\in\Gamma_h x} f_e(y_k, \tau_e)$. We have

$$f_e(y_k, \tau_e) \leq \frac{e^{-2\pi R(y_k,\tau)}}{R(y_k, \tau_e)} \leq e^{-2\pi R(y_k,\tau_e)}$$

for $R(y_k, \tau_e) \geq 1$, which happens for all except finitely many y_k's from Lemma 3.6. Furthermore, also from Lemma 3.6, since there are at most $O(z^{(n+2)/2})$ vectors y_k in our sum with $z \leq R(y_k, \tau_e) \leq z+1$, we are reduced to the convergence of

$$\sum_{z=1}^{\infty} e^{-2\pi z} z^{((n+2)/2)},$$

which converges using the integral test.

(2) Now we show the convergence of (19) for the partial derivatives in τ_e for the term $f_e(y_k, \tau_e)$. Note first that we can compute the derivatives:

$$\frac{\partial}{\partial\tau_e} f_e(y_k, \tau_e) = \frac{e^{-2\pi R(y_k,\tau_e)}}{2\pi R(y_k, \tau_e)} \frac{\partial}{\partial\tau_e} R(y_k, \tau_e),$$

$$\frac{\partial}{\partial\overline{\tau_e}} f_e(y_k, \tau_e) = \frac{e^{-2\pi R(y_k,\tau_e)}}{2\pi R(y_k, \tau_e)} \frac{\partial}{\partial\overline{\tau_e}} R(y_k, \tau_e).$$

In general terms we get

$$\frac{\partial}{\partial^R\tau_e \partial^S\overline{\tau_e}} f_e(y_k, \tau_e) = e^{-2\pi R(y_k,\tau_e)} \sum_i \frac{e^{-c_i R(y_k,\tau_e)}}{R(y_k, \tau_e)^{d_i}} P_i(\partial_{a_i,b_i} R),$$

where the above is a finite sum, $P_i(\partial R, y_k)$ are polynomials in

$$\frac{\partial}{\partial^{a_i}\tau_e \partial^{b_i}\overline{\tau_e}} R(y_k, \tau_e),$$

and the constants c_i, d_i are integers that satisfy $d_i \geq 1$, and $d_i > c_i \geq 0$. This can be easily shown by induction.

Excluding the terms for which $R(y_k, \tau_e) \leq 1$, note that if we fix a basis $(e_1, \ldots, e_{n+2})$ for V_{σ_e}, we have

$$\frac{\partial}{\partial^R \tau_e \partial^S \overline{\tau_e}} R(y_k, \tau_e) = -\sum_{j=1}^{n+2} (y_k^{(j),e})^2 \frac{\partial}{\partial^R \tau_e \partial^S \overline{\tau_e}} R(e_j, \tau_e),$$

thus we can further bound

$$\left| \frac{\partial}{\partial^a \tau_e \partial^b \overline{\tau_e}} R(y_k, \tau_e) \right| \leq M_{a,b}(q_e(x_k) + R(y_k, z_{0,e})),$$

where $M_{a,b}$ is the upper bound of the values

$$\frac{\partial}{\partial^a \tau_e \partial^b \overline{\tau_e}} R(e_j, \tau_e)$$

for $1 \leq j \leq n+2$ and τ_e in our compact.

As $d_i > c_i$, for $R(y_k, \tau_e) \geq 1$, we have

$$\frac{e^{-2\pi c_i R(y_k, \tau_e)}}{(2\pi R(y_k, \tau_e))^{d_i}} < 1$$

and using the above bound we have more generally

$$\left| \frac{\partial}{\partial^R \tau_e \partial^S \overline{\tau_e}} f_e(y_k, \tau_e) \right| \leq M e^{-2\pi R(y_k, \tau_e)} \widetilde{Q}(R(y_k, \tau_e)),$$

where $\widetilde{Q}$ is a polynomial in $R(y_k, z_{0,e})$. Let D be the degree of $\widetilde{Q}$ and let $\widetilde{Q_0}(x) := \sum |a_n| x^n$ if $Q := \sum a_n x^n$.

Similarly as before, we have at most $O(z^{(n+2)/2})$ values y_k such that $z \leq R(y_k, \tau_e) \leq z+1$ for τ_e inside a compact, and the above convergence is equivalent to the convergence of

$$\sum_{z=1}^{\infty} e^{-2\pi z} z^{(n+2)/2} \widetilde{Q_0}(z+1),$$

which converges by the integral test. □

Now we are also going to show:

Proposition 3.9. *For $x = (x_1, \ldots, x_k) \in V(F)^k$, the form*

$$\omega_3(x, \tau; g, h) = \sum_{\gamma \in G_x(F) \backslash G(F)} \omega_2(x, \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h)$$

converges.

Proof. Note that the above statement follows for $\dim U(x) = k$ from the proof of Proposition 3.2. For the general case the proof is similar to that of Lemma 3.8. Using the notation from Lemma 3.8, we can write

$$\omega_3(x, \tau; g, h) = \sum_{y \in \Gamma_h x} \varphi_1(y_1, \tau_1) \wedge \cdots \wedge \varphi_1(y_k, \tau_1) \wedge \cdots \wedge \varphi_1(y_1, \tau_e) \wedge \cdots \wedge \varphi_e(y_k, \tau_e).$$

Using the definition of $\varphi_i(y_j, \tau_i)$,

$$\varphi_i(y_j, \tau_i) = e^{-2\pi R(y_j, z_{0,i})}\Bigg(\sum_{1\le s,t\le n} y_j^{(s),i} y_j^{(t),i} p_i^*(\omega_{s,1i} \wedge \omega_{t,2i}) - \frac{1}{\pi} \sum_{1\le s\le n} p_i^*(\omega_{s,1i} \wedge \omega_{s,2i}) \Bigg),$$

the terms $p_i^*\omega_{s,1i} \wedge p_i^*\omega_{t,1i}$ are independent of y, and we are reduced to the convergence of the coefficients:

$$\sum_{y\in\Gamma_h x} e^{-2\pi \sum_{i=1}^{e}\sum_{j=1}^{k} R(y_j, z_{0,i})} P(y),$$

where

$$P(y) = \prod_{i=1}^{e}\prod_{j=1}^{k} \sum_{1\le s,t\le n} \sum_{f=0}^{1} (y_j^{(s),i} y_j^{(t),i})^f.$$

As in Lemma 3.8, we can bound

$$\sum_{1\le s,t\le n} \sum_{f=0}^{1} (y_j^{(s),i} y_j^{(t),i})^f \le (R(y_j, z_{0,i}) + q_i(x_j))^{n^2}.$$

Moreover, for $(i, j) \neq (e, k)$, by discarding finitely many terms from the lattice we have $R(y_k, \tau_e)$ large enough and we can bound $e^{-2\pi R(y_j,\tau_i)} R(y_j, z_{0,i})^m \le 1$, for any $1 \le m \le n^2$. Thus the convergence reduces to showing that

$$\sum_{y\in\Gamma_h x} e^{-2\pi R(y_k, z_{0,e})} (R(y_k, z_{0,e}) + q_e(x_k))^{n^2}$$

converges, or equivalently that any of the terms

$$\sum_{y\in\Gamma_h x} e^{-2\pi R(y_k, z_{0,e})} R(y_k, z_{0,e})^m,$$

converge for $1 \le m \le n^2$. Again we have at most $O(z^{(n+2)/2})$ values y_k such that $z \le R(y_k, \tau_e) \le z + 1$ for τ_e inside a compact, thus the above reduces to the convergence of

$$\sum_{y\in\Gamma_h x} e^{-2\pi z} (z + 1)^m z^{(n+2)/2},$$

which converges by the integral test. This finishes our proof. □

4. Modularity of $Z(g', \phi)$

We recall now the definition of the standard Whittaker function. Recall from Section 3C that we defined $\widetilde{\mathrm{Sp}}_{2r}(\mathbb{R})$ to be the metaplectic cover of $\mathrm{Sp}_{2r}(\mathbb{R})$, and K' the preimage under the projection map $\widetilde{\mathrm{Sp}}_{2r}(\mathbb{R}) \to \mathrm{Sp}_{2r}(\mathbb{R})$ of the compact subgroup $\left\{\left(\begin{smallmatrix} A & B \\ -B & A \end{smallmatrix}\right), A + iB \in U(r)\right\}$, where $U(r)$ is the unitary group. We also defined the character $\det^{1/2}$ on K' whose square descends to the determinant character of $U(r)$.

For (V_+, q_+) a quadratic space over $\mathbb{R}$ of signature $(n+2, 0)$, let $\varphi_+^\circ(x_+) \in S(V_+^r)$ be the standard Gaussian,

$$\varphi_+^\circ(x_+) = e^{-\pi \operatorname{tr}(x,x)_+},$$

where $\frac{1}{2}(x, x)_+ = \frac{1}{2}((x_i, x_j))_{1\le i,j\le r}$ is the intersection matrix of $x = (x_1, \dots, x_r) \in V_+^r$ for the inner product $(\cdot, \cdot)$ given by q_+ on V_+.

Then for $x \in V_+^r$ and $\beta = \frac{1}{2}(x, x)_+$ with β in $\mathrm{Sym}_r(\mathbb{R})$, the group of symmetric $r \times r$ matrices, we define the β-th "holomorphic" Whittaker function

$$W_\beta(g) = r(g)\varphi_+^\circ(x),$$

where $g \in \widetilde{\mathrm{Sp}}_{2r}(\mathbb{R})$ and r is the Weil representation of $\widetilde{\mathrm{Sp}}_{2r}(\mathbb{R}) \times O(V^r)$.

Using the Iwasawa decomposition of $\widetilde{\mathrm{Sp}}_{2r}(\mathbb{R})$, we can write each g in the form

$$g = \left(\begin{smallmatrix} 1 & u \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} v & 0 \\ 0 & (v^T)^{-1} \end{smallmatrix}\right)k', \quad v \in \mathrm{GL}_r(\mathbb{R})^+, k' \in K',$$

and we have

$$W_\beta(g) = \det(v)^{(n+2)/4} e^{2\pi i \operatorname{tr}\beta\tau} \det(k')^{(n+2)/2},$$

where $\tau = u + (v \cdot v^T)\sqrt{-1}$ is an element of $\mathcal{H}_r$, the Siegel upper half-space of genus r (see [Yuan et al. 2009] for a reference).

We can extend this definition for F_∞. For $g' = (g'_j)_{1\le j\le d} \in \widetilde{\mathrm{Sp}}_{2r}(F_\infty) = \prod_{\sigma_j : F\hookrightarrow\mathbb{R}} \widetilde{\mathrm{Sp}}_{2r}(\mathbb{R}_{\sigma_j})$, we take

$$W_\beta(g'_\infty) = \prod_{\sigma_j : F\hookrightarrow\mathbb{R}} W_{\sigma_j(\beta)}(g'_j).$$

Moreover, by writing each $g'_j = \left(\begin{smallmatrix} 1 & u_j \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} v_j & 0 \\ 0 & (v_j^T)^{-1} \end{smallmatrix}\right)k'_j$ using the Iwasawa decomposition and taking $\tau_j = u_j + i(v_j \cdot v_j^T)$ as above, we get

$$W_\beta(g'_\infty) = \prod_{\sigma_j : F\hookrightarrow\mathbb{R}} \det(v_j)^{(n+2)/2} e^{2\pi i \operatorname{tr}\sigma_j(\beta)\tau_j} \det(k'_j)^{(n+2)/2}.$$

Recall from the Introduction that we defined $T(x) = \frac{1}{2}(\langle x_i, x_j\rangle)_{1\le i,j\le r}$ to be the intersection matrix in $M_r(F)$. Note that for $1 \le i \le e$ the intersection matrix $T(x)$ is different from the intersection matrix $\frac{1}{2}(x, x)_+$ above, for which the inner product $(\,\cdot\,,\cdot\,)$ is positive-definite.

We extend the definition of W_β to $\sigma_j(\beta) \notin \mathrm{Sym}_r(\mathbb{R})$ for some σ_j, $1 \le j \le e$, by taking $W_\beta(g'_\infty) = 0$.

For $g' \in \widetilde{\mathrm{Sp}}_{2r}(\mathbb{A})$, $\phi \in (\mathcal{S}(V_{\mathbb{A}}^r))^K$, we defined in the introduction Kudla's generating series

$$Z(g', \phi) = \sum_{x \in G(F)\backslash V(F)^r} \sum_{g \in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K} r(g'_f)\phi_f(g^{-1}x) W_{T(x)}(g'_\infty) Z(x, g)_K. \tag{20}$$

We will show:

Theorem 4.1. *The function $Z(g', \phi)$ is an automorphic form parallel of weight $1 + n/2$ for $g' \in \widetilde{\mathrm{Sp}}_{2r}(\mathbb{A})$, $\phi \in \mathcal{S}(V_{\mathbb{A}}^r)$ with values in $H^{2er}(M_K, \mathbb{C})$.*

Recall that in $H^{2er}(M_K, \mathbb{C})$ we have $[Z(x, g)] = [\omega_4(x', \tau; g, h) \wedge ((-1)^e c_1(L_K^\vee))^{r-k}]$ as cohomology classes, where $c_1(L_K^\vee) = c_1(L_{K,1}^\vee) \cdots c_1(L_{K,e}^\vee)$. We are actually going to show in Section 4A that $[Z(x, g)] = [\omega_4(x, \tau; g, h)]$ and we will replace in the sum (20) the cohomology class of the special cycle $Z(x, g)$ with the cohomology class of $\omega_4(x, \tau; g, h)$. We are going to show first the following expansion of the pullback of $[Z(g', \phi)]$ to $D \times G(\mathbb{A}_f)/K$:

Lemma 4.2. *The pullback of the cohomology class $[Z(g', \phi)]$ to $D \times G(\mathbb{A}_f)/K$ is the cohomology class*

$$p^*[Z(g', \phi)] = \sum_{x \in V(F)^r} r(g')\phi_f(h^{-1}x) W_{T(x)}(g'_\infty)\omega_2(vx, \tau),$$

where $p : D \times G(\mathbb{A}_f)/K \to M_K$ is the natural projection map and $g'_i = \left(\begin{smallmatrix} 1 & u_i \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} v_i & 0 \\ 0 & (v_i^T)^{-1} \end{smallmatrix}\right)k'_i$ is the Iwasawa decomposition of $g'_i = \sigma_i(g')$ for $1 \le i \le d$.

We claim that this will imply Theorem 4.1. We will first discuss the pullback of cohomology classes in Section 4A and we will show Lemma 4.2 and Theorem 4.1 at the end of the section.

4A. ***Cohomology classes.*** First we would like to understand better how we take the pullback of the cohomology classes $[\omega_3(x, \tau; g, h)]$ to $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$.

Note that for $x \in V(F)^r$ with $U(x)$ a totally positive k-subspace of V, and $g \in G(\mathbb{A}_f)$, we have the equality of cohomology classes $[Z(U(x), g)] = [\omega_4(x', g)]$ in $H^{2ek}(M_K, \mathbb{C})$ and we can take the pullback $[\omega_3(x', g)]$ to $H^{2ek}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$. The pullback of $(-1)^e c_1(L_K^\vee)$ to $H^2(D \times G(\mathbb{A}_f)/K, \mathbb{C})$ is $\omega_3(0, \tau)$.

We are actually going to show that the pullbacks of the Kudla cycles $Z(U(x), g)c_1(L_K^\vee)^{r-k}$ can be represented by the cohomology class of $[\omega_3(x, g)]$ in $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$ in the lemma below:

Lemma 4.3. *In $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$ we have the equality of cohomology classes*:

$$[\omega_3(x') \wedge \omega_3(0)^{(r-k)}] = [\omega_3(x)].$$

To show this, we first recall from [Kudla 1997a, Lemma 7.3] how the pullback acts on the Kudla–Millson form $\varphi_{KM}^{(k)}$. For $1 \le i \le e$, recall that (V_{σ_i}, q_i) is a quadratic space of signature $(n, 2)$.

Lemma 4.4. *Let $U \subset V_{\sigma_i}$ be a positive k-plane. For $y \in U$, let $\varphi_+^\circ \in \mathcal{S}(U^k)$ be the standard Gaussian $\varphi_+^\circ(y) = e^{-\pi q_i(y)}$. Let $\iota_U : D_{U,i} \to D_i$ be the natural injection. Under the pullback $\iota_U^* : \Omega^k(D_i) \to \Omega^k(D_{U,i})$ of differential forms, we then have*

$$\iota_U^* \varphi_{KM}^{(k),\circ} = \varphi_+^\circ \otimes \varphi_{KM,V_U}^{(k),\circ},$$

where $\varphi_{KM,V_U}{}^{(k),\circ} \in (\mathcal{S}(U^k) \otimes \Omega^{k,k}(D_{U,i}))^K$ is the Kudla–Millson form for the vector space $V_{i,U} = \langle U \rangle^\perp$ and Hermitian symmetric domain $D_{U,i}$.

For $x \in V(F)^r$ such that $U(x)$ is a totally positive k-subspace of V we defined $x' = (x_{i_1}, \ldots, x_{i_k})$. Let $x'' = (x_{j_1}, \ldots, x_{j_{r-k}})$ consist of the remaining components of x.

Just for this section, we will use the notation $\omega_i^{(m)}(x, \tau)$ for $i = 2, 3$ when $x = (x_1, \ldots, x_m) \in V^m$. Using the above lemma, we are going to show:

Lemma 4.5. *With the above notation, the pullback of* $\omega_3^{(r-k)}(x'', \tau; g, h)$ *to* $D_U \times G_U(\mathbb{A}_f)gK/K$ *via the inclusion map* $\iota : D_U \times G_U(\mathbb{A}_f)gK/K \to D \times G(\mathbb{A}_f)/K$ *equals*

$$\iota^*\omega_3^{(r-k)}(x'', \tau; g, h) = \iota^*\omega_3^{(r-k)}(0, \tau; g, h). \tag{21}$$

Proof. From the definition of $\varphi_{KM}^{(r),\circ}$ we can write

$$\varphi_{KM}^{(r),\circ}(x) = \varphi_{KM}^{(k),\circ}(x') \wedge \varphi_{KM}^{(r-k),\circ}(x''). \tag{22}$$

Then from Lemma 4.4, for $\iota_U : D_{U,i} \to D_i$ the natural embedding, we have

$$i_U^*\varphi_{KM}^{(r-k),\circ}(x'') = (\varphi_+^\circ \otimes \varphi_{KM,V_{U,i}}^{(r-k),\circ})(x'') = \varphi_+^\circ(x'')\varphi_{KM,V_{U,i}}^{(r-k)}(0),$$

as $x'' \in U^{r-k}$. Note that this implies

$$i_U^*\varphi_{KM}^{(r-k)}(x'') = \varphi_{KM,V_{U,i}}^{(r-k)}(0). \tag{23}$$

We first want to pullback everything to D, via the projection maps $p_i : D \to D_i$. We have the maps $\iota_U : D_U \hookrightarrow D$, $p_i : D \to D_i$. Recall that

$$D_U = D_{U,1} \times \cdots \times D_{U,e},$$

and we can further define the embedding $\iota_{U,i} : D_{U,i} \hookrightarrow D_i$ and the projection map $p_{U,i} : D_U \to D_{U,i}$. It is easy to see that $\iota_{U,i} \circ p_{U,i} = p_i \circ \iota_U$ as maps from D_U to D_i, thus we also have the equality of pullbacks of differentials $\Omega^{r-k}(D_i) \to \Omega^{r-k}(D_U)$:

$$p_{U,i}^* \circ \iota_{U,i}^* = \iota_U^* \circ p_i^*.$$

Then we get the equality

$$\iota_U^* p_i^* \varphi_{KM}^{(r-k)}(x'', \tau_i) = p_{U,i}^* \circ \iota_{U,i}^* \varphi_{KM}^{(r-k)}(x'', \tau_i).$$

From (23), the right-hand side equals $p_{U,i}^*\varphi_{KM,V_{U,i}}^{(r-k)}(0, \tau_i)$. Applying the same steps also for $\varphi_{KM}^{(r-k)}(0)$, we get

$$\iota_U^* p_i^*(\varphi_{KM}^{(r-k)}(0, \tau_i)) = p_{U,i}^* \circ \iota_{U,i}^*(\varphi_{KM}^{(r-k)}(0, \tau_i)) = p_{U,i}^*(\varphi_{KM,V_{U,i}}^{(r-k)}(0, \tau_i)).$$

Thus we have

$$\iota_U^* p_i^* \varphi_{KM}^{(r-k)}(x, \tau_i) = \iota_U^* p_i^*(\varphi_{KM}^{(r-k)}(0, \tau_i)). \tag{24}$$

Note that we can further take the wedge product of $\iota_U^* p_i^* \varphi_{KM}^{(r-k)}(x, \tau_i)$ for $1 \le i \le e$ to get

$$\iota_U^*\omega_2^{(r-k)}(x'') = \iota_U^* \bigwedge_{i=1}^e p_i^* \varphi^{(r-k)}(x, \tau_i) = \bigwedge_{i=1}^e \iota_U^* p_i^* \varphi^{(r-k)}(x, \tau_i),$$

and using (24) this gives us $\iota_U^*(\omega_2^{(r-k)}(0, \tau))$. Note that this implies

$$\iota_U^*\omega_2^{(r-k)}(x'') = \iota_U^*(\omega_2^{(r-k)}(0, \tau)) \tag{25}$$

Finally, we are interested in the pullback of $\omega_3^{(r-k)}(x'', \tau; g, h)$ to $D_U \times G_U(\mathbb{A}_f)gK/K$ via the inclusion map $\iota : D_U \times G_U(\mathbb{A}_f)gK/K \to D \times G(\mathbb{A}_f)/K$. We have

$$\iota^*\omega_3^{(r-k)}(x'', \tau; g, h) = \sum_{\gamma \in G_U(F)\backslash G(F)} \iota_U^*\omega_2^{(r-k)}(x'', \gamma\tau)1_{G_U(\mathbb{A}_f)gK}(\gamma h),$$

and using the pullback above for the right-hand side we get

$$\sum_{\gamma \in G_U(F)\backslash G(F)} \iota_U^*\omega_2^{(r-k)}(0, \gamma\tau)1_{G_U(\mathbb{A}_f)gK}(\gamma h),$$

which equals $\iota^*\omega_3^{(r-k)}(0, \tau; g, h)$. Thus we have $\iota^*\omega_3^{(r-k)}(x'', \tau; g, h) = \iota^*\omega_3^{(r-k)}(0, \tau; g, h)$, which is the result of the lemma. □

Note that using (23) and (14) one can actually show that

$$[\varphi_{KM}^{(r)}(x)] = [\varphi_{KM}^{(k)}(x') \wedge \varphi_{KM}^{(r-k)}(0)]$$

as cohomology classes in $H^{2r}(D_i, \mathbb{C})$.

Moreover, using (25) and (16), one can further show that

$$[\omega_2^{(r)}(x)] = [\omega_2^{(k)}(x') \wedge \omega_2^{(r-k)}(0)]$$

as cohomology classes in $H^{2r}(D, \mathbb{C})$.

The proof of Lemma 4.3 below is based on the same principle.

Proof of Lemma 4.3. To show the equality of cohomology classes, we need to show that for a closed $(l-r, l-r)$-form μ with compact support, where l is the complex dimension of $D \times G(\mathbb{A}_f)/K$, we have

$$\int_{D\times G(\mathbb{A}_f)/K} \mu \wedge \omega_3^{(r)}(x) = \int_{D\times G(\mathbb{A}_f)/K} \mu \wedge \omega_3^{(k)}(x'') \wedge \omega_3^{(r-k)}(0). \tag{26}$$

From (5), for a closed form μ, as $\mu \wedge \omega_3^{(r-k)}$ is a closed $(l-k, l-k)$-form we have

$$\int_{D\times G(\mathbb{A}_f)/K} \mu \wedge \omega_3^{(r)}(x) = \int_{D_U\times G_U(\mathbb{A}_f)gK/K} \iota^*(\mu \wedge \omega_3^{(r-k)}(x'')).$$

From (21), we have $\iota^*(\mu \wedge \omega_3^{(r-k)}(x'')) = \iota^*(\mu \wedge \omega_3^{(r-k)}(0))$, thus we get

$$\int_{D\times G(\mathbb{A}_f)/K} \mu \wedge \omega_3^{(r)}(x) = \int_{D_U\times G_U(\mathbb{A}_f)gK/K} \iota^*(\mu \wedge \omega_3^{(r-k)}(0)). \tag{27}$$

Using (5) for $\mu \wedge \omega_3^{(r-k)}(0)$ we also get

$$\int_{D\times G(\mathbb{A}_f)/K} \mu \wedge \omega_3^{(k)}(x') \wedge \omega_3^{(r-k)}(0) = \int_{D_U\times G_U(\mathbb{A}_f)gK/K} \iota^*(\mu \wedge \omega_3^{(r-k)}(0)). \tag{28}$$

Combining the two equations (27) and (28) we get (26). □

Remarks on $\omega_3(vx)$ and $\omega_4(vx)$. We follow up with some remarks regarding $\omega_3(vx, \tau; g, h)$ and $\omega_4(vx, \tau; g, h)$ when $v \in \mathrm{GL}_r(F_\infty)$ and $x \in V(F)^r$ with $U(x)$ totally positive definite k-subspace of $V(F)$. We have defined them in Section 3E. Lemma 4.3 extends easily for $\omega_3(vx, \tau; g, h)$ and $\omega_4(vx, \tau; g, h)$ and we have, as cohomology classes in $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$,

$$[\omega_3(vx, \tau; g, h)] = [\omega_3((vx)', \tau; g, h) \wedge \omega_3^{(r-k)}(0, \tau)].$$

As actually $\omega_3((vx)')$ represents the same cohomology class as the preimages of $Z(U(vx), g)$ in $D \times G(\mathbb{A}_f)/K$, and as $Z(U(x), g) = Z(U(vx), g)$, we have:

Lemma 4.6. (i) *As cohomology classes in $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$, we have*

$$[\omega_3(vx, \tau; g, h)] = [\omega_3(x, \tau; g, h)]. \tag{29}$$

(ii) *Noting that* (29) *descends to M_K, we also have, as cohomology classes in $H^{2er}(M_K, \mathbb{C})$,*

$$[\omega_4(vx, \tau; g, h)] = [\omega_4(x, \tau; g, h)]. \tag{30}$$

Proof of modularity: We will finish below the proofs of Lemma 4.2 and Theorem 4.1.

