ON THE FREY-MAZUR CONJECTURE OVER LOW GENUS CURVES

BENJAMIN BAKKER AND JACOB TSIMERMAN

Abstract. The Frey–Mazur conjecture states that an elliptic curve over $\mathbb{Q}$ is determined up to isogeny by its p-torsion Galois representation for $p \geq 19$. We study a geometric analog of this conjecture, and show that the map from isogeny classes of "fake elliptic curves"—abelian surfaces with quaternionic multiplication—to their p-torsion Galois representations is one-to-one over function fields of small genus complex curves for sufficiently large p relative to the genus.

1. Introduction

The Frey-Mazur conjecture, originating in [MG78], states[1] that for a prime $p \geq 19$, an elliptic curve over $\mathbb{Q}$ is classified up to isogeny by its p-torsion, viewed as a Galois representation (or equivalently, as a finite flat group scheme). Geometrically, there is a surface $Z(p)$ that parameterizes pairs of elliptic curves (E, E') together with an isomorphism φ of their p-torsion, and this surface is endowed with natural Hecke divisors H_M parametrizing points for which φ is induced by an isogeny of degree M. The conjecture is equivalent to the statement that for $p \geq 19$, all rational points of $Z(p)$ lie on one of these divisors[2]. Since by work of Hermann [Her91] the surface $Z(p)$ is of general type for $p > 11$, the Bombieri-Lang conjecture implies that there are only finitely many rational points on the complement of the union of all rational and elliptic curves in $Z(p)$. Hence, it becomes natural to first consider the Frey-Mazur conjecture over the function fields of curves of genus at most 1.

Rather than work with elliptic curves themselves, we instead work with what are often called "fake elliptic curves": abelian surfaces with an action by a maximal order $\mathcal{O}_D$ in a quaternion algebra. Such abelian surfaces are also parametrized by a one-dimensional Shimura variety X^D, but crucially for us these curves are *compact*. There is an obvious natural analog of the Frey-Mazur conjecture in this setting as well. Our main result is:

Theorem 1 (see Theorem 19)**.** *For any $k > 0$, there exists $N > 0$ such that for any smooth quasiprojective complex curve B of genus $g < k$ and any two abelian surfaces A_1, A_2 over B with $\mathcal{O}_D$-actions whose p-torsion local systems $A_i[p]$ are isomorphic (as $\mathcal{O}_D$-modules), A_1 and A_2 are $\mathcal{O}_D$-isogenous provided $p > N$.*

The statement of Theorem 1 is equivalent to the assertion that any map from a curve of genus $g < k$ to the modular surface $Z^D(p)$ analogous to $Z(p)$ above

Date: November 20, 2015.

[1]The exact bound on p is often left inexplicit; a counterexample for $p = 17$ has been communicated to us by Billerey [Bil].

[2]Note that by work of Mazur, it is only necessary to consider $M \leq 163$

lies in a Hecke divisor H for $p \gg 0$. The proof of Theorem 1 is substantially easier assuming the base Shimura curve X^D has genus ≥ 2 (see Corollary 18), as the main difficulty is understanding rational and elliptic curves in $Z^D(p)$. A catalog of low genus Shimura curves can be found in [Voi09].

Theorem 1 is ostensibly about curves in $Z^D(p)$, however, it is easier to understand curves in the product $X^D(p) \times X^D(p)$, where $X^D(p)$ parameterizes such abelian surfaces A rigidified by an $\mathcal{O}_D$-isomorphism $A[p] \cong \mathcal{O}_D[p]$—equivalently an element $z \in A[p]$ generating $A[p]$ over $\mathcal{O}_D[p]$. The surface $Z^D(p)$ is naturally the quotient of $X^D(p) \times X^D(p)$ only remembering the composition $A_1[p] \xrightarrow{\cong} \mathcal{O}_D[p] \xrightarrow{\cong} A_2[p]$.