Proof of Lemma 4.2. The pullback to $D \times G(\mathbb{A}_f)/K$ of $\omega_4(x', \tau)$ is $\omega_3(x', \tau)$ and $\omega_3(0, \tau)$ is the pullback of $(-1)^{er} c_1^r(L_K^\vee) = Z(0, g)$. Then in (20) we can write

$$p^*[Z(g', \phi)] = \sum_{x \in G(F)\backslash V(F)^r} \sum_{g \in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K} r(g', g)\phi_f(x) W_{T(x)}(g'_\infty)[\omega_3(x', \tau; g, h) \wedge \omega_3^{(r-k)}(0)].$$

Furthermore, from Lemma 4.3 we have $[\omega_3(x', g; \tau, h) \wedge \omega_3^{(r-k)}(0, \tau)] = [\omega_3(x, \tau; g, h)]$ as classes in $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$. From (29) we also have the equality of cohomology classes $[\omega_3(x, \tau; g, h)] = [\omega_3(vx, \tau; g, h)]$. Thus we get

$$p^*[Z(g', \phi)] = \sum_{x \in G(F)\backslash V(F)^r} \sum_{g \in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K} r(g', g)\phi_f(x) W_{T(x)}(g'_\infty)[\omega_3(vx, g; \tau, h)].$$

By plugging in the definition

$$\omega_3(vx, \tau; g, h) = \sum_{\gamma \in G_x(F)\backslash G(F)} \omega_2(vx, \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h),$$

we get the cohomology class $p^*[Z(g', \phi)]$ equal to the cohomology class of

$$\sum_{x \in G(F)\backslash V(F)^r} \sum_{g \in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K} r(g'_f, g)\phi_f(x) W_{T(x)}(g'_\infty) \sum_{\gamma \in G_x(F)\backslash G(F)} \omega_2(vx, \gamma\tau) 1_{G_x(\mathbb{A}_f) g K h^{-1}}(\gamma).$$

We will unwind the sum below to get the result of the lemma. We interchange the summations to get

$$\sum_{x \in G(F)\backslash V(F)^r} \sum_{\gamma \in G_x(F)\backslash G(F)} \sum_{g \in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K} r(g'_f, g)\phi_f(x) W_{T(x)}(g'_\infty)\omega_2(vx, \gamma\tau) 1_{G_x(\mathbb{A}_f) g K}(\gamma h).$$

Note that $1_{G_x(\mathbb{A}_f)gK}(\gamma h) \neq 0$ if and only if $\gamma h \in G_x(\mathbb{A}_f)gK$, or equivalently if $g \in G_x(\mathbb{A}_f)\gamma hK$, and since we are summing for $g \in G_x(\mathbb{A}_f) \setminus G(\mathbb{A}_f)/K$, we can replace g by γh everywhere and get

$$p^*[Z(g', \phi)] = \sum_{x \in G(F)\setminus V(F)^r} \sum_{\gamma \in G_x(F)\setminus G(F)} r(g'_f, \gamma h)\phi_f(x) W_{T(x)}(g'_\infty)\omega_2(vx, \gamma\tau).$$

Since the action of $G(\mathbb{A}_f)$ on ϕ is given by $r(g'_f, \gamma h)\phi_f(x) = r(g'_f)\phi_f(h^{-1}\gamma^{-1}x)$ and $\omega_2(vx, \gamma\tau) = \omega_2(\gamma^{-1}vx, \tau) = \omega_2(v(\gamma^{-1}x), \tau)$, then we have

$$p^*[Z(g', \phi)] = \sum_{x \in V(F)^r} r(g'_f)\phi_f(h^{-1}x) W_{T(x)}(g'_\infty)\omega_2(vx, \tau),$$

which gives us the result of the lemma. □

Proof of Theorem 4.1. We would like to rewrite the sum of Lemma 4.2,

$$p^*[Z(g', \phi)] = \sum_{x \in V(F)^r} r(g')\phi_f(h^{-1}x) W_{T(x)}(g'_\infty)\omega_2(vx, \tau),$$

and show that this sum is automorphic with values in $H^{2er}(D \times G(\mathbb{A}_f)/K, \mathbb{C})$.

We recall the Iwasawa decomposition of $g' = (g'_i)_{1\le i\le d} \in \widetilde{\mathrm{Sp}}_{2r}(F_\infty)$ to be $g'_i = \left(\begin{smallmatrix} 1 & u_i \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} v_i & 0 \\ 0 & (v_i^T)^{-1} \end{smallmatrix}\right)k'_i$, where $v_i \in \mathrm{GL}_r(\mathbb{R}_{\sigma_i})^+$, $k'_i \in K'_i$.

Recall that, for $1 \le i \le e$, we have $\omega_1(x, \tau_i) = \varphi^{(r)}_{KM}(x, \tau_i)$ and $\omega_2(x, \tau) = p_1^*\omega_1(x, \tau_1) \wedge \cdots \wedge p_e^*\omega_1(x, \tau_e)$. From property (1) of the theorem of Kudla and Millson we presented in Section 3C, we have

$$r(k'_i)\varphi^{(r),\circ}_{KM} = \det(k'_i)^{(n+2)/2}\varphi^{(r),\circ}_{KM},$$

where $\varphi^{(r),\circ}_{KM}(x, \tau_i) = e^{-2\pi \operatorname{tr}\sigma_i(T(x))}\varphi_{KM}(x, \tau_i)$. Using the Weil representation this easily extends to

$$r(g'_i)\varphi^{(r),\circ}_{KM}(x, \tau_i) = \det(v_i)^{(n+2)/2}\det(k'_i)^{(n+2)/2}e^{-2\pi \operatorname{tr} T(\sigma_i(x))(u_i + iv_i\cdot v_i^T)}\varphi^{(r)}_{KM}(v_i x, \tau_i).$$

We take the pullback to D via the projection maps $p_i : D \to D_i$. We denote $\varphi_i(x, \tau_i) = p_i^*\varphi^{(r)}_{KM}(x, \tau_i)$ and $\varphi^\circ_i(x, \tau_i) = e^{-2\pi \operatorname{tr}\sigma_i(T(x))}\varphi_i(x, \tau_i)$ and thus we also have

$$r(g'_i)(\varphi^\circ_i(x, \tau_i)) = \det(v_i)^{(n+2)/2}\det(k'_i)^{(n+2)/2}e^{-2\pi \operatorname{tr} T(\sigma_i(x))(u_i + iv_i\cdot v_i^T)}\varphi_i(v_i x, \tau_i).$$

Note that on the right-hand side we got $W_{\sigma_i(T(x))}(g'_i)\varphi_i(v_i x, \tau_i)$, thus

$$r(g'_i)(\varphi^\circ_i(x, \tau_i)) = W_{\sigma_i(T(x))}(g'_i)\varphi_i(v_i x, \tau_i).$$

Furthermore, as we can rewrite

$$\begin{aligned} &W_{T(x)}(g'_\infty)\varphi_1(v_1x, \tau_1) \wedge \cdots \wedge \varphi_e(v_e x, \tau_e) \\ &\qquad = (W_{\sigma_1(T(x))}(g'_1)\varphi_1(v_1x, \tau_1) \wedge \cdots \wedge W_{\sigma_e(T(x))}(g'_e)\varphi_e(v_e x, \tau_e)) \prod_{i=e+1}^{d} W_{\sigma_i(T(x))}(g'_i), \end{aligned}$$

we get

$$W_{T(x)}(g'_\infty)\varphi_1(v_1x, \tau_1) \wedge \cdots \wedge \varphi_e(v_e x, \tau_e) = r(g'_\infty)\phi^\circ(x, \tau),$$

where

$$\phi^\circ(x,\tau)=\varphi_1^\circ(x,\tau_1)\wedge\cdots\wedge\varphi_e^\circ(x,\tau_e)\prod_{i=e+1}^{d}\varphi_{0,i}(x).$$

Recall that for $i\geq e+1$, we have $W_{T(\sigma_i(x))}(g_i)=r(g_i)\varphi_{0,i}(x)$. Here $\varphi_{0,i}(x)=e^{-\pi\,\mathrm{tr}\,T(\sigma_i(x))}$ is the standard Gaussian, as (V_{σ_i},q_i) is positive definite for $i\geq e+1$.

Going back to the sum of Lemma 4.2, we thus get

$$p^*[Z(g',\phi)]=\sum_{x\in V(F)^r}r(g'_f)\phi_f(h^{-1}x)r(g'_\infty)\phi^\circ(x,\tau),$$

and this is a theta function of weight $(n+2)/2$ with values in the cohomology group $H^{2er}(D\times G(\mathbb{A}_f)/K,\mathbb{C})$. This means that for any linear functional $l:H^{2er}(D\times G(\mathbb{A}_f)/K,\mathbb{C})\to\mathbb{C}$ acting on the cohomology part of $\phi^\circ(x,\tau)$, the generating series

$$l(p^*[Z(g',\phi)])=\sum_{x\in V(F)^r}r(g'_f)\phi_f(h^{-1}x)r(g'_\infty)l(\phi^\circ(x,\tau))$$

is a theta function of weight $(n+2)/2$. Note that this series is obtained by unwinding

$$p^*[Z(g',\phi)]=\sum_{x\in G(F)\backslash V(F)^r}\ \sum_{g\in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K}r(g',g)\phi_f(x)W_{T(x)}(g'_\infty)l(\omega_3(x,g)).$$

Denote

$$Z_0(g',\phi)=\sum_{x\in G(F)\backslash V(F)^r}\ \sum_{g\in G_x(\mathbb{A}_f)\backslash G(\mathbb{A}_f)/K}r(g',g)\phi_f(x)W_{T(x)}(g'_\infty)\omega_3(x,g).$$

For the natural projection $p:D\times G(\mathbb{A}_f)/K\to M_K$, recall the pullback

$$p^*:\Omega^{2er}(M_K)\to\Omega^{2er}(D\times G(\mathbb{A}_f)/K),$$

which further descends to the cohomology groups $p^*:H^{2er}_{dR}(M_K)\to H^{2er}_{dR}(D\times G(\mathbb{A}_f)/K)$ and the map is an injection.

We denote by $\mathrm{SC}^{2er}(M_K)$ the subspace of $H^{2er}_{dR}(M_K)$ generated by the classes $[\omega_4(x,g)]$ and by $\mathrm{SC}^{2er}(D\times G(\mathbb{A}_f)/K)$ the subspace of $H^{2er}_{dR}(M_K)$ generated by the classes $[\omega_3(x,g)]$. Then the above pullback map restricts to $p^*:\mathrm{SC}^{2er}(M_K)\to\mathrm{SC}^{2er}(D\times G(\mathbb{A}_f)/K)$ and it is an injection.

Then for any linear functional l of $\mathrm{SC}^{2er}(M_K)$, we are able to just define the linear functional $\tilde{l}$ on $\mathrm{SC}^{2er}(D\times G(\mathbb{A}_f)/K)$ given by $\tilde{l}(p^*[\omega])=\tilde{l}([\omega])$, and thus $\tilde{l}(Z_0(g',\phi))=l([Z(g',\phi)])$ is automorphic. Thus $[Z(g',\phi)]$ is a theta function valued in $H^{2er}(M_K)$.

We can also easily check the weight of the theta function by computing

$$r(k')\phi^\circ(x,\tau)=r(k'_1)\varphi_1^\circ(x,\tau_1)\wedge\cdots\wedge r(k'_e)\varphi_e^\circ(x,\tau_e)\prod_{i=e+1}^{d}r(k'_i)\phi_{0,i}(x),$$

which gives us the factor $\det(k'_i)^{(n+2)/2}$ at each place i. □

Acknowledgements

The authors would like to thank Xinyi Yuan for suggesting the problem and for very helpful discussions and insights regarding the problem. We would also like to thank the anonymous reviewer for detailed feedback. Rosu would also like to thank Max Planck Institute in Bonn for their hospitality, as well as to the IAS of Tsinghua University where part of the paper was written.

References

[Borcherds 1999] R. E. Borcherds, "The Gross–Kohnen–Zagier theorem in higher dimensions", *Duke Math. J.* **97**:2 (1999), 219–233. MR Zbl

[Chriss and Ginzburg 1997] N. Chriss and V. Ginzburg, *Representation theory and complex geometry*, Birkhäuser, Boston, 1997. MR Zbl

[Gillet and Soulé 1990] H. Gillet and C. Soulé, "Arithmetic intersection theory", *Inst. Hautes Études Sci. Publ. Math.* **72** (1990), 93–174. MR Zbl

[Griffiths and Harris 1978] P. Griffiths and J. Harris, *Principles of algebraic geometry*, Wiley, New York, 1978. 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

[Hirzebruch and Zagier 1976] F. Hirzebruch and D. Zagier, "Intersection numbers of curves on Hilbert modular surfaces and modular forms of Nebentypus", *Invent. Math.* **36** (1976), 57–113. MR Zbl

[Kudla 1997a] S. S. Kudla, "Algebraic cycles on Shimura varieties of orthogonal type", *Duke Math. J.* **86**:1 (1997), 39–78. MR Zbl

[Kudla 1997b] S. S. Kudla, "Central derivatives of Eisenstein series and height pairings", *Ann. of Math.* (2) **146**:3 (1997), 545–646. MR Zbl

[Kudla 2003] S. S. Kudla, "Integrals of Borcherds forms", *Compositio Math.* **137**:3 (2003), 293–349. MR Zbl

[Kudla and Millson 1986] S. S. Kudla and J. J. Millson, "The theta correspondence and harmonic forms, I", *Math. Ann.* **274**:3 (1986), 353–378. MR Zbl

[Kudla and Millson 1987] S. S. Kudla and J. J. Millson, "The theta correspondence and harmonic forms, II", *Math. Ann.* **277**:2 (1987), 267–314. MR Zbl

[Kudla and Millson 1990] S. S. Kudla and J. J. Millson, "Intersection numbers of cycles on locally symmetric spaces and Fourier coefficients of holomorphic modular forms in several complex variables", *Inst. Hautes Études Sci. Publ. Math.* **71** (1990), 121–172. MR Zbl

[de Rham 1955] G. de Rham, *Variétés différentiables: formes, courants, formes harmoniques*, Actualités Sci. Ind. **1222**, Hermann, Paris, 1955. MR Zbl

[Shih 1978] K.-y. Shih, "Existence of certain canonical models", *Duke Math. J.* **45**:1 (1978), 63–66. MR Zbl

[Soulé 1992] C. Soulé, *Lectures on Arakelov geometry*, Cambridge Stud. Adv. Math. **33**, Cambridge Univ. Press, 1992. MR Zbl

[Yuan et al. 2009] X. Yuan, S.-W. Zhang, and W. Zhang, "The Gross–Kohnen–Zagier theorem over totally real fields", *Compos. Math.* **145**:5 (2009), 1147–1162. MR Zbl

Communicated by Shou-Wu Zhang
Received 2019-09-12 Revised 2020-04-15 Accepted 2020-05-26

rosu@math.arizona.edu *Max Planck Institute for Mathematics, Bonn, Germany*
Department of Mathematics, University of Arizona, Tucson, AZ, United States
dyott@math.berkeley.edu *Department of Mathematics, UC Berkeley, Berkeley, CA, United States*

ALGEBRA AND NUMBER THEORY 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2773

Relative crystalline representations and p-divisible groups in the small ramification case

Tong Liu and Yong Suk Moon

Let k be a perfect field of characteristic $p > 2$, and let K be a finite totally ramified extension over $W(k)\big[\frac{1}{p}\big]$ of ramification degree e. Let R_0 be a relative base ring over $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$ satisfying some mild conditions, and let $R = R_0 \otimes_{W(k)} \mathcal{O}_K$. We show that if $e < p-1$, then every crystalline representation of $\pi_1^{\text{ét}}\big(\operatorname{Spec} R\big[\frac{1}{p}\big]\big)$ with Hodge–Tate weights in $[0, 1]$ arises from a p-divisible group over R.

1. Introduction

Let k be a perfect field of characteristic $p > 2$, and let $W(k)$ be its ring of Witt vectors. Let K be a finite totally ramified extension over $W(k)\big[\frac{1}{p}\big]$ with ramification degree e, and denote by $\mathcal{O}_K$ its ring of integers. If G is a p-divisible group over $\mathcal{O}_K$, then it is well-known that its Tate module $T_p(G)$ is a crystalline $\operatorname{Gal}(\overline{K}/K)$-representation with Hodge–Tate weights in $[0, 1]$. Conversely, Kisin [2006] showed the following result.

Theorem 1.1 [Kisin 2006, Corollary 2.2.6]. *Let T be a crystalline $\operatorname{Gal}(\overline{K}/K)$-representation finite free over $\mathbb{Z}_p$ whose Hodge–Tate weights lie in $[0, 1]$. Then there exists a p-divisible group G over $\mathcal{O}_K$ such that $T_p(G) \cong T$ as $\operatorname{Gal}(\overline{K}/K)$-representations.*

The result in Theorem 1.1 for the case $e \le p-1$ was first proved in [Laffaille 1980], in which the low ramification assumption is used to directly associate certain modules equipped with filtration and Frobenius endomorphism to p-divisible groups. This was one of the starting points of p-adic Hodge theory, to classify crystalline representations by weakly admissible filtered φ-modules and establish their connections to algebraic geometric objects.

The goal of this paper is to study the statement analogous to Theorem 1.1 in the relative case. When we work over a relative base ring, the situation becomes much more complicated, and it is unknown how to characterize crystalline representations by linear algebraic data. For example, [Hartl 2013] shows that a naive generalization of weakly admissible modules is not sufficient. In this paper, we obtain a partial result towards this direction for crystalline representations of Hodge–Tate weights in $[0, 1]$.

Let R_0 be a base ring over $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$ given as in Section 2A, and let $R = R_0 \otimes_{W(k)} \mathcal{O}_K$. Let $\mathcal{G}_R$ be the étale fundamental group of $\operatorname{Spec}\big(R\big[\frac{1}{p}\big]\big)$. For representations of $\mathcal{G}_R$, the condition of being *crystalline* is well-defined by [Brinon 2008; Kim 2015]. If G_R is a p-divisible group over R, its

MSC2010: primary 11F80; secondary 11S20, 14L05.
Keywords: crystalline representation, p-divisible group, relative p-adic Hodge theory.

Tate module $T_p(G_R)$ is a crystalline $\mathcal{G}_R$-representation with Hodge–Tate weights in $[0, 1]$ (see [Kim 2015]). Conversely, when the ramification index e is small, we prove that crystalline representations of Hodge–Tate weights in $[0, 1]$ can be associated with the linear algebraic data called Kisin modules of height 1, and show the following:

Theorem 1.2. *Suppose $e < p - 1$. Let T be a crystalline $\mathcal{G}_R$-representation finite free over $\mathbb{Z}_p$ whose Hodge–Tate weights lie in $[0, 1]$. Then there exists a p-divisible group G_R over R such that $T_p(G_R) \cong T$ as $\mathcal{G}_R$-representations.*

As an immediate corollary using the results in [Moon 2020], we obtain the following result on the geometry of the locus of crystalline $\mathcal{G}_R$-representations with Hodge–Tate weights in $[0, 1]$. For a fixed absolutely irreducible $\mathbb{F}_p$-representation V_0 of $\mathcal{G}_R$, there exists a universal deformation ring which parametrizes the deformations of V_0 [de Smit and Lenstra 1997]. By [Moon 2020, Theorem 5.7], we deduce:

Corollary 1.3. *Suppose R has Krull dimension 2 and $e < p - 1$. Then the locus of crystalline representations of $\mathcal{G}_R$ with Hodge–Tate weights in $[0, 1]$ cuts out a closed subscheme of the universal deformation space.*

We give a more precise statement of Corollary 1.3 in Section 6. The assumption that R has Krull dimension 2 appears in [Moon 2020, Theorem 5.7], since the construction of Barsotti–Tate deformation ring in [Moon 2020, Section 5] uses the result in [de Smit and Lenstra 1997] and relies on the assumption.

We now explain the major ingredients for the proof of Theorem 1.2. Firstly, Kim [2015] generalized the Breuil–Kisin classification in the relative setting, and showed that the category of p-divisible groups over R is anti-equivalent to the category of Kisin modules of height 1 over $R_0[\![u]\!]$. Using the classification, we reduce our problem to constructing desired Kisin modules. Secondly, Brinon and Trihan [2008] proved the generalization of Theorem 1.1 for the case when the base is a complete discrete valuation ring whose residue field has a finite p-basis. To construct appropriate Kisin modules, we use their result together with the fact that the p-adic completion of $R_{0,(p)}$ is an example of such a ring. We remark that our construction of Kisin modules relies on the assumption that the ramification index is small.

1A. ***Notations.*** We will reserve φ for various Frobenius. To be more precise, let A be an $W(k)$-algebra on which the arithmetic Frobenius φ on $W(k)$ extends, and M an A-module. We denote $\varphi_A : A \to A$ for such an extension. Let $\varphi_M : M \to M$ be a φ_A-semilinear map. This is equivalent to having an A-linear map $1 \otimes \varphi_M : \varphi_A^* M \to M$, where $\varphi_A^* M$ denotes $A \otimes_{\varphi_A, A} M$. We always drop the subscripts A and M from φ if no confusion arises. Let $f : A \to B$ be a ring map compatible with Frobenius, that is, $f \circ \varphi_A = \varphi_B \circ f$. Then φ_M naturally extends to $\varphi_{M_B} : M_B \to M_B$ for $M_B := B \otimes_A M$. It is easy to check that $\varphi_B^* M_B = B \otimes_A \varphi_A^* M$ and $1 \otimes \varphi_{M_B} : \varphi_B^* M_B \to M_B$ is equal to $B \otimes_A (1 \otimes \varphi_M)$.

2. Relative p-adic Hodge theory and étale φ-modules

2A. ***Base ring and crystalline period ring in the relative case.*** We follow the same notations as in the Introduction. We recall the assumptions on the base rings and the construction of crystalline period ring

in relative p-adic Hodge theory in [Kim 2015] (see also [Brinon 2008]), together with an additional assumption which will be needed later. Let R_0 be a ring obtained from $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$ by a finite number of iterations of the following operations:

- p-adic completion of an étale extension;
- p-adic completion of a localization;
- completion with respect to an ideal containing p.

We suppose that R_0 is an integral domain separated and complete with respect to some ideal $J \subset R_0$ containing p, such that R_0/J is finitely generated over some field k' (see [Kim 2015, Section 2.2.2]). We further assume that R_0/pR_0 is a unique factorization domain.

R_0/pR_0 has a finite p-basis given by $\{t_1, \dots, t_m\}$ in the sense of [de Jong 1995, Definition 1.1.1]. The Witt vector Frobenius on $W(k)$ extends (not necessarily uniquely) to R_0, and we fix such a Frobenius endomorphism $\varphi : R_0 \to R_0$. Let $\widehat{\Omega}_{R_0} := \varprojlim_n \Omega_{(R_0/p^n)/W(k)}$ be the module of p-adically continuous Kähler differentials. By [Brinon 2008, Proposition 2.0.2], $\widehat{\Omega}_{R_0} \cong \bigoplus_{i=1}^m R_0 \cdot dt_i$. We work over the base ring R given by $R := R_0 \otimes_{W(k)} \mathcal{O}_K$.

Let $\bar{R}$ denote the union of finite R-subalgebras R' of a fixed separable closure of $\mathrm{Frac}(R)$ such that $R'\big[\frac{1}{p}\big]$ is étale over $R\big[\frac{1}{p}\big]$. Then $\operatorname{Spec} \bar{R}\big[\frac{1}{p}\big]$ is a pro-universal covering of $\operatorname{Spec} R\big[\frac{1}{p}\big]$, and $\bar{R}$ is the integral closure of R in $\bar{R}\big[\frac{1}{p}\big]$. Let $\mathcal{G}_R := \mathrm{Gal}\big(\bar{R}\big[\frac{1}{p}\big]/R\big[\frac{1}{p}\big]\big) = \pi_1^{\text{ét}}\big(\operatorname{Spec} R\big[\frac{1}{p}\big]\big)$. By a representation of $\mathcal{G}_R$, we always mean a finite continuous representation.

The crystalline period ring $B_{\mathrm{cris}}(R)$ is constructed as follows. Let $\bar{R}^\flat = \varprojlim_\varphi \bar{R}/p\bar{R}$. There exists a natural $W(k)$-linear surjective map $\theta : W(\bar{R}^\flat) \to \widehat{\bar{R}}$ which lifts the projection onto the first factor. Here, $\widehat{\bar{R}}$ denotes the p-adic completion of $\bar{R}$. Let $\theta_{R_0} : R_0 \otimes_{W(k)} W(\bar{R}^\flat) \to \widehat{\bar{R}}$ be the R_0-linear extension of θ. Define the integral crystalline period ring $A_{\mathrm{cris}}(R)$ to be the p-adic completion of the divided power envelope of $R_0 \otimes_{W(k)} W(\bar{R}^\flat)$ with respect to $\ker(\theta_{R_0})$. Choose compatibly $\epsilon_n \in \bar{R}$ such that $\epsilon_0 = 1$, $\epsilon_n = \epsilon_{n+1}^p$ with $\epsilon_1 \neq 1$, and let $\tilde{\epsilon} = (\epsilon_n)_{n \geq 0} \in \bar{R}^\flat$. Then $\tau := \log[\tilde{\epsilon}] \in A_{\mathrm{cris}}(R)$. Define $B_{\mathrm{cris}}(R) = A_{\mathrm{cris}}(R)\big[\frac{1}{\tau}\big]$. $B_{\mathrm{cris}}(R)$ is equipped naturally with $\mathcal{G}_R$-action and Frobenius endomorphism, and $B_{\mathrm{cris}}(R) \otimes_{R_0[\frac{1}{p}]} R\big[\frac{1}{p}\big]$ is equipped with a natural filtration by $R\big[\frac{1}{p}\big]$-submodules. Furthermore, we have a natural integrable connection $\nabla : B_{\mathrm{cris}}(R) \to B_{\mathrm{cris}}(R) \otimes_{R_0} \widehat{\Omega}_{R_0}$ such that Frobenius is horizontal and Griffiths transversality is satisfied.

For a $\mathcal{G}_R$-representation V over $\mathbb{Q}_p$, let $D_{\mathrm{cris}}(V) := \mathrm{Hom}_{\mathcal{G}_R}(V, B_{\mathrm{cris}}(R))$. The natural morphism

$$\alpha_{\mathrm{cris}} : D_{\mathrm{cris}}(V) \otimes_{R_0[\frac{1}{p}]} B_{\mathrm{cris}}(R) \to V^\vee \otimes_{\mathbb{Q}_p} B_{\mathrm{cris}}(R)$$

is injective. We say V is *crystalline* if α_{cris} is an isomorphism. When V is crystalline, then $D_{\mathrm{cris}}(V)$ is a finite projective $R_0\big[\frac{1}{p}\big]$-module, and $D_{\mathrm{cris}}(V) \otimes_{R_0[\frac{1}{p}]} R\big[\frac{1}{p}\big]$ has the filtration induced by that on $B_{\mathrm{cris}}(R) \otimes_{R_0[\frac{1}{p}]} R\big[\frac{1}{p}\big]$. We define the Hodge–Tate weights similarly as in the classical p-adic Hodge theory. Frobenius and connection on $B_{\mathrm{cris}}(R)$ induce those structures on $D_{\mathrm{cris}}(V)$; for the Frobenius endomorphism on $D_{\mathrm{cris}}(V)$, $1 \otimes \varphi : \varphi^* D_{\mathrm{cris}}(V) \to D_{\mathrm{cris}}(V)$ is an isomorphism, and the connection $\nabla : D_{\mathrm{cris}}(V) \to D_{\mathrm{cris}}(V) \otimes_{R_0} \widehat{\Omega}_{R_0}$ is integrable and topologically quasinilpotent. Furthermore, Griffiths transversality is satisfied and φ is horizontal. For a $\mathcal{G}_R$-representation T which is free over $\mathbb{Z}_p$, we say it is crystalline if $T\big[\frac{1}{p}\big]$ is crystalline.

Suppose S_0 is another relative base ring over $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$ satisfying the above conditions and equipped with a choice of Frobenius, and let $b : R_0 \to S_0$ be a φ-equivariant $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$-algebra map. We also denote $b : R = R_0 \otimes_{W(k)} \mathcal{O}_K \to S := S_0 \otimes_{W(k)} \mathcal{O}_K$ the map induced $\mathcal{O}_K$-linearly. By choosing a common geometric point, this induces a map of Galois groups $\mathcal{G}_S \to \mathcal{G}_R$, and also a map of crystalline period rings $B_{\mathrm{cris}}(R) \to B_{\mathrm{cris}}(S)$ compatible with all structures. If V is a crystalline representation of $\mathcal{G}_R$ with certain Hodge–Tate weights, then via these maps V is also a crystalline representation of $\mathcal{G}_S$ with the same Hodge–Tate weights, and the construction of $D_{\mathrm{cris}}(V)$ is compatible with the base change.

We will consider the following base change maps in later sections. Let $\mathcal{O}_{L_0}$ be the p-adic completion of $R_{0,(p)}$, and let $b_L : R_0 \to \mathcal{O}_{L_0}$ be the natural φ-equivariant map. This induces $b_L : R \to \mathcal{O}_L := \mathcal{O}_{L_0} \otimes_{W(k)} \mathcal{O}_K$. Note that $L = \mathcal{O}_L\big[\frac{1}{p}\big]$ is an example of a complete discrete valuation field with a residue field having a finite p-basis, studied in [Brinon and Trihan 2008]. On the other hand, for each maximal ideal $\mathfrak{q} \in \mathrm{mSpec}\, R_0$, let $\widehat{R_{0,\mathfrak{q}}}$ be the $\mathfrak{q}$-adic completion of $R_{0,\mathfrak{q}}$. By the structure theorem of complete regular local rings, we have $\widehat{R_{0,\mathfrak{q}}} \cong \mathcal{O}_{\mathfrak{q}}[\![s_1, \dots, s_l]\!]$ where $\mathcal{O}_{\mathfrak{q}}$ is a Cohen ring with the maximal ideal (p) and $l \geq 0$ is an integer ($\widehat{R_{0,\mathfrak{q}}}$ is understood to be $\mathcal{O}_{\mathfrak{q}}$ when $l = 0$). We consider the natural φ-equivariant morphism $b_{\mathfrak{q}} : R_0 \to \widehat{R_{0,\mathfrak{q}}}$, which induces $b_{\mathfrak{q}} : R \to R_{\mathfrak{q}} := \widehat{R_{0,\mathfrak{q}}} \otimes_{W(k)} \mathcal{O}_K$.

2B. ***Étale φ-modules.*** We study étale φ-modules and associated Galois representations. Most of the material in this section is a review of [Kim 2015, Section 7], and the underlying geometry is based on perfectoid spaces as in [Scholze 2012].

Let R_0 be a relative base ring over $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$ and let $R = R_0 \otimes_{W(k)} \mathcal{O}_K$ as above. Choose a uniformizer $\varpi \in \mathcal{O}_K$. For integers $n \geq 0$, we choose compatibly $\varpi_n \in \overline{K}$ such that $\varpi_0 = \varpi$ and $\varpi_{n+1}^p = \varpi_n$, and let K_∞ be the p-adic completion of $\bigcup_{n \geq 0} K(\varpi_n)$. Then K_∞ is a perfectoid field and $\big(\widehat{\widetilde{R}}\big[\frac{1}{p}\big], \widehat{\widetilde{R}}\big)$ is a perfectoid affinoid K_∞-algebra. Let $K_\infty^\flat$ denote the tilt of K_∞ as defined in [Scholze 2012], and let $\underline{\varpi} := (\varpi_n) \in K_\infty^\flat$.

Let $\mathfrak{S} := R_0[\![u]\!]$ equipped with the Frobenius extending that on R_0 by $\varphi(u) = u^p$. Let $E^+_{R_\infty} = \mathfrak{S}/p\mathfrak{S}$, and let $\widetilde{E}^+_{R_\infty}$ be the u-adic completion of $\varinjlim_\varphi E^+_{R_\infty}$. Let $E_{R_\infty} = E^+_{R_\infty}\big[\frac{1}{u}\big]$ and $\widetilde{E}_{R_\infty} = \widetilde{E}^+_{R_\infty}\big[\frac{1}{u}\big]$. By [Scholze 2012, Proposition 5.9], $(\widetilde{E}_{R_\infty}, \widetilde{E}^+_{R_\infty})$ is a perfectoid affinoid $K_\infty^\flat$-algebra, and we have the natural injective map $(\widetilde{E}_{R_\infty}, \widetilde{E}^+_{R_\infty}) \hookrightarrow \big(\overline{R}^\flat\big[\frac{1}{\underline{\varpi}}\big], \overline{R}^\flat\big)$ given by $u \mapsto \underline{\varpi}$.

Let

$$\widetilde{R}_\infty := W(\widetilde{E}^+_{R_\infty}) \otimes_{W(K_\infty^{\flat\circ}),\theta} \mathcal{O}_{K_\infty}. \tag{2-1}$$

By [Scholze 2012, Remark 5.19], $\big(\widetilde{R}_\infty\big[\frac{1}{p}\big], \widetilde{R}_\infty\big)$ is a perfectoid affinoid K_∞-algebra whose tilt is $(\widetilde{E}_{R_\infty}, \widetilde{E}^+_{R_\infty})$. Furthermore, it is shown in [Kim 2015] that we have a natural injective map

$$\Big(\widetilde{R}_\infty\Big[\frac{1}{p}\Big], \widetilde{R}_\infty\Big) \hookrightarrow \Big(\widehat{\widetilde{R}}\Big[\frac{1}{p}\Big], \widehat{\widetilde{R}}\Big)$$

whose tilt is $(\widetilde{E}_{R_\infty}, \widetilde{E}^+_{R_\infty}) \hookrightarrow \big(\overline{R}^\flat\big[\frac{1}{\underline{\varpi}}\big], \overline{R}^\flat\big)$. For $\mathcal{G}_{\widetilde{R}_\infty} := \pi_1^{\text{ét}}\big(\mathrm{Spec}\, \widetilde{R}_\infty\big[\frac{1}{p}\big]\big)$, we then have a continuous map of Galois groups $\mathcal{G}_{\widetilde{R}_\infty} \to \mathcal{G}_R$, which is a closed embedding by [Gabber and Ramero 2003, Proposition 5.4.54]. By the almost purity theorem in [Scholze 2012], $\overline{R}^\flat\big[\frac{1}{\underline{\varpi}}\big]$ can be canonically identified with the

$\underline{\varpi}$-adic completion of the affine ring of a pro-universal covering of Spec $\widetilde{E}_{R_\infty}$, and letting $\mathcal{G}_{\widetilde{E}_{R_\infty}}$ be the Galois group corresponding to the pro-universal covering, there exists a canonical isomorphism $\mathcal{G}_{\widetilde{E}_{R_\infty}} \cong \mathcal{G}_{\widetilde{R}_\infty}$.

Lemma 2.1. *Consider the map of Galois groups $\mathcal{G}_{\mathcal{O}_L} \to \mathcal{G}_R$ induced by choosing a common geometric point for the base change map $b_L : R \to \mathcal{O}_L$ in Section 2A. Then the images of $\mathcal{G}_{\mathcal{O}_L}$ and $\mathcal{G}_{\widetilde{R}_\infty}$ inside $\mathcal{G}_R$ generate the group $\mathcal{G}_R$.*

Proof. $E^+_{R_\infty}$ has a finite p-basis given by $\{t_1, \dots, t_m, u\}$. Note that for any element of $g \in \mathcal{G}_R$, there exists an element $h \in \mathcal{G}_{\mathcal{O}_L}$ whose image in $\mathcal{G}_R$ induces the same actions on $t_1^{1/p^\infty}, \dots, t_m^{1/p^\infty}, \varpi^{1/p^\infty}$. Since $\widetilde{R}_\infty = W(\widetilde{E}^+_{R_\infty}) \otimes_{W(K^{\flat\circ}_\infty),\theta} \mathcal{O}_{K_\infty}$, the actions of g and h are the same on the elements of $\widetilde{R}_\infty$. Hence, the assertion follows. □

Now, let $\mathcal{O}_\mathcal{E}$ be the p-adic completion of $\mathfrak{S}\left[\frac{1}{u}\right]$. Note that φ on $\mathfrak{S}$ extends naturally to $\mathcal{O}_\mathcal{E}$.

Definition 2.2. An *étale* $(\varphi, \mathcal{O}_\mathcal{E})$*-module* is a pair $(\mathcal{M}, \varphi_\mathcal{M})$ where $\mathcal{M}$ is a finitely generated $\mathcal{O}_\mathcal{E}$-module and $\varphi_\mathcal{M} : \mathcal{M} \to \mathcal{M}$ is a φ-semilinear endomorphism such that $1 \otimes \varphi_\mathcal{M} : \varphi^*\mathcal{M} \to \mathcal{M}$ is an isomorphism. We say that an étale $(\varphi, \mathcal{O}_\mathcal{E})$-module is *projective* (resp. *torsion*) if the underlying $\mathcal{O}_\mathcal{E}$-module $\mathcal{M}$ is projective (resp. p-power torsion).

Let $\mathrm{Mod}_{\mathcal{O}_\mathcal{E}}$ denote the category of étale $(\varphi, \mathcal{O}_\mathcal{E})$-modules whose morphisms are $\mathcal{O}_\mathcal{E}$-module maps compatible with Frobenius. Let $\mathrm{Mod}^{\mathrm{pr}}_{\mathcal{O}_\mathcal{E}}$ and $\mathrm{Mod}^{\mathrm{tor}}_{\mathcal{O}_\mathcal{E}}$ respectively denote the full subcategories of projective and torsion objects. Note that we have a natural notion of a subquotient, direct sum, and tensor product for étale $(\varphi, \mathcal{O}_\mathcal{E})$-modules, and duality is defined for projective and torsion objects.

Lemma 2.3. *Let $\mathcal{M} \in \mathrm{Mod}^{\mathrm{tor}}_{\mathcal{O}_\mathcal{E}}$ be a torsion étale φ-module annihilated by p. Then $\mathcal{M}$ is a projective $\mathcal{O}_\mathcal{E}/p\mathcal{O}_\mathcal{E}$-module.*

Proof. This follows from essentially the same proof as in [Andreatta 2006, Lemma 7.10]. □

We consider $W\big(\overline{R}^\flat\big[\frac{1}{\underline{\varpi}}\big]\big)$ as an $\mathcal{O}_\mathcal{E}$-algebra via mapping u to the Teichmüller lift $[\underline{\varpi}]$ of $\underline{\varpi}$, and let $\mathcal{O}^{\mathrm{ur}}_\mathcal{E}$ be the integral closure of $\mathcal{O}_\mathcal{E}$ in $W\big(\overline{R}^\flat\big[\frac{1}{\underline{\varpi}}\big]\big)$. Let $\widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E}$ be its p-adic completion. Since $\mathcal{O}_\mathcal{E}$ is normal, we have $\mathrm{Aut}_{\mathcal{O}_\mathcal{E}}(\mathcal{O}^{\mathrm{ur}}_\mathcal{E}) \cong \mathcal{G}_{E_{R_\infty}} := \pi_1^{\text{ét}}(\mathrm{Spec}\, E_{R_\infty})$, and by [Gabber and Ramero 2003, Proposition 5.4.54] and the almost purity theorem, we have $\mathcal{G}_{E_{R_\infty}} \cong \mathcal{G}_{\widetilde{E}_{R_\infty}} \cong \mathcal{G}_{\widetilde{R}_\infty}$. This induces $\mathcal{G}_{\widetilde{R}_\infty}$-action on $\widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E}$. The following is proved in [Kim 2015].

Lemma 2.4 [Kim 2015, Lemmas 7.5 and 7.6]. *We have $(\widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E})^{\mathcal{G}_{\widetilde{R}_\infty}} = \mathcal{O}_\mathcal{E}$ and the same holds modulo p^n. Furthermore, there exists a unique $\mathcal{G}_{\widetilde{R}_\infty}$-equivariant ring endomorphism φ on $\widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E}$ lifting the p-th power map on $\widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E}/(p)$ and extending φ on $\mathcal{O}_\mathcal{E}$. The inclusion $\widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E} \hookrightarrow W\big(\overline{R}^\flat\big[\frac{1}{\underline{\varpi}}\big]\big)$ is φ-equivariant where the latter ring is given the Witt vector Frobenius.*

Let $\mathrm{Rep}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$ be the category of $\mathbb{Z}_p$-representations of $\mathcal{G}_{\widetilde{R}_\infty}$, and let $\mathrm{Rep}^{\mathrm{pr}}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$ and $\mathrm{Rep}^{\mathrm{tor}}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$ respectively denote the full subcategories of free and torsion objects. For $\mathcal{M} \in \mathrm{Mod}_{\mathcal{O}_\mathcal{E}}$ and $T \in \mathrm{Rep}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$, we define $T(\mathcal{M}) := (\mathcal{M} \otimes_{\mathcal{O}_\mathcal{E}} \widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E})^{\varphi=1}$ and $\mathcal{M}(T) := (T \otimes_{\mathbb{Z}_p} \widehat{\mathcal{O}}^{\mathrm{ur}}_\mathcal{E})^{\mathcal{G}_{\widetilde{R}_\infty}}$. For a torsion étale φ-module $\mathcal{M} \in \mathrm{Mod}^{\mathrm{tor}}_{\mathcal{O}_\mathcal{E}}$, we define its *length* to be the length of $\mathcal{M} \otimes_{\mathcal{O}_\mathcal{E}} (\mathcal{O}_\mathcal{E})_{(p)}$ as an $(\mathcal{O}_\mathcal{E})_{(p)}$-module.

Proposition 2.5 [Kim 2015, Proposition 7.7]. *The assignments $T(\cdot)$ and $\mathcal{M}(\cdot)$ are exact equivalences (inverse of each other) of $\otimes$-categories between $\mathrm{Mod}_{\mathcal{O}_\mathcal{E}}$ and $\mathrm{Rep}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$. Moreover, $T(\cdot)$ and $\mathcal{M}(\cdot)$ restrict to rank-preserving equivalence of categories between $\mathrm{Mod}^{\mathrm{pr}}_{\mathcal{O}_\mathcal{E}}$ and $\mathrm{Rep}^{\mathrm{pr}}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$ and length-preserving equivalence of categories between $\mathrm{Mod}^{\mathrm{tor}}_{\mathcal{O}_\mathcal{E}}$ and $\mathrm{Rep}^{\mathrm{tor}}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$. In both cases, $T(\cdot)$ and $\mathcal{M}(\cdot)$ commute with taking duals.*

Proof. This is [Kim 2015, Proposition 7.7]. We remark here for some additional details. Note that E_{R_∞} is a normal domain and $\pi_1^{\text{ét}}(\mathrm{Spec}\,\mathbb{E}_{R_\infty}) \cong \mathcal{G}_{\widetilde{R}_\infty}$. Given Lemma 2.3, the assertion therefore follows from the usual dévissage and [Katz 1973, Lemma 4.1.1]. Note that both functors $T(\cdot)$ and $\mathcal{M}(\cdot)$ are a priori left exact by definition, and exactness can be proved by the same argument as in the proof of [Andreatta 2006, Theorem 7.11]. □

Suppose S_0 is another relative base ring over $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$ as in Section 2A equipped with a choice of Frobenius, and suppose $b : R_0 \hookrightarrow S_0$ be a φ-equivariant $W(k)\langle t_1^{\pm 1}, \dots, t_m^{\pm 1}\rangle$-algebra map which is injective. Let $b : R = R_0 \otimes_{W(k)} \mathcal{O}_K \hookrightarrow S := S_0 \otimes_{W(k)} \mathcal{O}_K$ be the induced injective map. By choosing a common geometric point we have an injective map $\bar{R} \hookrightarrow \bar{S}$, and this induces an embedding $\widetilde{R}_\infty \hookrightarrow \widetilde{S}_\infty$ by the constructions given in (2-1). Hence, the corresponding map of Galois groups $\mathcal{G}_S \to \mathcal{G}_R$ restricts to $\mathcal{G}_{\widetilde{S}_\infty} \to \mathcal{G}_{\widetilde{R}_\infty}$. Let $\mathfrak{S}_S = S_0[\![u]\!]$ and let $\mathcal{O}_{\mathcal{E},S}$ be the p-adic completion of $\mathfrak{S}_S\big[\frac{1}{u}\big]$. Let $\mathcal{M}_S(\cdot)$ be the functor for the base ring S constructed similarly as above. Let $T \in \mathrm{Rep}^{\mathrm{pr}}_{\mathbb{Z}_p}(\mathcal{G}_{\widetilde{R}_\infty})$. Then T is also a $\mathcal{G}_{\widetilde{S}_\infty}$-representation via the map $\mathcal{G}_{\widetilde{S}_\infty} \to \mathcal{G}_{\widetilde{R}_\infty}$, and we have the natural isomorphism $\mathcal{M}(T) \otimes_{\mathcal{O}_\mathcal{E}} \mathcal{O}_{\mathcal{E},S} \cong \mathcal{M}_S(T)$ as étale $(\varphi, \mathcal{O}_{\mathcal{E},S})$-modules by the definition of the functors $\mathcal{M}(\cdot)$ and $T(\cdot)$ and by Proposition 2.5.

3. Relative Breuil–Kisin classification

We now explain the classification of p-divisible groups over $\mathrm{Spec}\,R$ via Kisin modules, which is proved in [Kisin 2006] when $R = \mathcal{O}_K$ and generalized in [Kim 2015] for the relative case. Denote by $E(u)$ the Eisenstein polynomial for the extension K over $W(k)\big[\frac{1}{p}\big]$, and let $\mathfrak{S} = R_0[\![u]\!]$ as above.

Definition 3.1. Denote by $\mathrm{Kis}^1(\mathfrak{S})$ the category of pairs $(\mathfrak{M}, \varphi_\mathfrak{M})$ where

- $\mathfrak{M}$ is a finitely generated projective $\mathfrak{S}$-module;
- $\varphi_\mathfrak{M} : \mathfrak{M} \to \mathfrak{M}$ is a φ-semilinear map such that $\mathrm{coker}(1 \otimes \varphi_\mathfrak{M})$ is annihilated by $E(u)$.

The morphisms are $\mathfrak{S}$-module maps compatible with Frobenius.

Note that for $(\mathfrak{M}, \varphi_\mathfrak{M}) \in \mathrm{Kis}^1(\mathfrak{S})$, $1 \otimes \varphi_\mathfrak{M} : \varphi^*\mathfrak{M} \to \mathfrak{M}$ is injective since $\mathfrak{M}$ is finite projective over $\mathfrak{S}$ and $\mathrm{coker}(1 \otimes \varphi_\mathfrak{M})$ is killed by $E(u)$. Consider the composite $\mathfrak{S} \twoheadrightarrow \mathfrak{S}/u\mathfrak{S} = R_0 \xrightarrow{\varphi} R_0$.

Definition 3.2. A *Kisin module* of height 1 is a tuple $(\mathfrak{M}, \varphi_\mathfrak{M}, \nabla_\mathfrak{M})$ such that:

- $(\mathfrak{M}, \varphi_\mathfrak{M}) \in \mathrm{Kis}^1(\mathfrak{S})$.
- Let $\mathcal{N} := \mathfrak{M} \otimes_{\mathfrak{S},\varphi} R_0$ equipped with the Frobenius $\varphi_\mathfrak{M} \otimes \varphi_{R_0}$. Then $\nabla_\mathfrak{M} : \mathcal{N} \to \mathcal{N} \otimes_{R_0} \widehat{\Omega}_{R_0}$ is a topologically quasinilpotent integrable connection commuting with Frobenius.

Here, $\nabla_{\mathfrak{M}}$ being topologically quasinilpotent means that the induced connection on $\mathcal{N}/p\mathcal{N}$ is nilpotent. Denote by $\mathrm{Kis}^1(\mathfrak{S}, \nabla)$ the category of Kisin modules of height 1 whose morphisms are $\mathfrak{S}$-module maps compatible with Frobenius and connection.

The following theorem classifying the p-divisible groups is proved in [Kim 2015].

Theorem 3.3 [Kim 2015, Corollary 6.7 and Remark 6.9]. *There exists an exact anti-equivalence of categories*

$$\mathfrak{M}^* : \{p\text{-divisible groups over } \operatorname{Spec} R\} \to \mathrm{Kis}^1(\mathfrak{S}, \nabla).$$

Let S_0 be another base ring satisfying the condition as in Section 2A and equipped with a Frobenius, and let $b : R_0 \to S_0$ be a φ-equivariant map. Then the formation of $\mathfrak{M}^*$ commutes with the base change $R \to S := S_0 \otimes_{W(k)} \mathcal{O}_K$ induced $\mathcal{O}_K$-linearly from b.

Note that if $(\mathfrak{M}, \varphi_{\mathfrak{M}}) \in \mathrm{Kis}^1(\mathfrak{S})$, then $(\mathfrak{M} \otimes_{\mathfrak{S}} \mathcal{O}_{\mathcal{E}}, \varphi_{\mathfrak{M}} \otimes \varphi_{\mathcal{O}_{\mathcal{E}}})$ is a projective étale $(\varphi, \mathcal{O}_{\mathcal{E}})$-module since $1 \otimes \varphi_{\mathfrak{M}}$ is injective and its cokernel is killed by $E(u)$ which is a unit in $\mathcal{O}_{\mathcal{E}}$. If G_R is a p-divisible group over R, its Tate module is given by $T_p(G_R) := \mathrm{Hom}_{\bar{R}}(\mathbb{Q}_p/\mathbb{Z}_p, G_R \times_R \bar{R})$, which is a finite free $\mathbb{Z}_p$-representation of $\mathcal{G}_R$. By [Kim 2015, Corollary 8.2], we have a natural $\mathcal{G}_{\widetilde{R}_\infty}$-equivariant isomorphism $T^\vee(\mathfrak{M}^*(G_R) \otimes_{\mathfrak{S}} \mathcal{O}_{\mathcal{E}}) \cong T_p(G_R)$, where $T^\vee(\mathfrak{M}^*(G_R) \otimes_{\mathfrak{S}} \mathcal{O}_{\mathcal{E}})$ denotes the dual of $T(\mathfrak{M}^*(G_R) \otimes_{\mathfrak{S}} \mathcal{O}_{\mathcal{E}})$.

4. Construction of Kisin modules

In this section, we will assume $e < p - 1$ from Proposition 4.3. We denote $\mathfrak{S}_n := \mathfrak{S}/p^n\mathfrak{S}$ for positive integers $n \geq 1$. Let T be a crystalline $\mathcal{G}_R$-representation which is free over $\mathbb{Z}_p$ of rank d with Hodge–Tate weights in $[0, 1]$. Let $\mathcal{M} := \mathcal{M}^\vee(T)$ be the associated étale $(\varphi, \mathcal{O}_{\mathcal{E}})$-module, where $\mathcal{M}^\vee(T)$ denotes the dual of $\mathcal{M}(T)$. For each integer $n \geq 1$, denote $\mathcal{M}_n = \mathcal{M}/p^n\mathcal{M}$. Note that $\mathcal{M}_n \cong \mathcal{M}^\vee(T/p^nT)$. On the other hand, consider the map $b_L : R \to \mathcal{O}_L$ as in Section 2A. T is also a crystalline $\mathcal{G}_{\mathcal{O}_L}$-representation with Hodge–Tate weights in $[0, 1]$, so by [Brinon and Trihan 2008, Theorem 6.10], there exists a p-divisible group $G_{\mathcal{O}_L}$ over $\mathcal{O}_L$ such that $T_p(G_{\mathcal{O}_L}) \cong T$ as $\mathcal{G}_{\mathcal{O}_L}$-representations. Let $(\mathfrak{M}_{\mathcal{O}_L}, \nabla_{\mathfrak{M}_{\mathcal{O}_L}}) := \mathfrak{M}^*(G_{\mathcal{O}_L}) \in \mathrm{Kis}^1(\mathfrak{S}_{\mathcal{O}_L}, \nabla)$ be the associated Kisin module over $\mathfrak{S}_{\mathcal{O}_L}$. Denote $\mathfrak{M}_{\mathcal{O}_L,n} = \mathfrak{M}_{\mathcal{O}_L}/p^n\mathfrak{M}_{\mathcal{O}_L}$. The map between the Galois groups $\mathcal{G}_{\mathcal{O}_L} \to \mathcal{G}_R$ restricts to $\mathcal{G}_{\widetilde{\mathcal{O}}_L,\infty} \to \mathcal{G}_{\widetilde{R}_\infty}$. Hence, we have the natural isomorphism $\mathcal{M} \otimes_{\mathcal{O}_{\mathcal{E}}} \mathcal{O}_{\mathcal{E},\mathcal{O}_L} \cong \mathfrak{M}_{\mathcal{O}_L} \otimes_{\mathfrak{S}_{\mathcal{O}_L}} \mathcal{O}_{\mathcal{E},\mathcal{O}_L}$ of étale $(\varphi, \mathcal{O}_{\mathcal{E},\mathcal{O}_L})$-modules. Let $\mathcal{M}_{\mathcal{O}_L} := \mathcal{M} \otimes_{\mathcal{O}_{\mathcal{E}}} \mathcal{O}_{\mathcal{E},\mathcal{O}_L}$ and $\mathcal{M}_{\mathcal{O}_L,n} := \mathcal{M}_{\mathcal{O}_L}/p^n\mathcal{M}_{\mathcal{O}_L}$.

For each $n \geq 1$, we define

$$\mathfrak{M}_n := \mathcal{M}_n \cap \mathfrak{M}_{\mathcal{O}_L,n},$$

where the intersection is taken as $\mathfrak{S}$-submodules of $\mathcal{M}_{\mathcal{O}_L,n}$. The Frobenius endomorphisms on $\mathcal{M}_n$ and $\mathfrak{M}_{\mathcal{O}_L,n}$ induce a Frobenius endomorphism on $\mathfrak{M}_n$. Since the Frobenius on $\mathcal{M}_{\mathcal{O}_L,n}$ is injective, we have the injective $\mathfrak{S}$-module morphism

$$1 \otimes \varphi : \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{M}_n \to \mathfrak{M}_n$$

for each n.

Lemma 4.1. *$\mathfrak{M}_n$ is a finitely generated $\mathfrak{S}_n$-module. Furthermore, we have φ-equivariant isomorphisms*

$$\mathfrak{M}_n \otimes_{\mathfrak{S}} \mathcal{O}_{\mathcal{E}} \cong \mathcal{M}_n \quad \text{and} \quad \mathfrak{M}_n \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathcal{O}_L} \cong \mathfrak{M}_{\mathcal{O}_L,n}.$$

Proof. We first prove that $\mathfrak{M}_n$ is finite over $\mathfrak{S}_n$. Note that $\mathfrak{M}_{\mathcal{O}_L,n}$ is free over $\mathfrak{S}_{\mathcal{O}_L,n}$ of rank d, and choose a basis $\{e_1, \ldots, e_d\}$ of $\mathfrak{M}_{\mathcal{O}_L,n}$. On the other hand, since $\mathcal{M}_n$ is projective over $\mathfrak{S}_n\big[\frac{1}{u}\big]$ of rank d, there exists a nonzero divisor $g \in \mathfrak{S}_n$ such that $\mathcal{M}_n\big[\frac{1}{g}\big]$ is free of rank d over $\mathfrak{S}_n\big[\frac{1}{u}\big]\big[\frac{1}{g}\big]$. Since $\mathcal{M}_n$ is finite over $\mathfrak{S}_n\big[\frac{1}{u}\big]$, we can choose a basis $\{f_1, \ldots, f_d\}$ of $\mathcal{M}_n\big[\frac{1}{g}\big]$ over $\mathfrak{S}_n\big[\frac{1}{u}\big]\big[\frac{1}{g}\big]$ such that letting $\mathfrak{N}$ to be the $\mathfrak{S}_n$-submodule of $\mathcal{M}_n\big[\frac{1}{g}\big]$ generated by $f_1, \ldots, f_d$, we have $\mathcal{M}_n \subset \mathfrak{N}\big[\frac{1}{u}\big]$ as $\mathfrak{S}_n\big[\frac{1}{u}\big]$-modules. It suffices to show that $\mathfrak{M}_n \subset \frac{1}{u^h} \cdot \mathfrak{N}$ as $\mathfrak{S}_n$-modules for some integer $h \geq 1$. We have

$$(f_1, \ldots, f_d)^t = A \cdot (e_1, \ldots, e_d)^t,$$

where A is an invertible $d \times d$ matrix with entries in $\mathfrak{S}_{\mathcal{O}_L,n}\big[\frac{1}{u}\big]\big[\frac{1}{g}\big]$. Consider the intersection $\mathfrak{N}\big[\frac{1}{u}\big] \cap \mathfrak{M}_{\mathcal{O}_L,n}$ as submodules of $\mathfrak{M}_{\mathcal{O}_L,n}\big[\frac{1}{u}\big]\big[\frac{1}{g}\big]$. For an element $x = b_1 f_1 + \cdots + b_d f_d \in \mathfrak{N}\big[\frac{1}{u}\big]$ with $b_1, \ldots, b_d \in \mathfrak{S}_n\big[\frac{1}{u}\big]$, we have $x \in \mathfrak{M}_{\mathcal{O}_L,n}$ if and only if

$$(b_1, \ldots, b_d) \cdot A = (c_1, \ldots, c_d)$$

for some $c_1, \ldots, c_d \in \mathfrak{S}_{\mathcal{O}_L,n}$. Then $(b_1, \ldots, b_d) = (c_1, \ldots, c_d)A^{-1}$, which implies that $\mathfrak{N}\big[\frac{1}{u}\big] \cap \mathfrak{M}_{\mathcal{O}_L,n} \subset \frac{1}{u^h} \cdot \mathfrak{N}$ as $\mathfrak{S}_n$-modules for some integer $h \geq 1$. Since $\mathfrak{M}_n \subset \mathfrak{N}\big[\frac{1}{u}\big] \cap \mathfrak{M}_{\mathcal{O}_L,n}$, this shows the first statement.