The main idea of the proof of Theorems 1 runs as follows: given a curve B in $Z^D(p)$ we lift it to a curve C in the surface $X^D(p) \times X^D(p)$ and estimate its genus using Riemann-Hurwitz in 2 different ways. We obtain a lower bound from the projections to the curves $X^D(p)$ simply by ignoring ramification. The lower bound only becomes useful once we know that the bidegree of non-Hecke curves has to be large. We deduce this from a theorem of Andre and Deligne [And92] which says that the image of the fundamental group of a non-Hecke curve is Zariski dense. This argument alone is enough to conclude Theorem 1 when $g(X^D) > 1$.

The upper bound requires a bound on the ramification divisor of $C \to B$, which can only be supported at the singular points of $X^D(p) \times X^D(p) \to Z^D(p)$, so we look to bound the number of times C can pass through this set. The singularities naturally split into two sets which, following Kani and Schanz [KS98], we label the "Heegner" and "anti-Heegner" CM points. We show that with respect to the hyperbolic metric, the Heegner CM points are far away from each other, except for a set which lie on low degree Hecke curves. We then use work of Hwang and To [HT02, HT12] to show that curves C with high incidence along the Heegner CM points must have large volume. The absence of Hecke curves passing through the anti-Heegner CM points requires us to prove an analogous bound on the volume of the curve C near the *conjugate* Hecke curves.

1.1. **The elliptic curve case.**

Since the writing of this preprint, the authors have proven [BT14] the analog of Theorem 1 for elliptic curves, namely that two elliptic curves over the function field of a complex curve B with isomorphic p-torsion are isogenous provided p is larger than a constant N. The proof follows the same strategy as that outlined above, though the analysis is substantially complicated by the existence of cusps on the modular curves $X(p)$. Furthermore, it is shown there that the constant N can be taken to depend only on the gonality of B, which is the analog of the degree of a number field in the function field setting. Though this preprint is largely subsumed and substantially generalized by [BT14], it has two advantages:

- The Shimura case simply exhibits the core idea;
- The argument of Section 5 is not needed in [BT14] but may still be of interest.

We note here that using the techniques of [BT14], the constant N of Theorem 1 can be likewise taken to depend only on the gonality of B (*cf.* Remark 20). We

also expect the method of this paper to work for all compact Shimura curves but do not pursue this here.

1.2. **Outline.**

We now give an outline of the rest of the paper. In Section 2 we recall background on quaternion algebras, Shimura modular curves and level structures. We carefully treat the uniformization of these curves, and classify the points with additional automorphisms (the "Heegner" and "anti-Heegner" CM points). In Section 3, we prove that those Heegner CM points that are not well spread out lie on low degree Hecke curves, and in Section 4 we use this to bound the incidence of non-Hecke curves $C \subset X^D(p) \times X^D(p)$ along the CM points. In Section 5 we show that the bidegree of non-Hecke curves $B \subset Z^D(p)$ grows with p. Section 6 contains the proof of Theorem 1.

1.3. **Acknowledgements.**

The authors benefited from many useful conversations with Fedor Bogomolov, Johan de Jong, Michael McQuillan, Allison Miller, and Peter Sarnak. The first named author was supported by NSF fellowship DMS-1103982.

1.4. **Notation.**

Throughout the paper we use the following notation regarding regarding asymptotic growth: For functions f, g we write $f \gg g$ if there is a positive constant $L > 0$ such that $f - Lg$ is a positive function; likewise for $\ll$. We may also sometimes write $f = O(g)$ instead of $f \ll g$. If f_t, g_t are functions depending on t, we write $f_t = o(g_t)$ as $t \to \infty$ to mean that for all $L > 0$, there exists $N > 0$ such that $g_t - Lf_t$ is positive, provided $t > N$.

2. Shimura Modular Curves

We begin by briefly recalling the theory of Shimura modular curves over $\mathbb{Q}$. Our main reference is [Mil97, Chapter 4] for quaterion algebras and [Cla03], [Elk98] for Shimura curves.