We have

$$\mathfrak{M}_n \otimes_{\mathfrak{S}} \mathcal{O}_{\mathcal{E}} \cong \mathfrak{M}_n\Big[\frac{1}{u}\Big] \cong \mathcal{M}_n \cap \mathcal{M}_{\mathcal{O}_L,n} = \mathcal{M}_n$$

and hence the first isomorphism. On the other hand, since $\mathfrak{S} \to \mathfrak{S}_{\mathcal{O}_L}$ is flat and $\mathfrak{M}_{\mathcal{O}_L,n}$ is finite free over $\mathfrak{S}_{\mathcal{O}_L,n}$, we have

$$\begin{aligned}\mathfrak{M}_n \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathcal{O}_L} &\cong (\mathcal{M}_n \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathcal{O}_L}) \cap (\mathfrak{M}_{\mathcal{O}_L,n} \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathcal{O}_L}) = \mathcal{M}_{\mathcal{O}_L,n} \cap (\mathfrak{M}_{\mathcal{O}_L,n} \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathcal{O}_L}) \\ &\cong \Big(\mathfrak{M}_{\mathcal{O}_L,n} \otimes_{\mathfrak{S}_n} \mathfrak{S}_n\Big[\frac{1}{u}\Big]\Big) \cap (\mathfrak{M}_{\mathcal{O}_L,n} \otimes_{\mathfrak{S}_n} \mathfrak{S}_{\mathcal{O}_L,n}) \cong \mathfrak{M}_{\mathcal{O}_L,n}\end{aligned}$$

by $\mathfrak{S}_n\big[\frac{1}{u}\big] \cap \mathfrak{S}_{\mathcal{O}_L,n} = \mathfrak{S}_n$. □

Lemma 4.2. *The cokernel of the $\mathfrak{S}$-module map $1 \otimes \varphi : \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{M}_n \to \mathfrak{M}_n$ is killed by $E(u)$.*

Proof. Let $x \in \mathfrak{M}_n$. There exists a unique $y_1 \in \mathcal{O}_{\mathcal{E}} \otimes_{\varphi,\mathcal{O}_{\mathcal{E}}} \mathcal{M}_n \cong \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathcal{M}_n$ such that $(1 \otimes \varphi)(y_1) = E(u)x$. On the other hand, there exists a unique $y_2 \in \mathfrak{S}_{\mathcal{O}_L} \otimes_{\varphi,\mathfrak{S}_{\mathcal{O}_L}} \mathfrak{M}_{\mathcal{O}_L,n}$ such that $(1 \otimes \varphi)(y_2) = E(u)x$. Then we have $y_1 = y_2 \in (\mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathcal{M}_n) \cap (\mathfrak{S}_{\mathcal{O}_L} \otimes_{\varphi,\mathfrak{S}_{\mathcal{O}_L}} \mathfrak{M}_{\mathcal{O}_L,n})$.

Since $\mathcal{O}_{L_0}/p\mathcal{O}_{L_0}$ has a finite p-basis given by $t_1, \ldots, t_m \in R_0/pR_0$ which also gives a p-basis of R_0/pR_0, the natural map $\mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{M}_{\mathcal{O}_L,n} \to \mathfrak{S}_{\mathcal{O}_L} \otimes_{\varphi,\mathfrak{S}_{\mathcal{O}_L}} \mathfrak{M}_{\mathcal{O}_L,n}$ is an isomorphism. Hence

$$y_1 \in (\mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathcal{M}_n) \cap (\mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{M}_{\mathcal{O}_L,n}) \cong \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} (\mathcal{M}_n \cap \mathfrak{M}_{\mathcal{O}_L,n}) = \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{M}_n$$

since $\varphi : \mathfrak{S} \to \mathfrak{S}$ is flat by [Brinon 2008, Lemma 7.1.8]. This proves the assertion. □

For any finite $\mathfrak{S}$-module $\mathfrak{N}$ equipped with a φ-semilinear endomorphism $\varphi : \mathfrak{N} \to \mathfrak{N}$, say $\mathfrak{N}$ has $E(u)$-*height* ≤ 1 if there exists an $\mathfrak{S}$-module map $\psi : \mathfrak{N} \to \varphi^*\mathfrak{N} = \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{N}$ such that the composite

$$\varphi^*\mathfrak{N} \xrightarrow{1\otimes\varphi} \mathfrak{N} \xrightarrow{\psi} \varphi^*\mathfrak{N}$$

is $E(u) \cdot \mathrm{Id}_{\varphi^*\mathfrak{N}}$. By Lemma 4.2, $\mathfrak{M}_n$ has $E(u)$-height ≤ 1.

For each maximal ideal $\mathfrak{q} \in \mathrm{mSpec} R_0$, consider $b_\mathfrak{q} : R \to R_\mathfrak{q}$ as in Section 2A. By choosing a common geometric point, we have the induced map of Galois groups $\mathcal{G}_{R_\mathfrak{q}} \to \mathcal{G}_R$ which restricts to $\mathcal{G}_{\widetilde{R}_{\mathfrak{q},\infty}} \to \mathcal{G}_{\widetilde{R},\infty}$, and T is a crystalline $\mathcal{G}_{R_\mathfrak{q}}$-representation with Hodge–Tate weights in $[0, 1]$. Denote $\mathfrak{S}_\mathfrak{q} := \widehat{R_{0,\mathfrak{q}}}[\![u]\!]$.

Proposition 4.3. *Assume $e < p - 1$. For each integer $n \geq 1$, $\mathfrak{M}_n$ is projective over $\mathfrak{S}_n$ of rank d.*

Proof. Let $\mathfrak{q}$ be a maximal ideal of R_0, and let $\mathfrak{N}_n := \mathfrak{M}_n \otimes_\mathfrak{S} \mathfrak{S}_\mathfrak{q}$ equipped with the induced Frobenius endomorphism. Then we have the induced $\mathfrak{S}_\mathfrak{q}$-linear map $\psi : \mathfrak{N}_n \to \mathfrak{S}_\mathfrak{q} \otimes_{\varphi,\mathfrak{S}_\mathfrak{q}} \mathfrak{N}_n$ such that the composite

$$\mathfrak{S}_\mathfrak{q} \otimes_{\varphi,\mathfrak{S}_\mathfrak{q}} \mathfrak{N}_n \xrightarrow{1\otimes\varphi} \mathfrak{N}_n \xrightarrow{\psi} \mathfrak{S}_\mathfrak{q} \otimes_{\varphi,\mathfrak{S}_\mathfrak{q}} \mathfrak{N}_n$$

is $E(u) \cdot \mathrm{Id}$. For the isomorphism $\widehat{R_{0,\mathfrak{q}}} \cong \mathcal{O}_\mathfrak{q}[\![s_1, \ldots, s_l]\!]$ as above, let us consider the projection $\mathfrak{S}_\mathfrak{q} \to \mathfrak{S}_\mathfrak{q}/(p, s_1, \ldots, s_l) \cong k_\mathfrak{q}[\![u]\!]$, where $k_\mathfrak{q} := \mathcal{O}_\mathfrak{q}/(p)$. Denote $\overline{\mathfrak{N}}_n = \mathfrak{N}_n \otimes_{\mathfrak{S}_\mathfrak{q}} k_\mathfrak{q}[\![u]\!]$ equipped with the induced Frobenius. Then we have the induced $k_\mathfrak{q}[\![u]\!]$-linear map $\psi : \overline{\mathfrak{N}}_n \to k_\mathfrak{q}[\![u]\!] \otimes_{\varphi,k_\mathfrak{q}[\![u]\!]} \overline{\mathfrak{N}}_n$ such that the composite

$$k_\mathfrak{q}[\![u]\!] \otimes_{\varphi,k_\mathfrak{q}[\![u]\!]} \overline{\mathfrak{N}}_n \xrightarrow{1\otimes\varphi} \overline{\mathfrak{N}}_n \xrightarrow{\psi} k_\mathfrak{q}[\![u]\!] \otimes_{\varphi,k_\mathfrak{q}[\![u]\!]} \overline{\mathfrak{N}}_n$$

is $u^e \cdot \mathrm{Id}$. Since $k_\mathfrak{q}[\![u]\!]$ is a principal ideal domain, $\overline{\mathfrak{N}}_n$ is a direct sum of its free part and u-torsion part $\overline{\mathfrak{N}}_n \cong \overline{\mathfrak{N}}_{n,\mathrm{free}} \oplus \overline{\mathfrak{N}}_{n,\mathrm{tor}}$ as $k_\mathfrak{q}[\![u]\!]$-modules. Furthermore, φ maps $\overline{\mathfrak{N}}_{n,\mathrm{tor}}$ into $\overline{\mathfrak{N}}_{n,\mathrm{tor}}$, and hence the above maps induce

$$k_\mathfrak{q}[\![u]\!] \otimes_{\varphi,k_\mathfrak{q}[\![u]\!]} \overline{\mathfrak{N}}_{n,\mathrm{tor}} \xrightarrow{1\otimes\varphi} \overline{\mathfrak{N}}_{n,\mathrm{tor}} \xrightarrow{\psi} k_\mathfrak{q}[\![u]\!] \otimes_{\varphi,k_\mathfrak{q}[\![u]\!]} \overline{\mathfrak{N}}_{n,\mathrm{tor}}$$

whose composite is $u^e \cdot \mathrm{Id}$.

We claim that $\overline{\mathfrak{N}}_{n,\mathrm{tor}} = 0$. Suppose otherwise. Then $\overline{\mathfrak{N}}_{n,\mathrm{tor}} \cong \bigoplus_{i=1}^b k_\mathfrak{q}[\![u]\!]/(u^{a_i})$ for some integers $a_i \geq 1$, and $k_\mathfrak{q}[\![u]\!] \otimes_{\varphi,k_\mathfrak{q}[\![u]\!]} \overline{\mathfrak{N}}_{n,\mathrm{tor}} \cong \bigoplus_{i=1}^b k_\mathfrak{q}[\![u]\!]/(u^{pa_i})$. By taking the appropriate wedge product and letting $a = a_1 + \cdots + a_b$, the above maps induce the map of $k_\mathfrak{q}[\![u]\!]$-modules

$$k_\mathfrak{q}[\![u]\!]/(u^{pa}) \xrightarrow{1\otimes\varphi} k_\mathfrak{q}[\![u]\!]/(u^a) \xrightarrow{\psi} k_\mathfrak{q}[\![u]\!]/(u^{pa})$$

whose composite is equal to $u^{eb} \cdot \mathrm{Id}$. Let $(1 \otimes \varphi)(1) = f(u) \in k_\mathfrak{q}[\![u]\!]/(u^a)$, and $\psi(1) = h(u) \in k_\mathfrak{q}[\![u]\!]/(u^{pa})$. Then $u^{pa} \mid u^a h(u)$, so $u^{(p-1)a} \mid h(u)$. On the other hand, $f(u)h(u) = u^{eb}$ in $k_\mathfrak{q}[\![u]\!]/(u^{pa})$. This implies $u^{(p-1)a} \mid u^{eb}$. But $e < p - 1$ and $a \geq b$, so we get a contradiction. Hence, $\overline{\mathfrak{N}}_{n,\mathrm{tor}} = 0$ and $\overline{\mathfrak{N}}_n$ is free over $k_\mathfrak{q}[\![u]\!]$ of rank d, since by Lemma 4.1 $\overline{\mathfrak{N}}_n\left[\frac{1}{u}\right] \cong (\mathcal{M}_n \otimes_\mathfrak{S} \mathfrak{S}_\mathfrak{q}) \otimes_{\mathfrak{S}_\mathfrak{q}} k_\mathfrak{q}[\![u]\!]$ which is projective over $k_\mathfrak{q}(\!(u)\!)$

of rank d. Let $b_1, \dots, b_d \in \mathfrak{N}_n$ be a lift of a basis elements of $\overline{\mathfrak{N}}_n$. By Nakayama's lemma, we have a surjection of $\mathfrak{S}_{\mathfrak{q},\mathfrak{n}}$-modules

$$f : \bigoplus_{i=1}^{d} \mathfrak{S}_{\mathfrak{q},n} \cdot e_i \twoheadrightarrow \mathfrak{N}_n$$

given by $e_i \mapsto b_i$. Since $\mathfrak{N}_n\big[\frac{1}{u}\big] \cong \mathcal{M}_n \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathfrak{q}}$ is projective over $\mathfrak{S}_{\mathfrak{q},n}\big[\frac{1}{u}\big]$ of rank d, f is also injective. Thus, $\mathfrak{N}_n = \mathfrak{M}_n \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathfrak{q}}$ is projective over $\mathfrak{S}_{\mathfrak{q},n}$ of rank d. Since this holds for every $\mathfrak{q} \in \mathrm{mSpec}\, R_0$, it proves the assertion. □

Lemma 4.4. *Assume $e < p-1$. Let $\mathfrak{N}$ and $\mathfrak{N}'$ be finite u-torsion free $\mathfrak{S}$-modules equipped with Frobenius endomorphisms such that $\mathfrak{N}\big[\frac{1}{u}\big]$ and $\mathfrak{N}'\big[\frac{1}{u}\big]$ are torsion étale φ-modules. Suppose that $\mathfrak{N}$ and $\mathfrak{N}'$ have $E(u)$-height ≤ 1 and $\mathfrak{N}\big[\frac{1}{u}\big] = \mathfrak{N}'\big[\frac{1}{u}\big]$ as étale φ-modules. Then $\mathfrak{N} = \mathfrak{N}'$.*

Proof. Consider $\mathfrak{N}$ and $\mathfrak{N}'$ as $\mathfrak{S}$-submodules of $\mathfrak{N}\big[\frac{1}{u}\big]$. Let $\mathfrak{L}$ be the cokernel of the embedding $\mathfrak{N} \hookrightarrow \mathfrak{N}+\mathfrak{N}'$ of $\mathfrak{S}$-modules. Note that $\mathfrak{S} \otimes_{\varphi,\mathfrak{S}} (\mathfrak{N}+\mathfrak{N}') \cong \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{N} + \mathfrak{S} \otimes_{\varphi,\mathfrak{S}} \mathfrak{N}'$ since $\varphi : \mathfrak{S} \to \mathfrak{S}$ is flat. Thus, $\mathfrak{N}+\mathfrak{N}'$ has $E(u)$-height ≤ 1, and $\mathfrak{L}$ has $E(u)$-height ≤ 1. Since $\mathfrak{L}\big[\frac{1}{u}\big] = 0$, we deduce similarly as in the proof of Proposition 4.3 that $\mathfrak{L} = 0$. So $\mathfrak{N} = \mathfrak{N} + \mathfrak{N}'$. Similarly, $\mathfrak{N}' = \mathfrak{N} + \mathfrak{N}'$. □

It is clear that both $p\mathfrak{M}_{n+1}$ and $\mathfrak{M}_n$ are u-torsion free, have $E(u)$-height ≤ 1 and

$$p\mathfrak{M}_{n+1}\Big[\frac{1}{u}\Big] = p\mathcal{M}_{n+1} \cong \mathcal{M}_n = \mathfrak{M}_n\Big[\frac{1}{u}\Big].$$

We conclude the following:

Proposition 4.5. *Assume $e < p-1$. For each $n \ge 1$, we have a φ-equivariant isomorphism*

$$p\mathfrak{M}_{n+1} \cong \mathfrak{M}_n.$$

By Lemma 4.2, Proposition 4.3 and 4.5, if we suppose $e < p-1$ and define the $\mathfrak{S}$-module

$$\mathfrak{M} := \varprojlim_n \mathfrak{M}_{\mathfrak{n}},$$

then $\mathfrak{M} \in \mathrm{Kis}^1(\mathfrak{S})$. Note that we have a φ-equivariant isomorphism $\mathfrak{M} \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathcal{O}_L} \cong \mathfrak{M}_{\mathcal{O}_L}$ by Lemma 4.1.

Remark 4.6. Analogous statements hold when T is a crystalline $\mathcal{G}_R$-representation with Hodge–Tate weights in $[0, r]$ for the case $er < p-1$, since [Brinon and Trihan 2008] constructs more generally a functor from crystalline representations with Hodge–Tate weights in $[0, r]$ to Kisin modules of height r when the base is a complete discrete valuation field whose residue field has a finite p-basis.

To study connections for $\mathfrak{M}$, we first consider the following general situation. Let A_0 be a k-algebra which is an integral domain. Consider n-variables $x_1, \dots, x_n$, and denote $\underline{x} = (x_1, \dots, x_n)^t$ and $\underline{x}^{[p]} := (x_1^p, \dots, x_n^p)^t$. An Artin–Schreier system of equations in n variables over A_0 is given by

$$\underline{x} = B\underline{x}^{[p]} + C, \tag{4-1}$$

where $B = (b_{ij})_{1\le i,j\le n} \in M_{n\times n}(A_0)$ is an $n \times n$ matrix with entries in A_0 and $C = (c_i)_{1\le i\le n} \in M_{n\times 1}(A_0)$.

Let

$$A_1 := A_0[x_1, \dots, x_n]/\Big(x_1 - c_1 - \sum_{i=1}^{n} b_{1i} x_i^p, \dots, x_n - c_n - \sum_{i=1}^{n} b_{ni} x_i^p\Big),$$

which is the A_0-algebra parametrizing the solutions of (4-1). $A_0 \to A_1$ is étale by [Vasiu 2013, Theorem 2.4.1(a)].

Lemma 4.7. *There exists a nonzero element $f \in A_0$ which depends only on B (and not on C) such that $A_1\big[\frac{1}{f}\big]$ is finite étale over $A_0\big[\frac{1}{f}\big]$.*

Proof. We induct on n. Suppose $n = 1$. If $\det B \neq 0$, then (4-1) is equivalent to

$$x_1^p = B^{-1} x_1 - B^{-1} C,$$

so the assertion holds with $f = \det B$. If $\det B = 0$, then $B = 0$ and $A_1 \cong A_0$, so the assertion holds trivially.

For $n \geq 2$, if $\det B \neq 0$, then (4-1) is equivalent to

$$\underline{x}^{[p]} = B^{-1} \underline{x} - B^{-1} C.$$

Hence, with $f = \det B$, $A_1\big[\frac{1}{f}\big]$ is finite étale over $A_0\big[\frac{1}{f}\big]$. Suppose $\det B = 0$. Denote by $B^{(i)}$ the i-th row of B. Then up to renumbering the index for x_i's, we have

$$\sum_{i=1}^{n} e_i B^{(i)} = 0$$

for some nonzero $f_1 \in A_0$ depending only on B and some $e_i \in A_0\big[\frac{1}{f_1}\big]$ with $e_n = 1$. From (4-1), we get

$$x_n = -\sum_{i=1}^{n-1} e_i x_i + c_n + \sum_{i=1}^{n-1} c_i e_i.$$

Hence, denoting $\underline{x}' = (x_1, \dots, x_{n-1})^t$, (4-1) is equivalent to an Artin–Schreier system of equations in $n-1$ variables over $A_0\big[\frac{1}{f_1}\big]$

$$\underline{x}' = B' \underline{x}'^{[p]} + C'$$

where $B' \in M_{(n-1)\times(n-1)}\big(A_0\big[\frac{1}{f_1}\big]\big)$ and $C' \in M_{(n-1)\times 1}\big(A_0\big[\frac{1}{f_1}\big]\big)$. Note that B' depends only on B and not on C. Hence, the assertion follows by induction. □

Let $\mathcal{N} := \mathfrak{M} \otimes_{\mathfrak{S},\varphi} R_0$ equipped with the Frobenius $\varphi_{\mathfrak{M}} \otimes \varphi_{R_0}$. From [Kim 2015, Equations (6.1), (6.2) and Remark 3.13], we have the R_0-submodule $\mathrm{Fil}^1\mathcal{N} \subset \mathcal{N}$ associated with $\mathfrak{M} \in \mathrm{Kis}^1(\mathfrak{S})$ such that $p\mathcal{N} \subset \mathrm{Fil}^1\mathcal{N}$, $\mathcal{N}/\mathrm{Fil}^1\mathcal{N}$ is projective over $R_0/(p)$, and $(1\otimes\varphi)(\varphi^*\mathrm{Fil}^1\mathcal{N}) = p\mathcal{N}$ as R_0-modules (see [Kim 2015, Definitions 3.4 and 3.6] for the frame $(R_0, pR_0, R_0/(p), \varphi_{R_0}, \varphi_{R_0}/p)$). Fix an R_0-direct factor $\mathcal{N}^1 \subset \mathcal{N}$ which lifts $\mathrm{Fil}^1\mathcal{N}/p\mathcal{N} \subset \mathcal{N}/p\mathcal{N}$, and let

$$\widetilde{\mathcal{N}} := R_0 \otimes_{\varphi,R_0} \Big(\mathcal{N} + \frac{1}{p}\mathcal{N}^1\Big) \subset R_0\Big[\frac{1}{p}\Big] \otimes_{\varphi,R_0} \mathcal{N}.$$

Let $\mathrm{Spf}(A, p) \to \mathrm{Spf}(R_0, p)$ be an étale morphism. Note that A is equipped with a unique Frobenius lifting that on R_0, and $\widehat{\Omega}_A \cong A\widehat{\otimes}_{R_0}\widehat{\Omega}_{R_0} \cong \bigoplus_{i=1}^m A \cdot dt_i$. For a connection

$$\nabla_{A,n} : A/(p^n) \otimes_{R_0} \mathcal{N} \to (A/(p^n) \otimes_{R_0} \mathcal{N}) \otimes_A \widehat{\Omega}_A$$

on $A/(p^n) \otimes_{R_0} \mathcal{N}$, we say that the Frobenius is *horizontal* if the following diagram commutes:

$$\begin{array}{ccc} A/(p^n) \otimes_A \widetilde{\mathcal{N}} & \xrightarrow{\varphi^*(\nabla_{A,n})} & A/(p^n) \otimes_A \widetilde{\mathcal{N}} \otimes_A \widehat{\Omega}_A \\ \downarrow{\scriptstyle 1\otimes\varphi} & & \downarrow{\scriptstyle (1\otimes\varphi)\otimes \mathrm{id}_{\widehat{\Omega}_A}} \\ A/(p^n) \otimes_A \mathcal{N} & \xrightarrow{\nabla_{A,n}} & A/(p^n) \otimes_A \mathcal{N} \otimes_A \widehat{\Omega}_A \end{array}$$

Here, $\varphi^*(\nabla_{A,n})$ is given by choosing an arbitrary lift of $\nabla_{A,n}$ on $A/(p^{n+1}) \otimes_A \mathcal{N}$, and $\varphi^*(\nabla_{A,n})$ does not depend on the choice of such a lift (see [Vasiu 2013, Section 3.1.1, Equation (9)]).

Proposition 4.8. *There exists $\tilde{f} \in R_0$ with $\tilde{f} \notin pR_0$ such that the following holds*:

Let S_0 be the p-adic completion of $R_0\big[\frac{1}{\tilde{f}}\big]$ equipped with the induced Frobenius, and let $\mathfrak{S}_S = S_0[\![u]\!]$. Let $\mathfrak{M}_S = \mathfrak{M} \otimes_{\mathfrak{S}} \mathfrak{S}_S$ equipped with the induced Frobenius, so $\mathfrak{M}_S \in \mathrm{Kis}^1(\mathfrak{S}_S)$. Then there exists a topologically quasinilpotent integrable connection

$$\nabla_{\mathfrak{M}_S} : (S_0 \otimes_{\varphi,\mathfrak{S}_S} \mathfrak{M}_S) \to (S_0 \otimes_{\varphi,\mathfrak{S}_S} \mathfrak{M}_S) \otimes_{S_0} \widehat{\Omega}_{S_0}$$

such that φ is horizontal, and thus $(\mathfrak{M}_S, \nabla_{\mathfrak{M}_S}) \in \mathrm{Kis}^1(\mathfrak{S}_S, \nabla)$. Furthermore, we can choose $\nabla_{\mathfrak{M}_S}$ so that $\mathfrak{M}_S \otimes_{\mathfrak{S}_S} \mathfrak{S}_{\mathcal{O}_L}$ equipped with the induced Frobenius and connection is isomorphic to $(\mathfrak{M}_{\mathcal{O}_L}, \nabla_{\mathfrak{M}_{\mathcal{O}_L}})$ as Kisin modules over $\mathfrak{S}_{\mathcal{O}_L}$.

Proof. Without loss of generality, we may pass to a Zariski open set of $\mathrm{Spf}(R_0, p)$ if necessary so that $\mathcal{N}^1$ and $\mathcal{N}/\mathcal{N}^1$ are free over R_0. Fix an R_0-basis of $\mathcal{N}$ adapted to the direct factor $\mathcal{N}^1$. Let $\mathrm{Spf}(A, p) \to \mathrm{Spf}(R_0, p)$ be an étale morphism. Consider a connection

$$\nabla_{A,1} : A/(p) \otimes_{R_0} \mathcal{N} \to (A/(p) \otimes_{R_0} \mathcal{N}) \otimes_A \widehat{\Omega}_A$$

such that the Frobenius is horizontal. By [Vasiu 2013, Section 3.2, Basic Theorem] and its proof, the set of such connections $\nabla_{A,1}$ corresponds to the set of solutions over $A/(p)$ of an Artin–Schreier system of equations

$$\underline{x} = B\underline{x}^{[p]} + C_1$$

for $\underline{x} = (x_1, \ldots, x_{dm})^t$, where $B \in M_{dm\times dm}(R_0/(p))$ and $C_1 \in M_{dm\times 1}(R_0/(p))$. When $A = \mathcal{O}_{L_0}$, it has a solution given by $\nabla_{\mathfrak{M}_{L_0}}$. Since $\mathcal{O}_{L_0}/(p) \cong \mathrm{Frac}(R_0/(p))$ and $R_0/(p)$ is a unique factorization domain, the solution lies in $(R_0/(p))\big[\frac{1}{f}\big]$ for some nonzero $f \in R_0/(p)$ depending only on B by Lemma 4.7 and its proof. Let $\tilde{f} \in R_0$ be a lift of f, and let S_0 be the p-adic completion of $R_0\big[\frac{1}{\tilde{f}}\big]$.

For $n \geq 1$, suppose we are given a connection

$$\nabla_{S_0,n} : S_0/(p^n) \otimes_{R_0} \mathcal{N} \to (S_0/(p^n) \otimes_{R_0} \mathcal{N}) \otimes_{S_0} \widehat{\Omega}_{S_0}$$

such that the Frobenius is horizontal and inducing $\nabla_{\mathfrak{M}_{L_0}} (\mathrm{mod}\ p^n)$ via the natural map $S_0 \to \mathcal{O}_{L_0}$. By [Vasiu 2013, Section 3.2, Basic Theorem] and its proof, for the choice of a basis of $\mathcal{N}$ as above, the set of connections

$$\nabla_{S_0,n+1} : S_0/(p^{n+1}) \otimes_{R_0} \mathcal{N} \to (S_0/(p^{n+1}) \otimes_{R_0} \mathcal{N}) \otimes_{S_0} \widehat{\Omega}_{S_0}$$

such that the Frobenius is horizontal and lifting $\nabla_{S_0,n}$ corresponds to the set of solutions over $S_0/(p)$ of an Artin–Schreier system of equations

$$\underline{x} = B\underline{x}^{[p]} + C_{n+1},$$

where B is the same matrix as above and $C_{n+1} \in M_{dm\times 1}(S_0/(p))$. The solution over $\mathcal{O}_{L_0}/(p)$ given by $\nabla_{\mathfrak{M}_{L_0}}$ lies in $S_0/(p)$ by Lemma 4.7 and its proof. This proves the assertion. □

Proposition 4.9. *Let S_0 be a ring as given in Proposition 4.8, and let $S = S_0 \otimes_{W(k)} \mathcal{O}_K$. Then there exists a p-divisible group G_S over S such that $T_p(G_S) \cong T$ as $\mathcal{G}_S$-representations.*

Proof. Let G_S be the p-divisible group over S given by $(\mathfrak{M}_S, \nabla_{\mathfrak{M}_S})$ in Proposition 4.8. Since $\mathfrak{M}_S \otimes_{\mathfrak{S}_S} \mathfrak{S}_{\mathcal{O}_L} \cong \mathfrak{M}_{\mathcal{O}_L}$ as Kisin modules, we have $T_p(G_S) \cong T$ as $\mathcal{G}_{\mathcal{O}_L}$-representations. On the other hand, $\mathfrak{M}_S \otimes_{\mathfrak{S}_S} \mathcal{O}_{\mathcal{E},S} \cong \mathcal{M} \otimes_{\mathcal{O}_\mathcal{E}} \mathcal{O}_{\mathcal{E},S}$ as étale φ-modules. Hence, $T_p(G_S) \cong T$ as $\mathcal{G}_{\widetilde{S},\infty}$-representations. Since $\mathcal{G}_{\widetilde{S},\infty}$ and $\mathcal{G}_{\mathcal{O}_L}$ generate the Galois group $\mathcal{G}_S$ by Lemma 2.1, we have $T_p(G_S) \cong T$ as $\mathcal{G}_S$-representations. □

5. Proof of the main theorem

In this section, we finish the proof of Theorem 1.2. We begin by recalling the following well-known lemma about p-divisible groups.

Lemma 5.1. *Let R_1 be an integral domain over $W(k)$ such that* $\mathrm{Frac}(R_1)$ *has characteristic* 0*. Then via the Tate module functor $T_p(\cdot)$, the category of p-divisible groups over $R_1\big[\frac{1}{p}\big]$ is equivalent to the category of finite free $\mathbb{Z}_p$-representations of $\mathcal{G}_{R_1} = \pi_1^{\text{ét}}\big(\mathrm{Spec}\ R_1\big[\frac{1}{p}\big]\big)$. Furthermore, such an equivalence is functorial in the following sense*:

Let $R_1 \to R_2$ be a map of integral domains over $W(k)$ such that $\mathrm{Frac}(R_1)$ *and* $\mathrm{Frac}(R_2)$ *have characteristic* 0*. Let G_{R_1} be a p-divisible group over R_1. Then $T_p(G_{R_1}) \cong T_p(G_{R_1} \times_{R_1} R_2)$ as $\mathcal{G}_{R_2}$-representations.*

We first consider the case when R is a formal power series ring of dimension 2.

Proposition 5.2. *Suppose $R_0 = \mathcal{O}[\![s]\!]$ for a Cohen ring $\mathcal{O}$, and $e \leq p-1$. Let T be a crystalline $\mathcal{G}_R$-representation which is finite free over $\mathbb{Z}_p$ and has Hodge–Tate weights in* $[0,1]$*. Then there exists a p-divisible group G_R over R such that $T_p(G_R) \cong T$ as $\mathcal{G}_R$-representations.*

Proof. Let G be a p-divisible group over $R\big[\frac{1}{p}\big]$ given by Lemma 5.1 such that $T_p(G) \cong T$ as $\mathcal{G}_R$-representations. It suffices to show that G extends to a p-divisible group G_R over R.

By [Brinon and Trihan 2008, Theorem 6.10], there exists a p-divisible group $G_{\mathcal{O}_L}$ over $\mathcal{O}_L$ extending $G \times_{R[\frac{1}{p}]} L$. For each integer $n \geq 1$, let A_n be the Hopf algebra over $R\big[\frac{1}{s}\big]\big[\frac{1}{p}\big]$ for the finite flat group scheme $\big(G \times_{R[\frac{1}{p}]} R\big[\frac{1}{s}\big]\big[\frac{1}{p}\big]\big)[p^n]$, and let B_n be the Hopf algebra over $\mathcal{O}_L$ for the finite flat group scheme $G_{\mathcal{O}_L}[p^n]$. Identify $A_n \otimes_{R[\frac{1}{s}][\frac{1}{p}]} L = B_n \otimes_{\mathcal{O}_L} L$ as Hopf algebras over L. Note that the p-adic completion of $R\big[\frac{1}{s}\big]$ is isomorphic to $\mathcal{O}_L$. By [Beauville and Laszlo 1995, Main Theorem] and its proof, the $R\big[\frac{1}{s}\big]$-subalgebra $C_n := A_n \cap B_n \subset B_n \otimes_{\mathcal{O}_L} L$ is finite flat over $R\big[\frac{1}{s}\big]$. Moreover, C_n is equipped with the Hopf algebra structure induced from (A_n, B_n) such that $C_n \otimes_{R[\frac{1}{s}]} R\big[\frac{1}{s}\big]\big[\frac{1}{p}\big] \cong A_n$ and $C_n \otimes_{R[\frac{1}{s}]} \mathcal{O}_L \cong B_n$. Hence, the datum of finite flat group schemes

$$\Big(\Big(G \times_{R[\frac{1}{p}]} R\Big[\frac{1}{s}\Big]\Big[\frac{1}{p}\Big]\Big)[p^n],\, G_{\mathcal{O}_L}[p^n]\Big)$$

descends to a finite flat group scheme over $R\big[\frac{1}{s}\big]$ (up to a unique isomorphism by [Beauville and Laszlo 1995, Main Theorem]).

Thus, we obtain a system of finite flat group schemes $(G_{U,n})_{n\geq 1}$ over $U := \operatorname{Spec} R \setminus \mathrm{pt}$ extending $(G[p^n])_{n\geq 1}$. Here, pt denotes the closed point given by the maximal ideal of R. The natural induced sequence of finite flat group schemes

$$0 \to G_{U,1} \to G_{U,n+1} \xrightarrow{\times p} G_{U,n} \to 0$$

is exact by fpqc descent. So $(G_{U,n})_{n\geq 1}$ is a p-divisible group over U extending G. Since $e \leq p-1$, G_U extends to a p-divisible group G_R over R by [Vasiu and Zink 2010, Theorem 3]. □

Remark 5.3. As illustrated in the above proof, this special case can be shown without using Kisin modules. However, the purity result for p-divisible groups [Vasiu and Zink 2010, Theorem 3] is proved only when R is regular local of Krull dimension 2 with low ramification (see [Vasiu and Zink 2010, Section 5.1]). So we use the construction of Kisin modules to show Theorem 1.2 for more general R with arbitrary dimensions.

Now, let R_0 be a general ring satisfying the assumptions in Section 2A, and let $R = R_0 \otimes_{W(k)} \mathcal{O}_K$ with $e < p-1$. Let T be a crystalline $\mathcal{G}_R$-representation free over $\mathbb{Z}_p$ with Hodge–Tate weights in $[0, 1]$. Denote by $\mathfrak{M}_{\mathfrak{S}}(T)$ the $\mathfrak{S}$-module in $\mathrm{Kis}^1(\mathfrak{S})$ constructed from T as in Section 4. Let $\tilde{f} \in R_0$ be an element as in Proposition 4.8, and let S_0 be the p-adic completion of $R_0\big[\frac{1}{\tilde{f}}\big]$ as in Proposition 4.9. Let $f \in R_0/pR_0$ be the image of $\tilde{f}$ in the projection $R_0 \to R_0/(p)$. Note that if f is a unit in R_0/pR_0, then $\tilde{f}$ is a unit in R_0 since R_0 is p-adically complete. So for such a case, $S_0 = R_0$ and Theorem 1.2 follows from Proposition 4.9. Now consider the case when f is not a unit in $R_0/(p)$. Since $R_0/(p)$ is a UFD, there exist prime elements $\bar{s}_1, \ldots, \bar{s}_l$ of $R_0/(p)$ dividing f. Let $s_1, \ldots, s_l \in R_0$ be any preimages of $\bar{s}_1, \ldots, \bar{s}_l$ respectively.

For each $i = 1, \ldots, l$, consider the prime ideal $\mathfrak{p}_i = (p, s_i) \subset R_0$ and let $R_0^{(i)} := \widehat{R_{0,\mathfrak{p}_i}}$ be the $\mathfrak{p}_i$-adic completion of $R_{0,\mathfrak{p}_i}$. Note that $R_0^{(i)}$ is a formal power series ring over a Cohen ring with Krull dimension 2. We consider the natural φ-equivariant map $b_i : R_0 \to R_0^{(i)}$, which induces $b_i : R \to R^{(i)} := R_0^{(i)} \otimes_{W(k)} \mathcal{O}_K$. On the other hand, let k_c be a field extension of $\mathrm{Frac}(R_0/pR_0)$ which is a composite of the fields

$\mathrm{Frac}(R_0^{(i)}/(p))$ for $i = 1, \dots, l$, and let $k_c^{\mathrm{perf}} = \varinjlim_{\varphi} k_c$ be its direct perfection. By the universal property of p-adic Witt vectors, there exists a unique φ-equivariant map $b_c : R_0 \to W(k_c^{\mathrm{perf}})$. Moreover, for each $i = 1, \dots, l$, we have a unique φ-equivariant embedding $R_0^{(i)} \to W(k_c^{\mathrm{perf}})$ whose composite with b_i is equal to b_c. We claim that $S_0/(p) \cap \bigcap_{i=1}^{l}(R_0^{(i)}/(p)) = R_0/(p)$ inside k_c^{perf}. To see the claim, let x be a nonzero element of $S_0/(p) = (R_0/(p))\left[\frac{1}{\bar{f}}\right]$ such that it also lies in $\bigcap_{i=1}^{l}(R_0^{(i)}/(p))$. We can write $x = \left(\prod_{i=1}^{l} \bar{s}_i^{n_i}\right) \cdot a$ for some integers n_i and some $a \in R_0/(p)$ which is not in the ideal $(\bar{s}_1, \dots, \bar{s}_l)$ of $R_0/(p)$. Suppose $n_1 < 0$. Then $(1/\bar{s}_1^{n_1})x$ lies in the maximal ideal of $R_0^{(1)}/(p)$. But $(1/\bar{s}_1^{n_1})x = \left(\prod_{i=2}^{l} \bar{s}_i^{n_i}\right) \cdot a$ is a unit in $R_0^{(1)}/(p)$, which is a contradiction. So $n_1 \geq 0$, and similarly $n_i \geq 0$ for each $i = 1, \dots, l$. So $x \in R_0/(p)$, which proves the claim. This implies that the natural embedding $R_0 \to S_0 \cap \bigcap_{i=1}^{l} R_0^{(i)}$ as subrings of $W(k_c^{\mathrm{perf}})$ is bijective.

By Proposition 5.2, there exists a p-divisible group G_i over $R^{(i)}$ such that $T_p(G_i) \cong T$ as $\mathcal{G}_{R^{(i)}}$-representations. We have

$$(\mathfrak{M}_{\mathfrak{S}}(T) \otimes_{\mathfrak{S}} \mathcal{O}_{\mathcal{E}}) \otimes_{\mathcal{O}_{\mathcal{E}}} \mathcal{O}_{\mathcal{E},R^{(i)}} \cong \mathfrak{M}^*(G_i) \otimes_{\mathfrak{S}_{R^{(i)}}} \mathcal{O}_{\mathcal{E},R^{(i)}}$$

as étale $(\varphi, \mathcal{O}_{\mathcal{E},R^{(i)}})$-modules. Applying Lemma 4.4, we can deduce that $\mathfrak{M}_{\mathfrak{S}}(T) \otimes_{\mathfrak{S}} \mathfrak{S}_{R^{(i)}} \cong \mathfrak{M}^*(G_i)$ compatibly with Frobenius.

Let $D = D_{\mathrm{cris}}\left(T\left[\frac{1}{p}\right]\right)$, and denote $\mathfrak{M} = \mathfrak{M}_{\mathfrak{S}}(T)$ and $\mathcal{N} = \mathfrak{M} \otimes_{\mathfrak{S},\varphi} R_0$. Let $\nabla : D \to D \otimes_{R_0} \widehat{\Omega}_{R_0}$ be the connection given by the functor $D_{\mathrm{cris}}(\cdot)$.

Proposition 5.4. *There exists a natural φ-equivariant embedding*

$$h : \mathcal{N} \hookrightarrow D$$

of R_0-modules. Furthermore, if we consider $\mathcal{N}$ as an R_0-submodule of D via h, then ∇ maps $\mathcal{N}$ into $\mathcal{N} \otimes_{R_0} \widehat{\Omega}_{R_0}$. Hence, $\mathfrak{M}$ is a Kisin module of height 1.

Proof. By [Kim 2015, Corollaries 5.3 and 6.7], there exists a natural φ-equivariant embedding

$$h_i : \mathcal{N} \to D \otimes_{R_0,b_i} R_0^{(i)}$$

for each $i = 1, \dots, l$ such that the connections given by $\mathfrak{M}^*(G_i)$ and D are compatible, and there exists a natural φ-equivariant embedding $h_c : \mathcal{N} \to D \otimes_{R_0,b_c} W(k_c^{\mathrm{perf}})$. Moreover, by Proposition 4.9, there exists a natural φ-equivariant embedding $h_S : \mathcal{N} \to D \otimes_{R_0} S_0$ such that the connections given by $\mathfrak{M}^*(G_S)$ and D are compatible. Since the construction of those natural maps is compatible with φ-equivariant base changes (see [Kim 2015, Section 5.5]), we deduce that the maps $h_1, \dots, h_l$ and h_S are compatible with one another, in the sense that their composites with the embedding into $D \otimes_{R_0,b_c} W(k_c^{\mathrm{perf}})$ are all equal to h_c. Hence, we obtain a φ-equivariant embedding

$$h : \mathcal{N} \hookrightarrow \left(D \otimes_{R_0\left[\frac{1}{p}\right]} S_0\left[\frac{1}{p}\right]\right) \cap \left(\bigcap_{i=1}^{l} D \otimes_{R_0\left[\frac{1}{p}\right],b_i} R_0^{(i)}\left[\frac{1}{p}\right]\right) \cong D \otimes_{R_0\left[\frac{1}{p}\right]} \left(S_0\left[\frac{1}{p}\right] \cap \bigcap_{i=1}^{l} R_0^{(i)}\left[\frac{1}{p}\right]\right) = D,$$

since D is flat over $R_0\left[\frac{1}{p}\right]$ and $S_0\left[\frac{1}{p}\right] \cap \bigcap_{i=1}^{l} R_0^{(i)}\left[\frac{1}{p}\right] = R_0\left[\frac{1}{p}\right]$.

Now, identify $\widehat{\Omega}_{R_0} = \bigoplus_{j=1}^m R_0 \cdot dt_j$. Then ∇ maps $\mathcal{N}$ to $\mathcal{N} \otimes_{R_0} \big(\bigoplus_{j=1}^m R_0\big[\frac{1}{p}\big] \cdot dt_j\big)$. On the other hand, by Propositions 4.8, 5.2, and the compatibility of $D_{\mathrm{cris}}(\cdot)$ with respect to φ-compatible base changes, we have that ∇ maps $\mathcal{N}$ into $\mathcal{N} \otimes_{R_0} \big(\bigoplus_{j=1}^m S_0 \cdot dt_j\big)$ and also into $\mathcal{N} \otimes_{R_0} \big(\bigoplus_{j=1}^m R_0^{(i)} \cdot dt_j\big)$ for each $i = 1, \ldots, l$. Since $\mathcal{N}$ is flat over R_0 and $S_0 \cap \bigcap_{i=1}^l R_0^{(i)} = R_0$, ∇ maps $\mathcal{N}$ into $\mathcal{N} \otimes_{R_0} \big(\bigoplus_{j=1}^m R_0 \cdot dt_j\big)$. □

Theorem 5.5. *There exists a p-divisible group G_R over R such that $T_p(G_R) \cong T$ as $\mathcal{G}_R$-representations.*

Proof. By Proposition 5.4, we have $\mathfrak{M} \in \mathrm{Kis}^1(\varphi, \nabla)$. Furthermore, $\mathfrak{M} \otimes_{\mathfrak{S}} \mathfrak{S}_{\mathcal{O}_L} \cong \mathfrak{M}_{\mathcal{O}_L}$ as Kisin modules over $\mathfrak{S}_{\mathcal{O}_L}$, since the Frobenius and connection structure on $\mathfrak{M}$ agree with those on D. Thus, if G_R is the p-divisible group corresponding to $\mathfrak{M}$, then $T_p(G_R) \cong T$ as $\mathcal{G}_{\mathcal{O}_L}$-representations as well as $\mathcal{G}_{\widetilde{R}_\infty}$-representations. The assertion follows from Lemma 2.1. □

6. Barsotti–Tate deformation ring

As an application of Theorem 5.5, we study the geometry of the locus of crystalline representations with Hodge–Tate weights in [0, 1] by using the results in [Moon 2020]. Note that in [Moon 2020, Section 2], R_0 is assumed to satisfy the additional conditions that $W(k)\langle t_1^{\pm 1}, \ldots, t_d^{\pm 1}\rangle \to R_0$ has geometrically regular fibers or R_0 has Krull dimension less than 2, and that $k \to R_0/pR_0$ is geometrically integral. These assumptions are only used to have the crystalline period ring as in [Brinon 2008]. However, the additional conditions are not necessary by [Kim 2015, Section 4], and the results in [Moon 2020] hold in our setting.

Denote by $\mathcal{C}$ the category of topological local $\mathbb{Z}_p$-algebras A satisfying the following conditions:

- The natural map $\mathbb{Z}_p \to A/\mathfrak{m}_A$ is surjective, where $\mathfrak{m}_A$ denotes the maximal ideal of A.
- The map from A to the projective limit of its discrete artinian quotients is a topological isomorphism.

By the first condition, the residue field of A is $\mathbb{F}_p$. The second condition is equivalent to that A is complete and its topology is given by a collection of open ideals $\mathfrak{a} \subset A$ for which $A/\mathfrak{a}$ is artinian. Morphisms in $\mathcal{C}$ are continuous $\mathbb{Z}_p$-algebra morphisms.

For $A \in \mathcal{C}$, we mean by an *A-representation of $\mathcal{G}_R$* a finite free A-module equipped with a continuous A-linear $\mathcal{G}_R$-action. Fix an $\mathbb{F}_p$-representation V_0 of $\mathcal{G}_R$ which is absolutely irreducible. For $A \in \mathcal{C}$, a *deformation* of V_0 in A is defined to be an isomorphism class of A-representations of V of $\mathcal{G}_R$ satisfying $V \otimes_A \mathbb{F}_p \cong V_0$ as $\mathbb{F}_p[\mathcal{G}_R]$-modules. Denote by $\mathrm{Def}(V_0, A)$ the set of such deformations. A morphism $f : A \to A'$ in $\mathcal{C}$ induces a map $f_* : \mathrm{Def}(V_0, A) \to \mathrm{Def}(V_0, A')$ sending the class of an A-representation V to the class of $V \otimes_{A,f} A'$. The following theorem on universal deformation ring is proved in [de Smit and Lenstra 1997].

Theorem 6.1 [de Smit and Lenstra 1997, Theorem 2.3]. *There exists a universal deformation ring $A_{\mathrm{univ}} \in \mathcal{C}$ and a deformation $V_{\mathrm{univ}} \in \mathrm{Def}(V_0, A_{\mathrm{univ}})$ such that for all $A \in \mathcal{C}$, we have a bijection*

$$\mathrm{Hom}_{\mathcal{C}}(A_{\mathrm{univ}}, A) \xrightarrow{\cong} \mathrm{Def}(V_0, A) \tag{6-1}$$

given by $f \mapsto f_*(V_{\mathrm{univ}})$.

We deduce that when R has dimension 2 and e is small, the locus of crystalline representations with Hodge–Tate weights in [0, 1] cuts out a closed subscheme of Spec A_{univ} in the following sense.

Theorem 6.2. *Suppose that $e < p-1$ and that the Krull dimension of R is 2. Then there exists a closed ideal $\mathfrak{a}_{\mathrm{BT}} \subset A_{\mathrm{univ}}$ such that the following holds*:

For any finite flat $\mathbb{Z}_p$-algebra A equipped with the p-adic topology and any continuous $\mathbb{Z}_p$-algebra map $f : A_{\mathrm{univ}} \to A$, the induced representation $V_{\mathrm{univ}} \otimes_{A_{\mathrm{univ}}, f} A\left[\frac{1}{p}\right]$ of $\mathcal{G}_R$ is crystalline with Hodge–Tate weights in $[0, 1]$ if and only if f factors through the quotient $A_{\mathrm{univ}}/\mathfrak{a}_{\mathrm{BT}}$.

Proof. This follows directly from Theorem 5.5 and [Moon 2020, Theorem 5.7]. Note that [Moon 2020, Theorem 5.7] assumes the Krull dimension of R is 2. The assumption was necessary in the argument of [Moon 2020, Section 5] to construct Barsotti–Tate deformation ring using the result in [de Smit and Lenstra 1997]. □

References

[Andreatta 2006] F. Andreatta, "Generalized ring of norms and generalized (ϕ, Γ)-modules", *Ann. Sci. École Norm. Sup.* (4) **39**:4 (2006), 599–647. MR Zbl

[Beauville and Laszlo 1995] A. Beauville and Y. Laszlo, "Un lemme de descente", *C. R. Acad. Sci. Paris Sér. I Math.* **320**:3 (1995), 335–340. MR Zbl

[Brinon 2008] O. Brinon, "Représentations p-adiques cristallines et de de Rham dans le cas relatif", *Mém. Soc. Math. Fr.* (*N.S.*) 112 (2008), vi+159. MR Zbl

[Brinon and Trihan 2008] O. Brinon and F. Trihan, "Représentations cristallines et F-cristaux: le cas d'un corps résiduel imparfait", *Rend. Semin. Mat. Univ. Padova* **119** (2008), 141–171. MR Zbl

[Gabber and Ramero 2003] O. Gabber and L. Ramero, *Almost ring theory*, Lecture Notes in Math. **1800**, Springer, 2003. MR Zbl

[Hartl 2013] U. Hartl, "On a conjecture of Rapoport and Zink", *Invent. Math.* **193**:3 (2013), 627–696. MR Zbl

[de Jong 1995] A. J. de Jong, "Crystalline Dieudonné module theory via formal and rigid geometry", *Inst. Hautes Études Sci. Publ. Math.* 82 (1995), 5–96. MR Zbl

[Katz 1973] N. M. Katz, "p-adic properties of modular schemes and modular forms", pp. 69–190 in *Modular functions of one variable, III* (Antwerp, 1972), edited by W. Kuyk and J.-P. Serre, Lecture Notes in Math. **350**, Springer, 1973. MR Zbl

[Kim 2015] W. Kim, "The relative Breuil–Kisin classification of p-divisible groups and finite flat group schemes", *Int. Math. Res. Not.* **2015**:17 (2015), 8152–8232. MR Zbl

[Kisin 2006] M. Kisin, "Crystalline representations and F-crystals", pp. 459–496 in *Algebraic geometry and number theory*, edited by V. Ginzburg, Progr. Math. **253**, Birkhäuser, Boston, 2006. MR Zbl

[Laffaille 1980] G. Laffaille, "Groupes p-divisibles et modules filtrés: le cas peu ramifié", *Bull. Soc. Math. France* **108**:2 (1980), 187–206. MR Zbl

[Moon 2020] Y. S. Moon, "Extending p-divisible groups and Barsotti–Tate deformation ring in the relative case", *Int. Math. Res. Not.* (2020).

[Scholze 2012] P. Scholze, "Perfectoid spaces", *Publ. Math. Inst. Hautes Études Sci.* **116** (2012), 245–313. MR Zbl

[de Smit and Lenstra 1997] B. de Smit and H. W. Lenstra, Jr., "Explicit construction of universal deformation rings", pp. 313–326 in *Modular forms and Fermat's last theorem* (Boston, MA, 1995), edited by G. Cornell et al., Springer, 1997. MR Zbl

[Vasiu 2013] A. Vasiu, "A motivic conjecture of Milne", *J. Reine Angew. Math.* **685** (2013), 181–247. MR Zbl

[Vasiu and Zink 2010] A. Vasiu and T. Zink, "Purity results for p-divisible groups and abelian schemes over regular bases of mixed characteristic", *Doc. Math.* **15** (2010), 571–599. MR Zbl

Communicated by Kiran S. Kedlaya
Received 2019-10-02 Revised 2020-05-14 Accepted 2020-06-24

tongliu@math.purdue.edu *Purdue University, West Lafayette, IN, United States*

ysmoon1@math.arizona.edu *University of Arizona, Tucson, AZ, United States*

msp

ALGEBRA AND NUMBER THEORY 14:10 (2020)
https://doi.org/10.2140/ant.2020.14.2791

Algorithms for orbit closure separation for invariants and semi-invariants of matrices

Harm Derksen and Visu Makam

We consider two group actions on m-tuples of $n \times n$ matrices with entries in the field K. The first is simultaneous conjugation by GL_n and the second is the left-right action of $\mathrm{SL}_n \times \mathrm{SL}_n$. Let $\overline{K}$ be the algebraic closure of the field K. Recently, a polynomial time algorithm was found to decide whether 0 lies in the Zariski closure of the $\mathrm{SL}_n(\overline{K}) \times \mathrm{SL}_n(\overline{K})$-orbit of a given m-tuple by Garg, Gurvits, Oliveira and Wigderson for the base field $K = \mathbb{Q}$. An algorithm that also works for finite fields of large enough cardinality was given by Ivanyos, Qiao and Subrahmanyam. A more general problem is the *orbit closure separation problem* that asks whether the orbit closures of two given m-tuples intersect. For the conjugation action of $\mathrm{GL}_n(\overline{K})$ a polynomial time algorithm for orbit closure separation was given by Forbes and Shpilka in characteristic 0. Here, we give a polynomial time algorithm for the orbit closure separation problem for both the conjugation action of $\mathrm{GL}_n(\overline{K})$ and the left-right action of $\mathrm{SL}_n(\overline{K}) \times \mathrm{SL}_n(\overline{K})$ in arbitrary characteristic. We also improve the known bounds for the degree of separating invariants in these cases.

1. Introduction

The algorithms we present will only use numbers from the field of definition, as opposed to its algebraic closure (see Section 5A). However, it will be convenient to assume that the field of definition is algebraically closed for stating and proving results.

In this paper, let K denote an algebraically closed field. For a vector space V over the field K, let $K[V]$ denote the ring of polynomial functions on V. Suppose that a group G acts on V by linear transformations. A polynomial $f \in K[V]$ is called an *invariant polynomial* if it is constant along orbits, i.e., $f(g \cdot v) = f(v)$ for all $g \in G$ and $v \in V$. The invariant polynomials form a graded subalgebra $K[V]^G = \bigoplus_{d=0}^{\infty} K[V]^G_d$, where $K[V]^G_d$ denotes the degree d homogeneous invariants. We will call $K[V]^G$ the *invariant ring* or the *ring of invariants*.

For a point $v \in V$, its orbit $G \cdot v = \{g \cdot v \mid g \in G\}$ is not necessarily closed with respect to the Zariski topology. We say that an invariant f separates two points $v, w \in V$ if $f(v) \neq f(w)$. It follows from continuity that any invariant polynomial must take the same value on all points of the closure of an orbit. Hence invariant polynomials cannot separate two points whose orbit closures intersect.

Derksen was supported by NSF grant DMS-1601229, DMS-2001460 and IIS-1837985. Makam was supported by the University of Melbourne and the NSF grants DMS-1601229, DMS-1638352, CCF-1412958 and CCF-1900460.