2.1. **Quaternion algebras.**

Let k be a field of characteristic $\operatorname{char} k \neq 2$. Recall that a quaterion algebra D/k over a field k is a central simple algebra over k with $\dim_k D = 4$. The trivial (or split) quaterion algebra is $D = M_2(k)$, the algebra of 2 by 2 matrices over k. Given an extension K/k, we say D is split over K if $D \otimes_k K \cong M_2(K)$, and we similarly define D to be split at a place v of k if $D \otimes k_v$ is split. A quaternion algebra D over $\mathbb{Q}$ is indefinite if it is split at the infinite place.

We can construct quaternion algebras analogously to the usual Hamiltonian quaternions. For $\alpha, \beta \in k$ we define $\left(\frac{\alpha,\beta}{k}\right)$ to be the quaternion algebra with k-basis $1, i, j, ij$ and relations

$$i^2 = \alpha, \qquad j^2 = \beta, \qquad ij = -ji$$

For example, $\left(\frac{1,1}{k}\right) \cong M_2(k)$ is split; $\left(\frac{-1,-1}{\mathbb{R}}\right)$ is the usual Hamiltonian quaternions. $D = \left(\frac{\alpha,\beta}{k}\right)$ comes endowed with a canonical involution $\bar{\cdot} : D \xrightarrow{\cong} D^{\mathrm{op}}$ given

by

$$\overline{a+bi+cj+dij} = a-bi-cj-dij$$

With our hypotheses on the characteristic, every quaternion algebra D/k is representable as $\left(\frac{\alpha,\beta}{k}\right)$ for some $\alpha, \beta \in k$. We define the reduced trace and norm to be the maps

$$\mathrm{tr} : D \to k : x \mapsto x + \overline{x}$$

$$\mathrm{N} : D \to k : x \mapsto x\overline{x}$$

So for $D = \left(\frac{\alpha,\beta}{k}\right)$, we have that $\mathrm{tr}(a+bi+cj+dij) = 2a$, and

$$\mathrm{N}(a+bi+cj+dij) = a^2 - \alpha b^2 - \beta c^2 + \alpha\beta d^2$$

For example, the involution of $M_2(k)$ is

$$\overline{\begin{pmatrix} a & b \\ c & d \end{pmatrix}} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

and the reduced trace and norm are simply the trace and determinant, respectively.

Note that $D = \left(\frac{\alpha,\beta}{k}\right)$ is naturally split over $K = k(\sqrt{\alpha})$. Indeed, $K \cong k \oplus ki$ is a subalgebra of D, and the left regular representation $L : D \to \mathrm{End}_K(D)$ mapping $x \in D$ to left multiplication by x becomes an isomorphism over K. Using the K-basis $1, j$, the representation is explicitly given by

$$L((a+bi)+(c+di)j) = \begin{pmatrix} a+bi & b(c-di) \\ c+di & a+bi \end{pmatrix}$$

If k is a number field or a p-adic field, an order $\mathcal{O}$ of a quaternion algebra D/k is a subring containing the ring of integers $\mathcal{O}_k$ of k which is finite as a module over $\mathcal{O}_k$. For example, $M_2(\mathcal{O}_k)$ and $\mathcal{O}_k[i,j]$ are orders in $M_2(k)$ and $\left(\frac{\alpha,\beta}{k}\right)$, respectively. For any order $\mathcal{O}$, we define $\mathcal{O}_+^*$ to be the group of units of positive norm, $\mathcal{O}_1^* \subset \mathcal{O}_+^*$ to be the norm 1 subgroup, and the discriminant $\mathrm{disc}\,\mathcal{O}$ to be its discriminant with respect to the reduced trace form.

2.2. **Shimura modular curves.**

Throughout the remainder of the paper, let $D/\mathbb{Q}$ be a nonsplit indefinite quaternion algebra of discriminant d, and let $\mathcal{O}_D$ be a maximal order of D. Note that because D is indefinite, all of its maximal orders are conjugate [Cla, Theorem 14].