MSC2010: primary 13A50; secondary 14L24, 68W30.

Keywords: orbit closure intersection, null cone, matrix semi-invariants, matrix invariants, separating invariants.

We can ask the converse question: if $v, w \in V$ such that $\overline{G \cdot v} \cap \overline{G \cdot w} = \varnothing$, then is there an invariant polynomial $f \in K[V]^G$ such that $f(v) \neq f(w)$? The answer to this question is in general negative; see [Derksen and Kemper 2002, Example 2.2.8]. However, if we enforce additional hypothesis, we get a positive answer as the theorem below shows; see [Mumford et al. 1994].

Theorem 1.1. *Let V be a rational representation of a reductive group G. Then for $v, w \in V$, there exists $f \in K[V]^G$ such that $f(v) \neq f(w)$ if and only if $\overline{G \cdot v} \cap \overline{G \cdot w} = \varnothing$.*

Henceforth, we shall assume that V is a rational representation of a reductive group G.

Problem 1.2 (orbit closure problem). Decide whether the orbit closures of two given points $v, w \in V$ intersect.

Definition 1.3. Two points $v, w \in V$ are said to be *closure equivalent* if $\overline{G \cdot v} \cap \overline{G \cdot w} \neq \varnothing$. We write $v \sim w$ if v and w are closure equivalent, and we write $v \not\sim w$ if they are not closure equivalent.

By Theorem 1.1, we have $v \sim w$ if and only if $f(v) = f(w)$ for all $f \in K[V]^G$. So $\sim$ is clearly an equivalence relation. Since closure equivalence can be detected by invariant polynomials, the existence of a small generating set of invariants, each of which can be computed efficiently would give an algorithm for the orbit closure problem. Fortunately, the invariant ring $K[V]^G$ is finitely generated; see [Haboush 1975; Hilbert 1890; 1893; Nagata 1963/64].

Definition 1.4. We define $\beta(K[V]^G)$ to be the smallest integer D such that invariants of degree $\leq D$ generate $K[V]^G$, i.e.,

$$\beta(K[V]^G) = \min\bigl\{D \in \mathbb{N} \mid \textstyle\bigcup_{d=1}^{D} K[V]_d^G \text{ generates } K[V]^G\bigr\},$$

where $\mathbb{N} = \{1, 2, \dots\}$.

We are not just interested in deciding whether orbit closures intersect — when they do not, we want to provide an explicit invariant that separates them. To be able to do this efficiently, there must exist an invariant of small enough degree that separates the two given points. A strong upper bound on $\beta(K[V]^G)$ would provide evidence that such invariants exist. Such a bound can be obtained for any rational representation V of a linearly reductive group G (see [Derksen 2001]), but this is often too large. For the cases of interest to us, stronger bounds exist, and we recall them in Section 2. Despite having strong degree bounds, it is a difficult problem to extract a small set of generators. On the other hand, we may only need a subset of the invariants to detect closure equivalence, prompting the definition of a separating set of invariants.

Definition 1.5. A subset of invariants $\mathcal{S} \subset K[V]^G$ is called a *separating set* of invariants if for every pair $v, w \in V$ such that $v \not\sim w$, there exists $f \in \mathcal{S}$ such that $f(v) \neq f(w)$.

We make another definition.

Definition 1.6. We define $\beta_{\mathrm{sep}}(K[V]^G)$ to be the smallest integer D such that the invariants of degree $\leq D$ form a separating set of invariants, i.e.,

$$\beta_{\mathrm{sep}}(K[V]^G) = \min\{D \in \mathbb{N} \mid \textstyle\bigcup_{d=1}^{D} K[V]_d^G \text{ is a separating set of invariants}\}.$$

Extracting a small set of separating invariants is also difficult; see [Kemper 2003] for a general algorithm. We now turn to a closely related problem, and to describe this we need to recall the null cone.

Definition 1.7. The null cone $\mathcal{N}(G, V) = \{v \in V \mid 0 \in \overline{G \cdot v}\}$.

For a set of polynomials $I \subset K[V]$ we define its vanishing set

$$\mathbb{V}(I) = \{v \in V \mid f(v) = 0 \text{ for all } f \in I\}.$$

The null cone can also be defined by $\mathcal{N}(G, V) = \mathbb{V}(K[V]_+^G)$, where $K[V]_+^G = \bigoplus_{d=1}^{\infty} K[V]_d^G$; see [Derksen and Kemper 2002, Definition 2.4.1, Lemma 2.4.2].

Problem 1.8 (null cone membership problem). Decide whether a given point $v \in V$ lies in the null cone $\mathcal{N}(G, V)$.

Since 0 is a closed orbit, a point $v \in V$ is in the null cone if and only if $0 \sim v$, and hence the null cone membership problem can be seen as a subproblem of the orbit closure problem. So, the null cone membership problem could potentially be easier than the orbit closure problem. On the other hand, an algorithm for the null cone membership problem may provide a stepping stone for the orbit closure problem.

In this paper, we are interested in giving efficient algorithms for the orbit closure problem in two specific cases — matrix invariants and matrix semi-invariants. These two cases have generated considerable interest over the past few years due to their connections to computational complexity; see [Derksen and Makam 2017b; Forbes and Shpilka 2013; Garg et al. 2016; Hrubeš and Wigderson 2014; Ivanyos et al. 2017; 2018; Mulmuley 2017].

Remark 1.9. For analyzing the run time of our algorithms, we will use the unit cost arithmetic model. This is also often referred to as algebraic complexity.

1A. ***Matrix invariants.*** Let $\mathrm{Mat}_{p,q}$ be the set of $p \times q$ matrices. The group GL_n acts by simultaneous conjugation on the space $V = \mathrm{Mat}_{n,n}^m$ of m-tuples of $n \times n$ matrices. This action is given by

$$g \cdot (X_1, X_2, \ldots, X_m) = (gX_1g^{-1}, gX_2g^{-1}, \ldots, gX_mg^{-1}).$$

We set $S(n, m) = K[V]^G$. The ring $S(n, m)$ is often referred to as the ring of matrix invariants. We will write $\sim_C$ for the orbit closure equivalence relation $\sim$ with respect to this simultaneous conjugation action.

1A1. *Representation theoretic view point.* Orbit closure intersection for matrix invariants has an interpretation in terms of finite-dimensional representations of the free algebra. Consider the free algebra $F_m = K\langle t_1, \ldots, t_m\rangle$ on m indeterminates. An m-tuple of matrices $X = (X_1, \ldots, X_m)$ gives an n-dimensional representation, i.e., an action of F_m on K^n where t_i acts via X_i. We will denote this representation by V_X. Two m-tuples X and Y are in the same GL_n orbit if and only if V_X and V_Y

are isomorphic representations of F_m. In other words, we have a correspondence between orbits and isomorphism classes of n-dimensional representations of F_m.

Finite-dimensional representations of F_m form an abelian category. A representation is called semisimple if it is a direct sum of simple representations. A composition series of a representation V is a filtration $0 = V_0 \subseteq V_1 \subseteq \cdots \subseteq V_l = V$ whose successive quotients V_i/V_{i-1} are simple. These simple subquotients are called composition factors and are independent of the choice of composition series. For the representation V, the direct sum $\bigoplus_{i=1}^{l}(V_i/V_{i-1})$ is called the associated semisimple representation of V. The following statements follow from [Artin 1969]:

Proposition 1.10 [Artin 1969]. *Consider the simultaneous conjugation action of $G = \mathrm{GL}_n$ on $\mathrm{Mat}_{n,n}^m$, and let $X, Y \in \mathrm{Mat}_{n,n}$.*

(1) *The orbit of X is closed if and only if the representation V_X is semisimple. In other words, we have a correspondence between closed orbits and semisimple representations of dimension n.*

(2) *There is a unique closed orbit in the orbit closure of X, and the representation corresponding to this unique closed orbit is the associated semisimple representation of V_X.*

(3) *The orbit closures of X and Y intersect if and only if the associated semisimple representations of V_X and V_Y are isomorphic.*

For the representation V_X, let a composition series be $0 = V_0 \subseteq V_1 \subseteq \cdots \subseteq V_l = V_X$. Suppose that $\dim V_i/V_{i-1} = n_i$ for all i. Then for an appropriate choice of basis of K^n, all the X_i's are in a block upper triangular form, with the sizes of the diagonal blocks being $n_1, \ldots, n_l$. Call $(n_1, \ldots, n_l)$ the type of the block upper triangularization. The diagonal blocks correspond to the composition factors V_i/V_{i-1} and the upper triangular blocks capture the information of the nontrivial extensions between these composition factors that make up the module V_X. In particular, the associated semisimple representation is then obtained by setting the strictly upper triangular blocks to 0. Hence, we may also rephrase the orbit closure problem for matrix invariants as follows:

Problem 1.11 (orbit closure for matrix invariants rephrased). Given $X, Y \in \mathrm{Mat}_{n,n}^m$, decide if there exist $g, h \in \mathrm{GL}_n$ such that the m-tuples $g \cdot X$ and $h \cdot Y$ are in block upper triangular form of the same type, such that for all $1 \le i \le m$, the diagonal blocks of $(g \cdot X)_i = gX_ig^{-1}$ and $(h \cdot Y)_i = hY_ih^{-1}$ are the same?

Remark 1.12. The more general question of when two representations V and W of a finitely generated algebra $\mathcal{F}$ have isomorphic associated semisimple representations can be reduced to the above problem. Indeed, we have a surjection $F_m \twoheadrightarrow \mathcal{F}$ for some m, and hence V and W can be viewed as representations of F_m. V and W have isomorphic associated semisimple representations as F_m representations if and only if they have isomorphic associated semisimple representations as $\mathcal{F}$ representations.

1A2. *Forbes–Shpilka algorithm.* Given any separating set $\mathcal{S}$, an obvious algorithm for the orbit closure problem would be to evaluate the two given points at every invariant function in the set $\mathcal{S}$. In characteristic 0, Forbes and Shpilka [2013] constructed a quasipolynomial sized set of explicit separating invariants in this case, but this is not sufficient to get a polynomial time algorithm.

Nevertheless, Forbes and Shpilka gave a deterministic parallel polynomial time algorithm for the orbit closure problem in characteristic 0. Given an input $X \in \mathrm{Mat}_{n,n}^m$, one can construct in polynomial time a noncommutative polynomial P_X with the feature that the coefficients of the monomials in P_X are the evaluations of a generating set of invariants on X. Hence, to check if the orbit closures of two points $X, Y \in \mathrm{Mat}_{n,n}^m$ intersect, one needs to determine whether the noncommutative polynomial $P_X - P_Y$ is the zero polynomial. There is an efficient algorithm to test whether $P_X - P_Y$ is the zero polynomial; see [Raz and Shpilka 2005].

1A3. *Our results.* Forbes and Shpilka's algorithm does not work in positive characteristic. In this paper, we provide an algorithm that works in all characteristics.

Theorem 1.13. *The orbit closure problem for the simultaneous conjugation action of* GL_n *on* $\mathrm{Mat}_{n,n}^m$ *can be decided in polynomial time. Further, if* $A, B \in \mathrm{Mat}_{n,n}^m$ *and* $A \not\sim_C B$, *then an explicit invariant* $f \in S(n, m)$ *that separates* A *and* B *can be found in polynomial time.*

Our algorithm has a remarkable and exciting feature — analyzing it allows us to prove a bound on the degree of separating invariants! The bounds we obtain beat the existing ones in literature; see [Mulmuley 2017].

Theorem 1.14. *We have* $\beta_{\mathrm{sep}}(S(n, m)) \le 4n^2 \log_2(n) + 12n^2 - 4n$. *If we assume* $\mathrm{char}(K) = 0$, *then we have* $\beta_{\mathrm{sep}}(S(n, m)) \le 4n \log_2(n) + 12n - 4$.

The bound in characteristic 0 is especially interesting because there are quadratic lower bounds for the degree of generating invariants in this case; see [Domokos 2018; Kuzmin 1975; Formanek 1986]. This also improves the bound in [Derksen and Makam 2017a] for the degree of invariants defining the null cone.

1B. ***Matrix semi-invariants.*** We consider the left-right action of $G = \mathrm{SL}_n \times \mathrm{SL}_n$ on the space $V = \mathrm{Mat}_{n,n}^m$ of m-tuples of $n \times n$ matrices. This action is given by

$$(P, Q) \cdot (X_1, X_2, \ldots, X_m) = (PX_1Q^{-1}, PX_2Q^{-1}, \ldots, PX_mQ^{-1}).$$

We set $R(n, m) = K[V]^G$. The ring $R(n, m)$ is often referred to as the ring of matrix semi-invariants. We will write $\sim_{LR}$ for the equivalence relation $\sim$ with respect to this left-right action.

Remark 1.15. Two m-tuples of $n \times n$ matrices $A = (\mathrm{Id}, A_2, \ldots, A_m)$ and $B = (\mathrm{Id}, B_2, \ldots, B_m)$ are in the same $\mathrm{SL}_n \times \mathrm{SL}_n$ orbit for the left-right action if and only if $\widetilde{A} = (A_2, \ldots, A_m)$ and $\widetilde{B} = (B_2, \ldots, B_m)$ are in the same GL_n orbit for the simultaneous conjugation action. This is compatible with orbit closure in the sense that the orbit closures of A and B intersect for the left-right action if and only if the orbit closures for $\widetilde{A}$ and $\widetilde{B}$ intersect for the simultaneous conjugation action; see Corollary 3.3 for the precise statement.

For $A, B \in \mathrm{Mat}_{n,n}^m$ with $A_1 = \mathrm{Id}$ it is easy to detect if $A \sim_{LR} B$. If $\det(B_1) \neq 1$, then $A \not\sim_{LR} B$. Otherwise, we have $\det(B_1) = 1$, i.e., $B_1 \in \mathrm{SL}_n$ and hence $\widetilde{B} = (B_1^{-1}, \mathrm{Id}) \cdot B$ is in the same orbit as B. Thus, it suffices to detect whether the orbit closures of A and $\widetilde{B}$ intersect. By design, we have $\widetilde{B}_1 = \mathrm{Id}$.

By the above remark, it suffices to detect whether the orbit closures for $(A_2, \dots, A_m)$ and $(\widetilde{B}_2, \dots, \widetilde{B}_m)$ intersect for the conjugation action.

In fact, if we can find a nonsingular matrix in the span of $(A_1, \dots, A_m)$, then a similar strategy can be used to detect orbit closure intersection; see Proposition 3.5. We can now highlight two important issues that need to be addressed.

(1) It is not known how to decide if the span of $A_1, \dots, A_m$ contains a nonsingular matrix in polynomial time. In [Valiant 1979], it was shown that this problem captures the problem of polynomial identity testing (PIT) (see also [Garg et al. 2016]). A polynomial time algorithm for PIT is a major open problem in computational complexity.

(2) There may not be a nonsingular matrix in the span of the matrices $A_1, \dots, A_m$. One might be tempted to hope that this condition would be equivalent to membership in the null cone, but this turns out to be erroneous. The simplest example is the 3-tuple of 3×3 matrices

$$S = \left(\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \right) \in \mathrm{Mat}_{3,3}^3 .$$

It is well known that S is not in the null cone (see [Domokos 2000]), but every matrix in the span of S_1, S_2, S_3 is singular. Similar examples can be found in [Derksen and Makam 2017b; 2018; Draisma 2006; Eisenbud and Harris 1988]. There are several equivalent characterizations of the null cone, and we refer to [Garg et al. 2016; Ivanyos et al. 2017] for details.

1B1. *Null cone membership problem.* The null cone membership problem for matrix semi-invariants has attracted a lot of attention due to its connections to noncommutative circuits and identity testing; see [Derksen and Makam 2017b; Garg et al. 2016; Hrubeš and Wigderson 2014; Ivanyos et al. 2017; 2018]. In characteristic 0, Gurvits' algorithm gives a deterministic polynomial time algorithm; see [Derksen and Makam 2017b; Garg et al. 2016]. There is a different algorithm which works for any sufficiently large field in [Ivanyos et al. 2018].

Theorem 1.16 [Derksen and Makam 2017b; Garg et al. 2016; Ivanyos et al. 2018]. *The null cone membership problem for the left-right action of* $\mathrm{SL}_n \times \mathrm{SL}_n$ *on* $\mathrm{Mat}_{n,n}^m$ *can be decided in polynomial time.*

1B2. *Our results.* The above theorem allows us to bypass the two issues mentioned above, and we are able to show a polynomial time reduction from the orbit closure problem for matrix semi-invariants to the orbit closure problem for matrix invariants. In fact, the converse also holds, i.e., there is a polynomial time reduction from the orbit closure problem for matrix invariants to the orbit closure problem for matrix semi-invariants. As a consequence, we have a polynomial time algorithm for the orbit closure problem for matrix semi-invariants as well. Moreover, due to the nature of the reduction, we will be able to find a separating invariant when the orbit closures of two points do not intersect.

Theorem 1.17. *The orbit closure problem for the left-right action of* $\mathrm{SL}_n \times \mathrm{SL}_n$ *on* $\mathrm{Mat}_{n,n}^m$ *can be decided in polynomial time. Further for* $A, B \in \mathrm{Mat}_{n,n}^m$, *if* $A \not\sim_{LR} B$, *an explicit invariant* $f \in R(n, m)$ *that separates* A *and* B *can be found in polynomial time.*

In characteristic 0, an analytic algorithm for the orbit closure problem for matrix semi-invariants has also been obtained by Allen-Zhu, Garg, Li, Oliveira and Wigderson [Allen-Zhu et al. 2018]. Our algorithm is algebraic, independent of characteristic, and provides a separating invariant when the orbit closures do not intersect.

In [Derksen and Makam 2017a], bounds on $\beta_{\mathrm{sep}}(R(n, m))$ were given. In this paper, we give better bounds using a reduction to matrix invariants.

Theorem 1.18. *We have* $\beta_{\mathrm{sep}}(R(n, m)) \leq n^2 \beta_{\mathrm{sep}}(S(n, mn^2))$.

Using the bounds on matrix invariants in Theorem 1.14, we get bounds for matrix semi-invariants.

Corollary 1.19. *We have* $\beta_{\mathrm{sep}}(R(n, m)) \leq 4n^4 \log_2(n) + 12n^4 - 4n^3$. *If we assume* $\mathrm{char}(K) = 0$, *then we have* $\beta_{\mathrm{sep}}(R(n, m)) \leq 4n^3 \log_2(n) + 12n^3 - 4n^2$.

Remark 1.20. There is a representation theoretic viewpoint for orbit closure intersection for matrix semi-invariants in terms of semistable representations of the m-Kronecker quiver. We will not recall it as it is not useful for our purposes and refer the interested reader to [King 1994].

Remark 1.21. We will say the null cone membership problem and orbit closure problem for matrix invariants (resp. matrix semi-invariants) to refer to the corresponding problem for the simultaneous conjugation action of GL_n (resp. left-right action of $\mathrm{SL}_n \times \mathrm{SL}_n$) on $\mathrm{Mat}_{n,n}^m$.

Remark 1.22. Another interesting problem is to determine if two tuples $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_m)$ are in the same orbit for the simultaneous conjugation action of GL_n (also for left-right action). An obvious algorithm to do this would be to solve the equations $X_i Z = Z Y_i$ for all i. This is a linear system of equations that can be solved efficiently. However, we need such a Z to be invertible, so we would need to be able to verify whether the space of solutions to the equations $X_i Z = Z Y_i$ has an invertible matrix in it. As pointed out in the discussion after Remark 1.15, it is not known how to do this in polynomial time. Nevertheless, there is a polynomial time algorithm to test if the two tuples X and Y are in the same orbit! We refer the interested reader to [Brooksbank and Luks 2008; Chistov et al. 1997].

1C. ***Organization.*** In Section 2, we collect a number of preliminary results on matrix invariants and matrix semi-invariants. In Section 3, we show polynomial time reductions in both directions between the orbit closure problems for matrix invariants and matrix semi-invariants. We give a polynomial time algorithm for finding a basis of a subalgebra of matrices in Section 4. In Section 5, we give the algorithm for the orbit closure problem for matrix invariants, and prove bounds on separating invariants. Finally in Section 6, we prove Theorem 1.18.

2. Preliminaries on matrix invariants and matrix semi-invariants

2A. ***Matrix invariants.*** Let us recall that the ring of matrix invariants $S(n, m)$ is the invariant ring for the simultaneous conjugation action of GL_n on $\mathrm{Mat}_{n,n}^m$, the space of m-tuples of $n \times n$ matrices. Sibirskiĭ [1968] showed that in characteristic 0, the ring $S(n, m)$ is generated by traces of words in the matrices; see also [Procesi 1976].

A word in an alphabet set Σ is an expression of the form $i_1 i_2 \dots i_k$ with $i_j \in \Sigma$. We denote the set of all words in an alphabet Σ by $\Sigma^\star$ (the Kleene closure of Σ). The set $\Sigma^\star$ includes the empty word ϵ. For a word $w = i_1 i_2 \dots i_k$, we define its length $l(w) = k$. For a positive integer m, we write $[m] := \{1, 2, \dots, m\}$, the set of all positive integers less equal m. For a word $w = i_1 i_2 \dots i_k \in [m]^\star$, and for $X = (X_1, \dots, X_m) \in \mathrm{Mat}_{n,n}^m$, we define $X_w = X_{i_1} X_{i_2} \dots X_{i_k}$. The function $T_w : \mathrm{Mat}_{n,n}^m \to K$ given by $T_w(X) := \mathrm{Tr}(X_w)$ is an invariant polynomial.

Theorem 2.1 [Sibirskiĭ 1968; Procesi 1976]. *Assume* $\mathrm{char}(K) = 0$. *The invariant functions of the form* T_w, $w \in [m]^\star$ *generate* $S(n, m)$.

Razmyslov studied trace identities, and as a consequence of his work, we have:

Theorem 2.2 [Razmyslov 1974]. *Assume* $\mathrm{char}(K) = 0$. *Then* $\beta(S(n, m)) \le n^2$.

In positive characteristic, generators of the invariant ring were given by Donkin [1992; 1993]. In simple terms, we have to replace traces with coefficients of characteristic polynomial. For an $n \times n$ matrix X, let $\mathrm{c}(X) = \det(\mathrm{Id} + tX) = \sum_{i=0}^n \sigma_j(X) t^j$ denote its characteristic polynomial. The function $X \mapsto \sigma_j(X)$ is a polynomial in the entries of X, and is called the j-th characteristic coefficient of X. Note that $\sigma_0 = 1$, $\sigma_1(X) = \mathrm{Tr}(X)$ and $\sigma_n(X) = \det(X)$. For any word w, we define the invariant polynomial $\sigma_{j,w} \in S(n, m)$ by $\sigma_{j,w}(X) := \sigma_j(X_w)$ for $X = (X_1, X_2, \dots, X_m) \in \mathrm{Mat}_{n,n}^m$.

Theorem 2.3 [Donkin 1992; 1993]. *The set of invariant functions* $\{\sigma_{j,w} \mid w \in [m]^\star,\ 1 \le j \le n\}$ *is a generating set for the invariant ring* $S(n, m)$.

In a radically different approach from the case of characteristic 0, we recently proved a polynomial bound on the degree of generators.

Theorem 2.4 [Derksen and Makam 2017a]. *We have* $\beta(S(n, m)) \le (m + 1)n^4$.

2B. ***Matrix semi-invariants.*** The ring of matrix semi-invariants $R(n, m)$ is the ring of invariants for the left-right action of $\mathrm{SL}_n \times \mathrm{SL}_n$ on $\mathrm{Mat}_{n,n}^m$. There is a determinantal description for semi-invariants of quivers; see [Derksen and Weyman 2000; Domokos and Zubkov 2001; Schofield and van den Bergh 2001]. Matrix semi-invariants is a special case — it is the ring of semi-invariants for the generalized Kronecker quiver, for a particular choice of a dimension vector; see for example [Derksen and Makam 2017b].

Given two matrices $A = (a_{ij})$ of size $p \times q$, and $B = (b_{ij})$ of size $r \times s$, we define their tensor (or Kronecker) product to be

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{m1}B & \cdots & \cdots & a_{mn}B \end{bmatrix} \in \mathrm{Mat}_{pr,qs}\,.$$

Associated to each $T = (T_1, T_2, \ldots, T_m) \in \mathrm{Mat}_{d,d}^m$, we define a homogeneous invariant $f_T \in R(n, m)$ of degree dn by

$$f_T(X_1, X_2, \ldots, X_m) = \det(T_1 \otimes X_1 + T_2 \otimes X_2 + \cdots + T_m \otimes X_m).$$

Theorem 2.5 [Derksen and Weyman 2000; Domokos and Zubkov 2001; Schofield and van den Bergh 2001]. *The invariant ring $R(n, m)$ is spanned by all f_T with $T \in \mathrm{Mat}_{d,d}^m$ and $d \geq 1$.*

In particular, notice that if d is not a multiple of n, then there are no degree d invariants. In other words, we have $R(n, m) = \bigoplus_{d=0}^{\infty} R(n, m)_{dn}$. A polynomial bound on the degree of generators in characteristic 0 was shown in [Derksen and Makam 2017b], and the restriction on characteristic was removed in [Derksen and Makam 2017a].

Theorem 2.6 [Derksen and Makam 2017a; 2017b]. *We have $\beta(R(n, m)) \leq mn^4$. If $\mathrm{char}(K) = 0$, then $\beta(R(n, m)) \leq n^6$.*

Let $\mathcal{N}(n, m)$ denote the null cone for the left-right action of $\mathrm{SL}_n \times \mathrm{SL}_n$ on $\mathrm{Mat}_{n,n}^m$. The following is proved in [Derksen and Makam 2017b].

Theorem 2.7 [Derksen and Makam 2017b]. *For $X \in \mathrm{Mat}_{n,n}^m$, the following are equivalent*:

(1) $X \notin \mathcal{N}(n, m)$.

(2) *For some $d \in \mathbb{N}$, there exists $T \in \mathrm{Mat}_{d,d}^m$ such that $f_T(X) \neq 0$.*

(3) *For any $d \geq n - 1$, there exists $T \in \mathrm{Mat}_{d,d}^m$ such that $f_T(X) \neq 0$.*

The above theorem relies crucially on the regularity lemma proved in [Ivanyos et al. 2017]. A more conceptual proof of the regularity lemma is given in [Derksen and Makam 2018] using universal division algebras, although it lacks the constructiveness of the original proof.

An algorithmic version of the above theorem appears in [Ivanyos et al. 2018].

Theorem 2.8 [Ivanyos et al. 2018]. *For $X \in \mathrm{Mat}_{n,n}^m$, there is a deterministic polynomial time (in n and m) algorithm which determines if $X \notin \mathcal{N}(n, m)$. Further, for $X \notin \mathcal{N}(n, m)$ and any $n - 1 \leq d \leq \mathrm{poly}(n)$, the algorithm provides in polynomial time, an explicit $T \in \mathrm{Mat}_{d,d}^m$ such that $f_T(X) \neq 0$.*

Remark 2.9. We will henceforth refer to the algorithm in Theorem 2.8 above as the IQS algorithm.

For $1 \leq j, k \leq d$, we define $E_{j,k} \in \mathrm{Mat}_{d,d}$ to be the $d \times d$ matrix which has a 1 in the (j, k)-th entry, and 0 everywhere else.

Definition 2.10. If $X = (X_1, \dots, X_m) \in \mathrm{Mat}_{n,n}^m$, we define $X^{[d]} = (X_i \otimes E_{j,k})_{i,j,k} \in \mathrm{Mat}_{nd,nd}^{md^2}$, where the tuples $(i, j, k) \in [m] \times [d] \times [d]$ are ordered lexicographically.

Proposition 2.11. *The following are equivalent*:

(1) *There exists* $f \in R(n, m)$ *such that* $f(A) \neq f(B)$.

(2) *There exists* $g \in R(nd, md^2)$ *such that* $g(A^{[d]}) \neq g(B^{[d]})$ *for either* $d = n - 1$ *or* $d = n$.

Proof. We first show (1) $\Longrightarrow$ (2). We can assume $f = f_T$ for some $T \in \mathrm{Mat}_{e,e}^m$ for some $e \geq 1$. Without loss of generality, assume $f(A) \neq 0$. Then we have $\mu = f(B)/f(A) \neq 1$. For any $\mu \neq 1$, both μ^{n-1} and μ^n cannot be 1. Hence for at least one of $d \in \{n-1, n\}$, we have $\mu^d = f(B)^d/f(A)^d \neq 1$, and hence $f(A)^d \neq f(B)^d$. Now, it suffices to show the existence of $g \in R(nd, md^2)$ such that $g(A^{[d]}) = f(A)^d$ for all $A \in \mathrm{Mat}_{n,n}^m$.

But now, consider

$$\begin{aligned} f_T(A)^d &= \det\big(\textstyle\sum_{i=1}^m T_i \otimes A_i\big)^d \\ &= \det\big(\textstyle\sum_{i=1}^m T_i^{\oplus d} \otimes A_i\big) \\ &= \det\big(\textstyle\sum_{i=1}^m \big(\sum_{k=1}^d T_i \otimes E_{k,k} \otimes A_i\big)\big) \\ &= \det\big(\textstyle\sum_{i,k} T_i \otimes (A_i \otimes E_{k,k})\big). \end{aligned}$$

Let $S \in \mathrm{Mat}_{e,e}^{md^2}$ given by $S_{i,j,k} = \delta_{j,k} T_i$. We can take $g = f_S$.

We now show (2) $\Longrightarrow$ (1). Indeed, we can choose $g = f_S$ for some $S \in \mathrm{Mat}_{e,e}^{md^2}$, $e \geq 1$. We have

$$\begin{aligned} f_S(A^{[d]}) &= \det\big(\textstyle\sum_{i,j,k} S_{i,j,k} \otimes (A^{[d]})_{i,j,k}\big) \\ &= \det\big(\textstyle\sum_{i,j,k} S_{i,j,k} \otimes A_i \otimes E_{j,k}\big) \\ &= \det\big(\textstyle\sum_i \big(\sum_{j,k} S_{i,j,k} \otimes E_{j,k}\big) \otimes A_i\big) \\ &= \det\big(\textstyle\sum_i \widetilde{S}_i \otimes A_i\big), \end{aligned}$$

where $\widetilde{S}_i = \sum_{j,k} S_{i,j,k} \otimes E_{j,k}$. Let $\widetilde{S} = (\widetilde{S}_1, \dots, \widetilde{S}_m) \in \mathrm{Mat}_{de,de}^m$. Then the above calculation tells us that $f_{\widetilde{S}}(A) = f_S(A^{[d]}) = g(A^{[d]})$. Hence we have

$$f_{\widetilde{S}}(A) = g(A^{[d]}) \neq g(B^{[d]}) = f_{\widetilde{S}}(B).$$

We can take $f = f_{\widetilde{S}}$. □

Corollary 2.12. *The orbit closures of* A *and* B *do not intersect if and only if the orbit closures of* $A^{[d]}$ *and* $B^{[d]}$ *do not intersect for at least one choice of* $d \in \{n-1, n\}$.

2C. ***Commuting action of another group.*** Let G be a group acting on V. Suppose we have another group H acting on V, and the actions of G and H commute. To distinguish the actions, we will denote the action of H by $\star$. The orbit closure problem for the action of G on V also commutes with the action of H. More precisely, we have the following:

Lemma 2.13. *Let $v, w \in V$ and $h \in H$. Then $v \sim w$ if and only if $h \star v \sim h \star w$.*

We have a natural identification of $V = \mathrm{Mat}_{n,n}^m$ with $\mathrm{Mat}_{n,n} \otimes K^m$. The latter viewpoint illuminates an action of GL_m on V that commutes with the left-right action of $\mathrm{SL}_n \times \mathrm{SL}_n$, as well as the simultaneous conjugation action of GL_n. In explicit terms, for $P = (p_{i,j}) \in \mathrm{GL}_m$ and $X = (X_1, \ldots, X_m) \in \mathrm{Mat}_{n,n}^m$, we have

$$P \star (X_1, \ldots, X_m) = \big(\textstyle\sum_j p_{1,j} X_j, \sum_j p_{2,j} X_j, \ldots, \sum_j p_{m,j} X_j\big).$$

Corollary 2.14. *The orbit closure problem for both the left-right action of $\mathrm{SL}_n \times \mathrm{SL}_n$ and the simultaneous conjugation action of GL_n on $\mathrm{Mat}_{n,n}^m$ commutes with the action of GL_m.*

2D. ***A useful surjection.*** We consider the map

$$\phi : \mathrm{Mat}_{n,n}^m \to \mathrm{Mat}_{n,n}^{m+1}, \quad (X_1, \ldots, X_m) \mapsto (\mathrm{Id}, X_1, \ldots, X_m).$$

This gives a surjection on the coordinate rings $\phi^* : K[\mathrm{Mat}_{n,n}^{m+1}] \to K[\mathrm{Mat}_{n,n}^m]$, which descends to a surjective map on invariant rings as below; see [Domokos 2000; Derksen and Makam 2017a].

Proposition 2.15 [Domokos 2000]. *The map $\phi^* : R(n, m+1) \twoheadrightarrow S(n, m)$ is surjective.*

We recall the proof of this proposition because the construction in the proof plays a significant role in some of the algorithms below. Before proving the proposition, let us recall some basic linear algebra. For a matrix $X \in \mathrm{Mat}_{n,n}$, let us denote the adjoint (or adjugate) matrix by $\mathrm{Adj}(X)$.

Lemma 2.16. *Let $X, Y \in \mathrm{Mat}_{n,n}$. Then we have*:

(1) $\mathrm{Adj}(XY) = \mathrm{Adj}(Y)\,\mathrm{Adj}(X)$.

(2) $X\,\mathrm{Adj}(X) = \det(X)\,\mathrm{Id}$. *In particular, if* $\det(X) = 1$, *then* $\mathrm{Adj}(X) = X^{-1}$.

(3) *For* $(P, Q) \in \mathrm{SL}_n \times \mathrm{SL}_n$, *we have* $\mathrm{Adj}(PXQ^{-1})(PYQ^{-1}) = Q(\mathrm{Adj}(X)Y)Q^{-1}$.

Proof. The first two are well known. The last one follows from the first two. □

Proof of Proposition 2.15. We want to first show that we have an inclusion $\phi^*(R(n, m+1)) \subseteq S(n, m)$.

Indeed for $f \in R(n, m+1)$ and $g \in \mathrm{GL}_n$, we have

$$\begin{aligned}
\phi^*(f)(gX_1g^{-1}, \ldots, gX_mg^{-1}) &= f(\mathrm{Id}, gX_1g^{-1}, \ldots, gX_mg^{-1}) \\
&= f(g\mathrm{Id}g^{-1}, gX_1g^{-1}, \ldots, gX_mg^{-1}) \\
&= f(\mathrm{Id}, X_1, \ldots, X_m) \\
&= \phi^*(f)(X_1, \ldots, X_m).
\end{aligned}$$

The third equality is the only nontrivial one. Even though g may not be in SL_n, we can replace g by $g' = \lambda g \in \mathrm{SL}_n$ for a suitable $\lambda \in K^*$. Then, one has to observe that conjugation by g and conjugation by g' are the same.

Now, we show that the image of ϕ^* surjects onto $S(n, m)$. For $f \in S(n, m)$, define $\tilde{f}$ by

$$\tilde{f}(X_1, \ldots, X_{m+1}) = f(\mathrm{Adj}(X_1)X_2, \mathrm{Adj}(X_1)X_3, \ldots, \mathrm{Adj}(X_1)X_{m+1}).$$

We claim that $\tilde{f}$ is invariant with respect to the left-right action of $\mathrm{SL}_n \times \mathrm{SL}_n$. Indeed for $(P, Q) \in \mathrm{SL}_n \times \mathrm{SL}_n$, we have

$$\begin{aligned}\tilde{f}(PX_1Q^{-1}, \ldots, PX_{m+1}Q^{-1}) &= f(\mathrm{Adj}(PX_1Q^{-1})PX_2Q^{-1}, \ldots, \mathrm{Adj}(PX_1Q^{-1})PX_{m+1}Q^{-1}) \\ &= f(Q(\mathrm{Adj}(X_1)X_2)Q^{-1}, \ldots, Q(\mathrm{Adj}(X_1)X_{m+1})Q^{-1}) \\ &= f(\mathrm{Adj}(X_1)X_2, \ldots, \mathrm{Adj}(X_1)X_{m+1}) \\ &= \tilde{f}(X_1, \ldots, X_{m+1}).\end{aligned}$$

The second equality follows from the above lemma, and the third follows because f is invariant under simultaneous conjugation.

Further, we have

$$\begin{aligned}(\phi^*(\tilde{f}))(X_1, \ldots, X_m) &= \tilde{f}(\mathrm{Id}, X_1, \ldots, X_m) \\ &= f(\mathrm{Adj}(\mathrm{Id})X_1, \ldots, \mathrm{Adj}(\mathrm{Id})X_m) \\ &= f(X_1, \ldots, X_m)\end{aligned}$$

Hence for each $f \in S(n, m)$, we have constructed a preimage $\tilde{f} \in R(n, m+1)$. Thus ϕ^* is a surjection from $R(n, m+1)$ onto $S(n, m)$. □

In fact, from the above proof, we can see that for $f \in S(n, m)$, we can construct a preimage easily. We record this as a corollary.

Corollary 2.17 [Domokos 2000]. *For $f \in S(n, m)$, the invariant polynomial $\tilde{f} \in R(n, m+1)$ defined by*

$$\tilde{f}(X_1, \ldots, X_{m+1}) = f(\mathrm{Adj}(X_1)X_2, \mathrm{Adj}(X_1)X_3, \ldots, \mathrm{Adj}(X_1)X_{m+1})$$

is a preimage of f under ϕ^*, i.e., $\phi^*(\tilde{f}) = f$.

3. Time complexity equivalence of orbit closure problems

In this section, we will show polynomial reductions between the orbit closure problem for matrix invariants and the orbit closure problem for matrix semi-invariants. We will in fact show a more robust reduction.

Let G be a group acting on V.

Definition 3.1. An algorithm for the *orbit closure problem with witness* is an algorithm that decides if $v \sim w$ for any two points $v, w \in V$, and if $v \not\sim w$, provides a witness $f \in K[V]^G$ such that $f(v) \neq f(w)$.

3A. ***Reduction from matrix invariants to matrix semi-invariants.*** Let $A, B \in \mathrm{Mat}_{n,n}^m$. We can consider $\phi(A), \phi(B) \in \mathrm{Mat}_{n,n}^{m+1}$, where $\phi : \mathrm{Mat}_{n,n}^m \to \mathrm{Mat}_{n,n}^{m+1}$ is the map described in Section 2D.

Proposition 3.2. *The following are equivalent*:

(1) *There exists $f \in S(n, m)$ such that $f(A) \neq f(B)$.*

(2) *There exists $g \in R(n, m+1)$ such that $g(\phi(A)) \neq g(\phi(B))$.*

Proof. Recall the surjection $\phi^* : R(n, m+1) \twoheadrightarrow S(n, m)$ from Proposition 2.15. Let's first prove (1) $\Longrightarrow$ (2). Given $f \in S(n, m)$ such that $f(A) \neq f(B)$, take g to be a preimage of f, i.e., $\phi^*(g) = f$. Now,

$$g(\phi(A)) = \phi^*(g)(A) = f(A) \neq f(B) = \phi^*(g)(B) = g(\phi(B)).$$

To prove (2) $\Longrightarrow$ (1), simply take $f = \phi^*(g)$. □

Corollary 3.3. *Let $A, B \in \mathrm{Mat}_{n,n}^m$. Then we have*

$$A \sim_C B \text{ if and only if } \phi(A) \sim_{LR} \phi(B).$$

Corollary 3.4. *There is a polynomial reduction that reduces the orbit closure problem with witness for matrix invariants to the orbit closure problem with witness for matrix semi-invariants.*

Proof. Given $A, B \in \mathrm{Mat}_{n,n}^m$, we construct $\phi(A)$ and $\phi(B)$. Appeal to the orbit closure problem with witness for matrix semi-invariants with input $\phi(A)$ and $\phi(B)$. There are two possible outcomes. If $\phi(A) \sim_{LR} \phi(B)$, then we conclude that $A \sim_C B$. If $\phi(A) \not\sim_{LR} \phi(B)$ and $f \in R(n, m+1)$ separates $\phi(A)$ and $\phi(B)$, then $\phi^*(f)$ is an invariant that separates A and B. The reduction is clearly polynomial time. □

3B. ***Reduction from matrix semi-invariants to matrix invariants.*** We will show that the orbit closure problem for matrix semi-invariants can be reduced to the orbit closure problem for matrix invariants. Let $A, B \in \mathrm{Mat}_{n,n}^m$. Recall the discussion in Section 1B, in particular, that if we can find efficiently a nonsingular matrix in the span of $A_1, \ldots, A_m$, we would be done. We must address the two issues indicated in Section 1B. The IQS algorithm (Theorem 2.8) can determine whether A is in the null cone for the left-right action. Further, when A is not in the null cone, it constructs efficiently a nonsingular matrix of the form $\sum_{i=1}^m T_i \otimes A_i$, with $T_i \in \mathrm{Mat}_{d,d}$ for any $n - 1 \leq d < \mathrm{poly}(n)$. Roughly speaking, these nonsingular matrices will address both issues. We will now make precise statements.

Proposition 3.5. *Assume $A, B \in \mathrm{Mat}_{n,n}^m$ such that $\det(A_1) = \det(B_1) \neq 0$. If we denote*

$$\widetilde{A} = (A_1^{-1}A_2, \ldots, A_1^{-1}A_m) \quad \text{and} \quad \widetilde{B} = (B_1^{-1}B_2, \ldots, B_1^{-1}B_m),$$

then we have

$$A \sim_{LR} B \Longleftrightarrow \widetilde{A} \sim_C \widetilde{B}.$$

Proof. Let us first suppose that $\det(A_1) = \det(B_1) = 1$. Then for $g = (A_1^{-1}, \mathrm{Id}) \in \mathrm{SL}_n \times \mathrm{SL}_n$, we have $g \cdot A = (\mathrm{Id}, A_1^{-1}A_2, \ldots, A_1^{-1}A_m) = \phi(\widetilde{A})$. Similarly for $h = (B_1^{-1}, \mathrm{Id}) \in \mathrm{SL}_n \times \mathrm{SL}_n$, we have $h \cdot B = \phi(\widetilde{B})$. Now, we have

$$A \sim_{LR} B \Longleftrightarrow g \cdot A \sim_{LR} h \cdot B \Longleftrightarrow \phi(\widetilde{A}) \sim_{LR} \phi(\widetilde{B}) \Longleftrightarrow \widetilde{A} \sim_C \widetilde{B}.$$

The last statement follows from Corollary 3.3. The general case for $\det(A_1) \neq 0$ follows because the orbit closures of A and B intersect if and only if the orbit closures of $\lambda \cdot A = (\lambda A_1, \ldots, \lambda A_m)$ and $\lambda \cdot B = (\lambda B_1, \ldots, \lambda B_m)$ intersect for any $\lambda \in K^*$; see Lemma 2.13. □

Lemma 3.6. *For any nonzero row vector $\boldsymbol{v} = (v_1, \ldots, v_m)$, we can construct efficiently a matrix $P \in \mathrm{GL}_m$ such that the top row of the matrix P is $\boldsymbol{v}$.*

Proof. This is straightforward and left to the reader. □

Algorithm 3.7. Now we give an algorithm to reduce the orbit closure problem with witness for matrix semi-invariants to the orbit closure problem with witness for matrix invariants.

Input: $A, B \in \mathrm{Mat}_{n,n}^m$.

Step 1: Check if A or B are in the null cone by the IQS algorithm. If both of them are in the null cone, then $A \sim_{LR} B$. If precisely one of them is in the null cone, then $A \not\sim_{LR} B$ and the IQS algorithm gives an invariant that separates A and B. If neither are in the null cone, then we proceed to Step 2.

Step 2: Neither A nor B in the null cone. Now, for $d \in \{n-1, n\}$, the IQS algorithm constructs $T(d) \in \mathrm{Mat}_{d,d}^m$ such that $f_{T(d)}(A) \neq 0$ in polynomial time. We denote $f_d := f_{T(d)}$. If $f_d(A) \neq f_d(B)$, then $A \not\sim_{LR} B$ and f_d is the separating invariant. Else $f_d(A) = f_d(B)$ for both choices of $d \in \{n-1, n\}$, and we proceed to Step 3.

Step 3: For $d \in \{n-1, n\}$, we have

$$\begin{aligned} f_d(A) &= \det\big(\textstyle\sum_i T(d)_i \otimes A_i\big) \\ &= \det\big(\textstyle\sum_i \big(\sum_{j,k} (T(d)_i)_{j,k} E_{j,k}\big) \otimes A_i\big) \\ &= \det\big(\textstyle\sum_{i,j,k} (T(d)_i)_{j,k} (E_{j,k} \otimes A_i)\big) \\ &= \det\big(\textstyle\sum_{i,j,k} (T(d)_i)_{j,k} (A_i \otimes E_{j,k})\big). \end{aligned}$$

We can construct efficiently a matrix $P \in \mathrm{Mat}_{md^2, md^2}$ such that the first row is $(T(d)_i)_{j,k})_{i,j,k}$ by Lemma 3.6. Consider $U = P \star A^{[d]}$, $V = P \star B^{[d]} \in \mathrm{Mat}_{nd,nd}^{md^2}$. By construction, this has the property that $\det(U_1) = f_d(A) \neq 0$, and $\det(V_1) = f_d(B)$. Since we did not terminate in Step 2, we know that $\det(U_1) = \det(V_1)$. Let us recall that by Corollary 2.12, $A \sim_{LR} B$ if and only $A^{[d]} \sim_{LR} B^{[d]}$ for both $d = n-1$ and $d = n$. By Lemma 2.13, $A^{[d]} \sim_{LR} B^{[d]}$ if and only if $U \sim_{LR} V$.

To decide whether $U \sim_{LR} V$, we do the following. Let $\widetilde{U} = (U_1^{-1}U_2, \ldots, U_1^{-1}U_{md^2})$ and $\widetilde{V} = (V_1^{-1}V_2, \ldots, V_1^{-1}V_{md^2})$. By Proposition 3.5, we have $U \sim_{LR} V$ if and only if $\widetilde{U} \sim_C \widetilde{V}$. But this can be seen as an instance of an orbit closure problem with witness for matrix invariants. Also note the fact if we get an invariant separating $\widetilde{U}$ and $\widetilde{V}$, the steps can be traced back to get an invariant separating A and B.

Corollary 3.8. *There is a polynomial time reduction from the orbit closure problem with witness for matrix semi-invariants to the orbit closure problem with witness for matrix invariants.*

4. A polynomial time algorithm for finding a subalgebra basis

Let $\{C_1, \ldots, C_m\} \subseteq \mathrm{Mat}_{n,n}$ be a finite subset of $\mathrm{Mat}_{n,n}$. Consider the (unital) subalgebra $\mathcal{C} \subseteq \mathrm{Mat}_{n,n}$ generated by $C_1, \ldots, C_m$. In other words, $\mathcal{C}$ is the smallest subspace of $\mathrm{Mat}_{n,n}$ containing the identity matrix Id and the matrices $C_1, \ldots, C_m$ that is closed under multiplication. For a word $i_1 i_2 \ldots i_b$ we define

$C_w = C_{i_1} C_{i_2} \cdot C_{i_b}$. We also define $C_\epsilon = \mathrm{Id}$ for the empty word ϵ. We will describe a polynomial time algorithm for finding a basis for $\mathcal{C}$. First observe that $\mathcal{C}$ is spanned by $\{C_w \mid w \in [m]^\star\}$. While this is an infinite spanning set, we will extract a basis from this, in polynomial time. We define a total order on $[m]^\star$.

Definition 4.1. For words $w_1 = i_1 i_2 \dots i_b$ and $w_2 = j_1 j_2 \dots j_c$, we write $w_1 \prec w_2$ if either

(1) $l(w_1) < l(w_2)$ or

(2) $l(w_1) = l(w_2)$ and for the smallest integer m for which $i_m \neq j_m$, we have $i_m < j_m$.

Remark 4.2. If $w \prec w'$, we will say w is smaller than w'.

We call a word w a pivot if C_w does not lie in the span of all C_u, $u \prec w$. Otherwise, we call w a nonpivot.

Lemma 4.3. *Let $P = \{w \mid w \text{ is pivot}\}$. Then $\{C_w \mid w \in P\}$ is a basis for $\mathcal{C}$. We call this the pivot basis.*

Definition 4.4. For words $w = i_1 i_2 \dots i_b$ and $w' = j_1 j_2 \dots j_c$, we define the concatenation

$$w w' = i_1 i_2 \dots i_b j_1 j_2 \dots j_c.$$

Lemma 4.5. *If w is a nonpivot, then xwy is a nonpivot for all words $x, y \in [m]^\star$.*

Proof. If w is nonpivot, then $C_w = \sum_k a_k C_{w_k}$ for $w_k \prec w$ and $a_k \in K$. Then we have $C_{xwy} = \sum_k a_k C_{x w_k y}$. Hence, xwy is nonpivot as well. □

Corollary 4.6. *Every subword of a pivot word is a pivot.*

Lemma 4.7. *The length of the longest pivot is at most $2n \log_2(n) + 4n - 4$.*

Proof. This follows from the main result of [Shitov 2019]. For a collection $S \subseteq \mathrm{Mat}_{n,n}$, we define $l(S)$ as the smallest integer k such that all the words of length $\leq k$ in S span the subalgebra of $\mathrm{Mat}_{n,n}$ generated by S. In particular, if we take $S = \{C_1, \dots, C_m\}$, this means that any pivot word has length at most $l(S)$. Moreover, $l(S) \leq 2n \log_2(n) + 4n - 4$ is the statement of [Shitov 2019, Theorem 3] (a strong improvement over the previous known bound from [Pappacena 1997]). Thus every pivot word has length at most $2n \log_2(n) + 4n - 4$ as required. □

Now, we describe an efficient algorithm to construct the set of pivots.

Algorithm 4.8 (finding a basis for a subalgebra of $\mathrm{Mat}_{n,n}$).

Input: $n \times n$ matrices $C_1, C_2, \dots, C_m$.

Step 1: Set $t = 1$ and $P = P_0 = [(\epsilon, \mathrm{Id})]$.

Step 2: If $P_{t-1} = [w_1, w_2, \dots, w_s]$, define

$$P_t = [w_1 1, \dots, w_1 m, w_2 1, \dots, w_2 m, \dots, w_s 1, \dots, w_s m].$$

Step 3: Proceeding through the list P_t, check if an entry (w, C_w) is a pivot. This can be done in polynomial time, as we have to simply check if C_w is a linear combination of smaller pivots. If it is a pivot, add it to P. If it is not a pivot, then remove it from P_t. Upon completing this step, the list P_t contains all the pivots of length t, and the list P contains all pivots of length $\leq t$.

Step 4: If $P_t \neq []$, set $t = t + 1$ and go back to Step 2. Else, return P and terminate.

Corollary 4.9. *There is a polynomial time algorithm to construct the set of pivots. Further, this algorithm also records the word associated to each pivot.*

Proof. To show that the above algorithm runs in polynomial time, it suffices to show that the number of words we consider is at most polynomial. Indeed, if there are k pivots of length d, then we only consider km words of length $d+1$. Since $k \leq n^2$, the number of words we consider in each degree is at most n^2m. We only consider words of length up to $2n\log_2(n)+4n-4$. Hence, the number of words considered is polynomial (in n and m). □

5. Orbit closure problem for matrix invariants

Let $A, B \in \mathrm{Mat}_{n,n}^m$ with $A = (A_1, \ldots, A_m)$ and $B = (B_1, \ldots, B_m)$. Define

$$C_i = \begin{pmatrix} A_i & 0 \\ 0 & B_i \end{pmatrix}$$

for all i. Let $\mathcal{C}$ be the algebra generated by $C_1, C_2, \ldots, C_m$. Let $Z_1, Z_2, \ldots, Z_s$ be the pivot basis of $\mathcal{C}$ and write

$$Z_j = \begin{pmatrix} X_j & 0 \\ 0 & Y_j \end{pmatrix}$$

for all j.

Proposition 5.1. *Suppose* $\mathrm{char}(K) = 0$. *Then we have* $A \sim_C B$ *if and only if* $\mathrm{Tr}(X_j) = \mathrm{Tr}(Y_j)$ *for all* j.

Proof. Two orbit closures do not intersect if and only if there is an invariant that separates them. By Theorem 2.1, the invariant ring is generated by invariants of the form $X \mapsto \mathrm{Tr}\, X_w$ for some word w in the alphabet $\{1, 2, \ldots, m\}$. Note that $\mathcal{C}$ is the span of all

$$C_w = \begin{pmatrix} A_w & 0 \\ 0 & B_w \end{pmatrix},$$

where w is a word. Now the proposition follows by linearity of trace. □

We will appeal to a result from [Cohen et al. 1997] in order to get a version of the above proposition in arbitrary characteristic; see also [Procesi 1974].

Theorem 5.2. *We have* $A \sim_C B$ *if and only if* $\det(\mathrm{Id} + tX_j) = \det(\mathrm{Id} + tY_j)$ *as a polynomial in* t *for all* j.

Proof. Let F_m denote free algebra generated by m elements $f_1, \ldots, f_m$. From Section 1A1, recall that A (resp. B) gives rise to a representation V_A (resp. V_B) of F_m. Recall from Proposition 1.10 that the orbit closures of A and B intersect if and only if V_A and V_B have the same associated semisimple representation. It is clear that for both V_A and V_B, the action of F_m factors through the surjection $F_m \to \mathcal{C}$ given by $f_i \mapsto C_i$.

Thus it suffices to check whether V_A and V_B have the same associated semisimple representation as $\mathcal{C}$-modules; see Remark 1.12. The theorem now is just the statement of [Cohen et al. 1997, Corollary 12] for the finite-dimensional algebra $\mathcal{C}$. □

Proof of Theorem 1.13. Given $A, B \in \mathrm{Mat}_{n,n}^m$, let $C_i = \left(\begin{smallmatrix} A_i & 0 \\ 0 & B_i \end{smallmatrix}\right)$. Let $\mathcal{C}$ be the subalgebra generated by $C_1, \ldots, C_m$. Construct the pivot basis $Z_1, \ldots, Z_s$ of $\mathcal{C}$. For all j, let $Z_j = \left(\begin{smallmatrix} X_j & 0 \\ 0 & Y_j \end{smallmatrix}\right)$. Further for each j, we have $Z_j = C_{w_j}$ for some word $w_j \in [m]^\star$, and consequently $X_j = A_{w_j}$ and $Y_j = B_{w_j}$.

If $\mathrm{char}(K) = 0$, we only need to check if $\mathrm{Tr}(X_j) = \mathrm{Tr}(Y_j)$. If they are equal for all j, then we have $A \sim_C B$. Else, we have $\mathrm{Tr}(X_j) \neq \mathrm{Tr}(Y_j)$ for some j, i.e., $T_{w_j}(A) \neq T_{w_j}(B)$ and $A \not\sim_C B$.

For arbitrary characteristic, we need to check instead if $\det(\mathrm{Id}+tX_j) = \det(\mathrm{Id}+tY_j)$ as a polynomial in t for each j. But this can be done efficiently. When $A \not\sim_C B$, the algorithm finds j with $1 \leq j \leq n$ and $w \in [m]^\star$ such that $\sigma_{j,w}(A) \neq \sigma_{j,w}(B)$. This means that $\sigma_{j,w} \in S(n, m)$ is an invariant that separates A and B. □

We will now prove the bounds for separating invariants. For $A, B \in \mathrm{Mat}_{n,n}^m$ with $A \not\sim_C B$, we will write $C_i = \left(\begin{smallmatrix} A_i & 0 \\ 0 & B_i \end{smallmatrix}\right)$ and define $\mathcal{C} \subseteq \mathrm{Mat}_{2n,2n}$ to be the subalgebra generated by $C_1, \ldots, C_m$.