For a variety S, an abelian surface over S with an $\mathcal{O}_D$-action is an abelian scheme A/S of relative dimension 2 with an injective ring homomorphism $\iota : \mathcal{O}_D \hookrightarrow \mathrm{End}_S(A)$. Let $\mathcal{X}^D/\mathbb{Q}$ be the stack of such families. The associated coarse space $X^D/\mathbb{Q}$ is a smooth proper curve, called a Shimura curve.

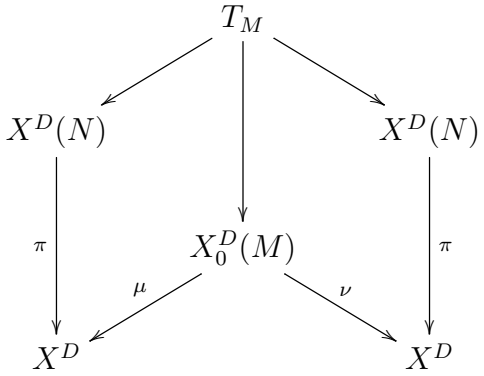
Shimura curves can be thought of loosely as generalized elliptic modular curves. Indeed, one can view the elliptic modular curve $X(1)$ as constructed via the above procedure by taking the maximal order $\mathcal{O}_D = M_2(\mathbb{Z})$ in the split quaternion algebra $D = M_2(\mathbb{Q})$. An abelian surface A with an $\mathcal{O}_D$-action is then forced to be the square of an elliptic curve with the obvious inclusion $\mathcal{O}_D \hookrightarrow \mathrm{End}(A)$.

For N coprime to d, we define $X_0^D(N)/\mathbb{Q}$ to be the coarse space associated to the stack of abelian surfaces with an $\mathcal{O}_D$-action together with a $\mathbb{Z}/N\mathbb{Z}$-rank 2 (left) $\mathcal{O}_D$-submodule V of the N-torsion $A[N]$. $X_0^D(N)$ is a smooth proper curve

and admits two maps: $\mu : X_0^D(N) \to X^D$ forgetting the torsion submodule, and $\nu : X_0^D(N) \to X^D$ sending A to A/V. Since N is coprime to d, D splits over $\mathbb{Q}_p$ for each $p|N$, and therefore $\mathcal{O}_D(\mathbb{Z}/N\mathbb{Z}) \cong M_2(\mathbb{Z}/N\mathbb{Z})$. Thus there are $\prod_{p^e||N}(p^e + p^{e-1})$ such modules V. By analogy with the elliptic modular curve case, we say that A/V is cyclically isogenous to A.

Again for N coprime to d, a full level N structure on an abelian surface A with an $\mathcal{O}_D$ action is an isomorphism $A[N] \cong \mathcal{O}_D[N]$ of $\mathcal{O}_D$-modules, and two level structures are equivalent if the isomorphisms are equal up to scale (*cf.* Remark 6). Equivalently, a full level N structure is an element $v \in A[N]$ such that $\mathcal{O}_D[N]v = A[N]$, defined up to scaling by $(\mathbb{Z}/N)^*$. The Shimura curves with full level structure $X^D(N)$ are then defined as the coarse space associated to the stack $\mathcal{X}^D(N)$ of abelian surfaces A with an $\mathcal{O}_D$-action and endowed with full level N structure.

There is similarly an obvious forgetful map $\pi : X^D(N) \to X^D$, and for any M coprime to N, the two maps $X_0^D(M) \to X^D$ induce a Hecke correspondence T_M (where we drop the N by abuse of notation):



Explicitly, points in the image $T_M \to X^D(N) \times X^D(N)$ are pairs of cyclically isogenous abelian surfaces with an $\mathcal{O}_D$-action and full level N structure such that the isogeny is of degree M and induces an isomorphism of level N structures. By the above, the degree of μ and ν is $\prod_{p^e||N}(p^e + p^{e-1})$.

2.3. **Uniformization.**

Like elliptic modular curves, Shimura curves can also be represented explicitly as quotients of $\mathbb{H}^{\pm} = \mathbb{C}\backslash\mathbb{R}$ by discrete groups of isometries.