Proof of Theorem 1.14. Given $A, B \in \mathrm{Mat}_{n,n}$ with $A \not\sim_C B$, let $\{C_1, \ldots, C_m\} \subseteq \mathrm{Mat}_{2n,2n}$ be as above, and construct the pivot basis for $\mathcal{C}$. We know, by Lemma 4.7, that the length of every pivot is at most $2(2n)\log_2(2n) + 4(2n) - 4 = 4n\log_2(n) + 12n - 4$.

If $\mathrm{char}(K) = 0$, then an invariant T_w separates A and B for some pivot w. This means there is an invariant of degree $\deg(T_w) = l(w) \leq 4n\log_2(n) + 12n - 4$ that separates them.

If $\mathrm{char}(K) > 0$, we must have $\det(\mathrm{Id}+tA_w) \neq \det(\mathrm{Id}+tB_w)$ for some pivot w. Hence for some $1 \leq j \leq n$, $\sigma_{j,w}(A) \neq \sigma_{j,w}(B)$. This gives an invariant of degree $\leq 4n^2\log_2(n) + 12n^2 - 4n$ that separates them. □

Remark 5.3. The null cone for the simultaneous conjugation action of GL_n on $\mathrm{Mat}_{n,n}^m$ is in fact defined by invariants of degree $\leq 2n\log_2(n) + 4n - 4$ in characteristic 0. To see this, we will use a similar argument as in the proof of Theorem 1.14 above. For A that is not in the null cone, simply consider the subalgebra $\mathcal{A} \subseteq \mathrm{Mat}_{n,n}$ generated by $A_1, \ldots, A_m$. For some pivot w, the invariant T_w does not vanish on A. Every pivot has length at most $2n\log_2(n) + 4n - 4$, so this gives the bound on the null cone. Similarly, in positive characteristic, we can get a bound of $2n^2\log_2(n) + 4n^2 - 4n$, but better bounds are already known; see [Derksen and Makam 2017a].

5A. ***Nonalgebraically closed fields.*** Suppose L is a subfield of (an algebraically closed field) K, and suppose $A, B \in \mathrm{Mat}_{n,n}^m(L)$. Let us assume L is infinite and that we use the unit cost arithmetic model for operations in L.

First, we observe that the entire algorithm for both matrix invariants and matrix semi-invariants can be run using only operations in L, and is polynomial time in this unit cost arithmetic model. However, we should point out that the algorithm does not check whether the orbit closures of A and B for the action of $\mathrm{GL}_n(L)$ intersect. Instead, it checks whether the orbit closures of A and B for the action of $\mathrm{GL}_n(K)$ intersect.

Finally, if we take $L = \mathbb{Q}$, the run times of our algorithms for matrix invariants as well as matrix semi-invariants will be polynomial in the bit length of the inputs.

Remark 5.4. We can relax the hypothesis on L by asking for L to be sufficiently large. For fields that are too small, the algorithms will run into issues — for example, the IQS algorithm (Theorem 2.8) requires a sufficiently large field.

6. Bounds for separating matrix semi-invariants

The reduction given in Section 3B is good enough for showing that the orbit closure problems for matrix invariants and matrix semi-invariants are in the same complexity class. In this section we give a stronger reduction with the aim of finding better bounds for the degree of separating invariants for matrix semi-invariants. This reduction can also be made algorithmic, and can replace the reduction in Section 3B. However, we will only focus on obtaining bounds for separating invariants.

Let $T \in \mathrm{Mat}_{d,d}^m$. For $X \in \mathrm{Mat}_{n,n}^m$, consider

$$L_T(X) = \sum_{k=1}^{m} T_k \otimes X_k = \begin{pmatrix} L_{1,1}(X) & \dots & L_{1,d}(X) \\ \vdots & \ddots & \vdots \\ L_{d,1}(X) & \dots & L_{d,d}(X) \end{pmatrix},$$

where $L_{i,j}(X)$ represents an $n \times n$ block. From the definition of Kronecker product of matrices, one can check that $L_{i,j}(X) = \sum_{k=1}^m (T_k)_{i,j} X_k$, i.e., a linear combination of the X_i. By definition $f_T(X) = \det\big(\sum_{k=1}^m T_k \otimes X_k\big) = \det(L_T(X))$. Let

$$M_T(X) = \mathrm{Adj}(L_T(X)) = \begin{pmatrix} M_{1,1}(X) & \dots & M_{1,d}(X) \\ \vdots & \ddots & \vdots \\ M_{d,1}(X) & \dots & M_{d,d}(X) \end{pmatrix},$$

where $M_{i,j}(X)$ represents an $n \times n$ block. The entries of $M_T(X)$ are not linear in the entries of the matrices X_k. Instead the entries are polynomials of degree $dn-1$ in the $(X_k)_{i,j}$'s. We first compute how $M_{i,j}$ change under the action of $\mathrm{SL}_n \times \mathrm{SL}_n$.

Lemma 6.1. *Let $\sigma = (P, Q^{-1}) \in \mathrm{SL}_n \times \mathrm{SL}_n$. Then we have $M_{i.j}(\sigma \cdot X) = Q^{-1} M_{i,j}(X) P^{-1}$.*

Proof. First, observe that $L_T(\sigma \cdot X) = (P \otimes \mathrm{Id}) L_T(X) (Q \otimes \mathrm{Id})$ follows because $L_T(X)$ is a block matrix where each block is a linear combination of the X_i's. Thus we have

$$\begin{aligned} M_T(\sigma \cdot X) &= \mathrm{Adj}(L_T(\sigma \cdot X)) \\ &= \mathrm{Adj}((P \otimes \mathrm{Id}) L_T(X) (Q \otimes \mathrm{Id})) \\ &= \mathrm{Adj}(Q \otimes \mathrm{Id}) M_T(X) \,\mathrm{Adj}(P \otimes \mathrm{Id}) \\ &= (Q^{-1} \otimes \mathrm{Id}) M_T(X) (P^{-1} \otimes \mathrm{Id}) \end{aligned}$$

The last equality follows from Lemma 2.16 because $\det(P \otimes \mathrm{Id}) = \det(Q \otimes \mathrm{Id}) = 1$. We deduce that $M_{i.j}(\sigma \cdot X) = Q^{-1} M_{i,j}(X) P^{-1}$. □

For $X \in \mathrm{Mat}_{n,n}^m$, let us define

$$X_{i,j,k} = X_k M_{i,j}(X),$$

for $1 \le k \le m$ and $1 \le i, j \le d$.

The $X_{i,j,k}$'s have been designed in such a way that the left-right action on X_i's turns into a conjugation action on the $X_{i,j,k}$'s. Further, the entries of $X_{i,j,k}$ are degree dn polynomials in the entries of the X_l's.

Corollary 6.2. $(\sigma \cdot X)_{i,j,k} = PX_{i,j,k}P^{-1}$.

Proof. It follows from the above lemma that

$$(\sigma \cdot X)_{i,j,k} = (\sigma \cdot X)_k M_{i,j}(\sigma \cdot X) = (PX_kQ)(Q^{-1}M_{i,j}(X)P^{-1}) = PX_{i,j,k}P^{-1}. \qquad \square$$

Consider the map $\zeta : \mathrm{Mat}_{n,n}^m \to \mathrm{Mat}_{n,n}^{md^2}$ given by $X \mapsto (X_{i,j,k})_{i,j,k}$. This gives a map on the coordinate rings $\zeta^* : K[\mathrm{Mat}_{n,n}^{md^2}] \to K[\mathrm{Mat}_{n,n}^m]$. We note that ζ is a map of degree dn because the entries of $X_{i,j,k}$ are degree dn polynomials in the entries of the X_l's.

The above corollary can be now reformulated as:

Corollary 6.3. *Let $\sigma = (P, Q^{-1}) \in \mathrm{SL}_n \times \mathrm{SL}_n$. Then we have $\zeta(\sigma \cdot X) = P\zeta(X)P^{-1}$.*

Proposition 6.4. *The map ζ^* descends to a map on invariant rings $\zeta^* : S(n, md^2) \to R(n, m)$.*

Proof. Let $\sigma = (P, Q^{-1}) \in \mathrm{SL}_n \times \mathrm{SL}_n$. For $g \in S(n, md^2)$, by the above corollary, we have $g(\zeta(\sigma \cdot X)) = g(P\zeta(X)P^{-1}) = g(\zeta(X))$. Now observe that $\zeta^*(g) \in R(n, m)$ since $\zeta^*(g)(\sigma \cdot X) = g(\zeta(\sigma \cdot X)) = g(\zeta(X)) = \zeta^*(g)(X)$. $\square$

Observe that this is a very different map from the one in Proposition 2.15. We will still be able to use it to get separating invariants for left-right action from separating invariants for the conjugation action. We make an obvious observation.

Corollary 6.5. *Suppose we have $g \in S(n, md^2)$ such that $\zeta^*(g)(A) \neq \zeta^*(g)(B)$, then $A \not\sim_{LR} B$.*

Remark 6.6. In order for the above corollary to be useful to get separating invariants, we need to be able to guarantee that separating invariants will arise this way. In other words, for $A \not\sim_{LR} B$, we want $g \in S(n, md^2)$ such that $\zeta^*(g)$ separates A and B. We will only be able to do it under certain conditions, but that will be sufficient.

The first issue to notice is that since ζ^* is a map of degree dn, any homogeneous invariant of the form $\zeta^*(g)$ must have degree dkn for some $k \in \mathbb{Z}_{\geq 0}$. For a graded ring $R = \oplus_{t \in \mathbb{Z}} R_t$, let us define its k-th Veronese subring $v_k(R) := \bigoplus_{t \in \mathbb{Z}} R_{tk}$.

Lemma 6.7. *We have $\zeta^* : S(n, md^2) \to v_{dn}(R(n, m)) \hookrightarrow R(n, m)$.*

It is certainly possible that for some d, no invariant of degree dkn separates A and B. A simple example is given by taking any A not in the null cone, and taking B such that $B_i = \mu_d A_i$, where μ_d is a d-th root of unity for some d coprime to n. Hence, we may have to consider more than one choice of d.

For the following lemma, any two coprime numbers can be used in place of $n-1$ and n, but this is the smallest pair of coprime numbers larger than $n-1$. The significance of $n-1$ is that as long as $d \geq n-1$, for any A not in the null cone, we can guarantee the existence of an invariant f_T, with $T \in \mathrm{Mat}_{d,d}^m$ such that $f_T(A) \neq 0$; see Theorem 2.7.

Lemma 6.8. *Assume $A, B \in \mathrm{Mat}_{n,n}^m$ and assume $A \not\sim_{LR} B$. Then $\bigcup_{d \in \{n-1, n\}} v_{dn}(R(n, m))$ form a set of separating invariants.*

Proof. Since $A \not\sim_{LR} B$, there is a choice of $S \in \mathrm{Mat}_{k,k}^m$, for some $k \geq 1$, such that $f_S(A) \neq f_S(B)$. Without loss of generality, assume $f_S(B) \neq 0$. Hence $f_S(A)/f_S(B) \neq 1$. Once again we must have $f_S(A)^d/f_S(B)^d \neq 1$ for at least one choice of $d \in \{n-1, n\}$. In particular, for such a d, $(f_S)^d \in \nu_{dn}(R(n,m))$ separates A and B. □

Once we have d such $\nu_{dn}(R(n,m))$ separates A and B, we still need to produce such an invariant that separates A and B. Once, we restrict our attention to invariants whose degree is a multiple of dn, the best case scenario is that there is a degree dn invariant that separates A and B. We will construct an invariant of the form $\zeta^*(g)$ that separates A and B when degree dn invariants fail to separate A and B. The following lemma completes the strategy outlined in Remark 6.6.

Lemma 6.9. *Let $A, B \in \mathrm{Mat}_{n,n}^m$ such that $A \not\sim_{LR} B$. Suppose we have $d \geq n-1$ such that $\nu_{dn}(R(n,m))$ separates A and B. Then $R(n,m)_{dn} \cup \zeta^*(S(n, md^2))$ will separate A and B.*

Proof. Assume that $R(n,m)_{dn}$ fails to separate A and B. We will find $g \in S(n, md^2)$ such that $\zeta^*(g)$ separates A and B.

Since both A and B cannot be in the null cone, we can assume without loss of generality that A is not in the null cone. By Theorem 2.7, we have $T \in \mathrm{Mat}_{d,d}^m$, such that $f_T(A) \neq 0$. Now, since degree dn invariants fail to separate A and B, we must have $f_T(A) = f_T(B) \neq 0$.

There exists $U \in \mathrm{Mat}_{dk,dk}^m$ such that $f_U(A) \neq f_U(B)$ since such invariants span $\nu_{dn}(R(n,m))$, which by assumption separates A and B. Now for $X \in \mathrm{Mat}_{n,n}^m$, define $\mathcal{L}(X) := \sum_{k=1}^m U_k \otimes X_k$ and $\mathcal{R}(X) := \mathrm{Id}_k \otimes M_T(X)$. Let

$$N(X) := \mathcal{L}(X)\mathcal{R}(X) = \left(\sum_{k=1}^m U_k \otimes X_k\right)(\mathrm{Id}_k \otimes M_T(X))$$

Let us make some observations to help understand $N(X)$.

- The matrix $\mathcal{L}(X) = \sum_{k=1}^m U_k \otimes X_k$ can be seen as a $dk \times dk$ block matrix, where each block has size $n \times n$. Further, each block is a linear combination of the X_k's.
- The matrix $\mathcal{R}(X) = \mathrm{Id}_k \otimes M_T(X)$ can be seen as a $k \times k$ block matrix, where the off diagonal blocks are 0, and the diagonal blocks are a copy of $M_T(X)$. Observe further that $M_T(X)$ is a $d \times d$ block matrix, where each block $M_{i,j}$ is of size $n \times n$ as shown above. Hence, we can see $\mathcal{R}(X)$ as a $dk \times dk$ block matrix, where each block is of size $n \times n$ and is either $M_{i,j}$ or 0.
- A product of a block from $\mathcal{L}(X)$ and a block from $\mathcal{R}(X)$ yields a linear combination of terms of the form $X_k M_{i,j}$'s, i.e., a linear combination of the $X_{i,j,k}$'s.
- We can obtain $N(X)$ as a $dk \times dk$ block matrix by block multiplying $\mathcal{L}(X)$ and $\mathcal{R}(X)$. Hence, we see that each block of $N(X)$ is a linear combination of the $X_{i,j,k}$'s.

To summarize, $N(X)$ is a $dk \times dk$ block matrix and the size of each block is $n \times n$. Further, the (p,q)-th block $N(X)_{p,q}$ is a linear combination $\sum_{i,j,k} \lambda_{p,q}^{i,j,k} X_{i,j,k}$ for some $\lambda_{p,q}^{i,j,k} \in K$. Now we can define an invariant $g \in S(n, md^2)$. For $Z = (Z_{i,j,k})_{i,j,k} \in \mathrm{Mat}_{n,n}^{md^2}$, we define N_Z to be the $dk \times dk$ block matrix, where the (p,q)-th block is given by $\sum_{i,j,k} \lambda_{p,q}^{i,j,k} Z_{i,j,k}$. Let $g(Z) = \det(N_Z)$. This is the required g.

The point to note here is that by construction, we have $N_{\zeta(X)} = N(X)$. Thus $\zeta^*(g)(X) = g(\zeta(X)) = \det(N_{\zeta(X)}) = \det(N(X))$.

There are two things we need to check. First that g as defined is indeed invariant under simultaneous conjugation, and then that $\zeta^*(g)(X) = \det(N(X))$ does separate A and B.

The function g is invariant under the simultaneous conjugation action of GL_n on $\mathrm{Mat}_{n,n}^{md^2}$ because it is given by the determinant of a block matrix whose blocks are linear combinations of matrices from the input md^2-tuple.

Observe that $\det(\mathcal{L}(X)) = f_U(X)$ and $\det(\mathcal{R}(X)) = \det(M_T(X))^k$. Therefore we have that $\det(N(X)) = f_U(X)\det(M_T(X))^k$. Recall that $f_T(X) = \det(L_T(X))$, and that $M_T(X) = \mathrm{Adj}(L_T(X))$. Now, since $f_T(A) = f_T(B) \neq 0$, we have that $\det(M_T(A)) = \det(M_T(B)) \neq 0$. In particular, since $f_U(A) \neq f_U(B)$, we have $\det(N(A)) \neq \det(N(B))$ as required.

Thus $\zeta^*(g)(A) = \det(N(A)) \neq \det(N(B)) = \zeta^*(g)$ showing that $\zeta^*(g)$ indeed separates A and B. □

Now, we can finally prove Theorem 1.18.

Proof of Theorem 1.18. Suppose $A, B \in \mathrm{Mat}_{n,n}^m$ with $A \not\sim_{LR} B$. By Lemma 6.8, for at least one choice of $d \in \{n-1, n\}$, we have that $\nu_{dn}(R(n,m))$ separates A and B. Fix this d. By Lemma 6.9, either $R(n,m)_{dn}$ or $\zeta^*(S(n, md^2))$ separates A and B. In the former case, we have an invariant of degree $dn \leq n^2$ that separates A and B. In the latter case, $\zeta^*(S(n, md^2))$ separates A and B which implies that $S(n, md^2)$ separates $\zeta(A)$ and $\zeta(B)$. Hence, we have an invariant $g \in S(n, md^2)$ of degree $\leq \beta_{\mathrm{sep}}(S(n, md^2))$ such that $g(\zeta(A)) \neq g(\zeta(B))$.

Now, since ζ is a map of degree dn, we have $\zeta^*(g) \in R(n,m)$ is a polynomial of degree $\deg(g)dn \leq n^2\beta_{\mathrm{sep}}(S(n, md^2)) \leq n^2\beta_{\mathrm{sep}}(S(n, mn^2))$ that separates A and B. □

Remark 6.10. It is easy to see from Theorem 2.3 that the statement of Theorem 2.1 holds if we assume $\mathrm{char}(K) > n$; see also [Zubkov 1993]. Hence, the statements in Theorem 1.14 and Corollary 1.19 that assumed $\mathrm{char}(K) = 0$ also hold under the assumption that $\mathrm{char}(K) > n$.

Acknowledgements

We thank the authors of [Allen-Zhu et al. 2018] for sending us an early version of their paper. We thank Gábor Ivanyos and Gregor Kemper for providing useful references. Finally, we thank the anonymous referee for several useful suggestions on improving the exposition.

References

[Allen-Zhu et al. 2018] Z. Allen-Zhu, A. Garg, Y. Li, R. Oliveira, and A. Wigderson, "Operator scaling via geodesically convex optimization, invariant theory and polynomial identity testing", pp. 172–181 in *STOC'18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (Los Angeles, CA), edited by I. Diakonikolas et al., ACM, New York, 2018. MR Zbl

[Artin 1969] M. Artin, "On Azumaya algebras and finite dimensional representations of rings", *J. Algebra* **11** (1969), 532–563. MR Zbl

[Brooksbank and Luks 2008] P. A. Brooksbank and E. M. Luks, "Testing isomorphism of modules", *J. Algebra* **320**:11 (2008), 4020–4029. MR Zbl

[Chistov et al. 1997] A. Chistov, G. Ivanyos, and M. Karpinski, "Polynomial time algorithms for modules over finite dimensional algebras", pp. 68–74 in *Proceedings of the 1997 International Symposium on Symbolic and Algebraic Computation* (Kihei, HI), edited by W. W. Küchlin, ACM, New York, 1997. MR Zbl

[Cohen et al. 1997] A. M. Cohen, G. Ivanyos, and D. B. Wales, "Finding the radical of an algebra of linear transformations", *J. Pure Appl. Algebra* **117/118** (1997), 177–193. MR Zbl

[Derksen 2001] H. Derksen, "Polynomial bounds for rings of invariants", *Proc. Amer. Math. Soc.* **129**:4 (2001), 955–963. MR Zbl

[Derksen and Kemper 2002] H. Derksen and G. Kemper, *Computational invariant theory*, Encyclopaedia of Mathematical Sciences **130**, Springer, 2002. MR Zbl

[Derksen and Makam 2017a] H. Derksen and V. Makam, "Generating invariant rings of quivers in arbitrary characteristic", *J. Algebra* **489** (2017), 435–445. MR Zbl

[Derksen and Makam 2017b] H. Derksen and V. Makam, "Polynomial degree bounds for matrix semi-invariants", *Adv. Math.* **310** (2017), 44–63. MR Zbl

[Derksen and Makam 2018] H. Derksen and V. Makam, "On non-commutative rank and tensor rank", *Linear and Multilinear Algebra* **66**:6 (2018), 1069–1084. MR Zbl

[Derksen and Weyman 2000] H. Derksen and J. Weyman, "Semi-invariants of quivers and saturation for Littlewood–Richardson coefficients", *J. Amer. Math. Soc.* **13**:3 (2000), 467–479. MR Zbl

[Domokos 2000] M. Domokos, "Relative invariants of 3×3 matrix triples", *Linear and Multilinear Algebra* **47**:2 (2000), 175–190. MR Zbl

[Domokos 2018] M. Domokos, "Polynomial bound for the nilpotency index of finitely generated nil algebras", *Algebra Number Theory* **12**:5 (2018), 1233–1242. MR Zbl

[Domokos and Zubkov 2001] M. Domokos and A. N. Zubkov, "Semi-invariants of quivers as determinants", *Transform. Groups* **6**:1 (2001), 9–24. MR Zbl

[Donkin 1992] S. Donkin, "Invariants of several matrices", *Invent. Math.* **110**:2 (1992), 389–401. MR Zbl

[Donkin 1993] S. Donkin, "Invariant functions on matrices", *Math. Proc. Cambridge Philos. Soc.* **113**:1 (1993), 23–43. MR Zbl

[Draisma 2006] J. Draisma, "Small maximal spaces of non-invertible matrices", *Bull. London Math. Soc.* **38**:5 (2006), 764–776. MR Zbl

[Eisenbud and Harris 1988] D. Eisenbud and J. Harris, "Vector spaces of matrices of low rank", *Adv. in Math.* **70**:2 (1988), 135–155. MR Zbl

[Forbes and Shpilka 2013] M. A. Forbes and A. Shpilka, "Explicit Noether normalization for simultaneous conjugation via polynomial identity testing", pp. 527–542 in *Approximation, randomization, and combinatorial optimization*, edited by P. Raghavendra et al., Lecture Notes in Comput. Sci. **8096**, Springer, 2013. MR Zbl

[Formanek 1986] E. Formanek, "Generating the ring of matrix invariants", pp. 73–82 in *Ring theory* (Antwerp, 1985), edited by F. M. J. Van Oystaeyen, Lecture Notes in Math. **1197**, Springer, 1986. MR Zbl

[Garg et al. 2016] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson, "A deterministic polynomial time algorithm for non-commutative rational identity testing", pp. 109–117 in *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016* (New Brunswick, NJ), IEEE Computer Soc., Los Alamitos, CA, 2016. MR

[Haboush 1975] W. J. Haboush, "Reductive groups are geometrically reductive", *Ann. of Math.* (2) **102**:1 (1975), 67–83. MR Zbl

[Hilbert 1890] D. Hilbert, "Ueber die Theorie der algebraischen Formen", *Math. Ann.* **36**:4 (1890), 473–534. MR Zbl

[Hilbert 1893] D. Hilbert, "Ueber die vollen Invariantensysteme", *Math. Ann.* **42**:3 (1893), 313–373. MR Zbl

[Hrubeš and Wigderson 2014] P. Hrubeš and A. Wigderson, "Non-commutative arithmetic circuits with division", pp. 49–65 in *ITCS'14—Proceedings of the 2014 Conference on Innovations in Theoretical Computer Science* (Princeton, NJ), ACM, New York, 2014. MR Zbl

[Ivanyos et al. 2017] G. Ivanyos, Y. Qiao, and K. V. Subrahmanyam, "Non-commutative Edmonds' problem and matrix semi-invariants", *Comput. Complexity* **26**:3 (2017), 717–763. MR Zbl

[Ivanyos et al. 2018] G. Ivanyos, Y. Qiao, and K. V. Subrahmanyam, "Constructive non-commutative rank computation is in deterministic polynomial time", *Comput. Complexity* **27**:4 (2018), 561–593. MR Zbl

[Kemper 2003] G. Kemper, "Computing invariants of reductive groups in positive characteristic", *Transform. Groups* **8**:2 (2003), 159–176. MR Zbl

[King 1994] A. D. King, "Moduli of representations of finite-dimensional algebras", *Quart. J. Math. Oxford Ser.* (2) **45**:180 (1994), 515–530. MR Zbl

[Kuzmin 1975] E. N. Kuzmin, "On the Nagata–Higman theorem", pp. 101–107 in *Mathematical Structures–Computational Mathematics–Mathematical Modelling*, edited by B. Sendov, Bulgarian Acad. Sci., Sofia, 1975. In Russian.

[Mulmuley 2017] K. D. Mulmuley, "Geometric complexity theory V: Efficient algorithms for Noether normalization", *J. Amer. Math. Soc.* **30**:1 (2017), 225–309. MR Zbl

[Mumford et al. 1994] D. Mumford, J. Fogarty, and F. Kirwan, *Geometric invariant theory*, 3rd ed., Ergebnisse der Mathematik und ihrer Grenzgebiete (2) **34**, Springer, 1994. MR Zbl

[Nagata 1963/64] M. Nagata, "Invariants of a group in an affine ring", *J. Math. Kyoto Univ.* **3** (1963/64), 369–377. MR

[Pappacena 1997] C. J. Pappacena, "An upper bound for the length of a finite-dimensional algebra", *J. Algebra* **197**:2 (1997), 535–545. MR Zbl

[Procesi 1974] C. Procesi, "Finite dimensional representations of algebras", *Israel J. Math.* **19** (1974), 169–182. MR Zbl

[Procesi 1976] C. Procesi, "The invariant theory of $n \times n$ matrices", *Advances in Math.* **19**:3 (1976), 306–381. MR Zbl

[Raz and Shpilka 2005] R. Raz and A. Shpilka, "Deterministic polynomial identity testing in non-commutative models", *Comput. Complexity* **14**:1 (2005), 1–19. MR Zbl

[Razmyslov 1974] Y. Razmyslov, "Trace identities of full matrix algebras over a field of characteristic zero", *Math. USSR-Izv.* **8**:4 (1974), 727–760.

[Schofield and van den Bergh 2001] A. Schofield and M. van den Bergh, "Semi-invariants of quivers for arbitrary dimension vectors", *Indag. Math.* (*N.S.*) **12**:1 (2001), 125–138. MR Zbl

[Shitov 2019] Y. Shitov, "An improved bound for the lengths of matrix algebras", *Algebra Number Theory* **13**:6 (2019), 1501–1507. MR Zbl

[Sibirskiĭ 1968] K. S. Sibirskiĭ, "Algebraic invariants of a system of matrices", *Sibirsk. Mat. Ž.* **9** (1968), 152–164. MR

[Valiant 1979] L. G. Valiant, "The complexity of computing the permanent", *Theoret. Comput. Sci.* **8**:2 (1979), 189–201. MR Zbl

[Zubkov 1993] A. N. Zubkov, "Matrix invariants over an infinite field of finite characteristic", *Sibirsk. Mat. Zh.* **34**:6 (1993), 68–74, ii, viii. MR Zbl

Communicated by Michel Van den Bergh
Received 2019-10-24 Revised 2020-03-15 Accepted 2020-06-20

hderksen@umich.edu *Department of Mathematics, University of Michigan, Ann Arbor, MI, United States*

visu@ias.edu *School of Mathematics, Institute for Advanced Study, Princeton, NJ, 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. 10 2020

1937-0652(2020)14:10;1-2