Define $\Gamma^D = \mathcal{O}_D^*/\pm 1$. Since D is indefinite, we may choose an isomorphism $\varphi_D : D \otimes \mathbb{R} \xrightarrow{\cong} M_2(\mathbb{R})$, and under this isomorphism $(D \otimes \mathbb{R})^* \cong \mathrm{GL}_2(\mathbb{R})$. Moreover, this induces an inclusion $\Gamma^D \hookrightarrow \mathrm{PGL}_2(\mathbb{R})$ as a discrete cocompact subgroup, and in fact we have the following

Lemma 2. $X^D(\mathbb{C}) \cong \Gamma^D\backslash\mathbb{H}^{\pm}$.

Proof. We can make the above isomorphism explicit. For $z \in \mathbb{H}^{\pm}$ set

$$L_z := \varphi_D(\mathcal{O}_D) \cdot (1, z) \subset \mathbb{C}^2$$

and set $A_z := \mathbb{C}^2/L_z$. This is a complex torus with an $\mathcal{O}_D$-action given by φ_D, and by choosing $\mu \in \mathcal{O}_D$ such that $\mu^2 = -\mathrm{disc(D)}$, A_z can be given the structure of an abelian surface by the Riemann form $(x(1,z), y(1,z)) = \mathrm{tr}(\mu x\overline{y})$

for $x, y \in D \otimes \mathbb{R}$. It is easy to check that acting on z by $\mathcal{O}_D^*$ preserves the lattice L_z up to a right $\mathcal{O}_D^*$-action. Thus we have a well defined map

$$\psi : \Gamma^D \backslash \mathbb{H}^{\pm} \to X^D(\mathbb{C}), \qquad \psi(z) = (A_z, \varphi_D)$$

Likewise, given an abelian surface $A/\mathbb{C}$ with an $\mathcal{O}_D$-action ι, we can pick a uniformization $A(\mathbb{C}) \cong \mathbb{C}^2/L$ in such a way that the induced $\mathcal{O}_D$-action on $\mathbb{C}^2$ is given by φ_D. Then, as the Picard group of $\mathcal{O}_D$ is trivial by [Cla, Theorem 15] we can pick an element $(v, w) \in L$ which generates L over $\mathcal{O}_D$, and this element is unique up to the $\mathcal{O}_D^*$-action. Setting $\xi(A) = z$ gives us a well defined map from $X^D(\mathbb{C})$ to $\Gamma^D \backslash \mathbb{H}^{\pm}$ and it is easy to check that ξ and ψ are inverse to each other. □

Note that $X^D(\mathbb{C})$ can have 2 connected components. In fact, this will be the case precisely when $\mathcal{O}_D^*$ has an element of norm -1, as can easily be deduced from the above.

In uniformizing $X^D(p)(\mathbb{C})$, it turns out to be convenient to use two copies of $\mathbb{H}^{\pm}$ instead of one. Briefly, the reason for this is that the square class of the Weil pairing of any two particular p-torsion elements is an invariant, and thus $X^D(p)(\mathbb{C})$ has twice as many connected components as X^D.

Define

$$\Gamma^D(p) := \ker(\Gamma^D \to (\mathcal{O}_D \otimes \mathbb{F}_p)^* / \pm 1)$$

Fix a non-square element $\alpha \in \mathbb{F}_p^*$ and an element $g_0 \in \mathcal{O}_D \otimes \mathbb{F}_p$ such that $g_0^2 = \alpha$. Note that g_0 exists because $\mathbb{F}_p^2$ embeds into $M_2(\mathbb{F}_p)$ which is isomorphic to $\mathcal{O}_D \otimes \mathbb{F}_p$ for $p \gg 1$. We define two maps, $\psi_1, \psi_2 : \mathbb{H}^{\pm} \to X^D(p)$. Both maps send z to the abelian surface (A_z, φ_D) as in the proof of Lemma 2, but ψ_1 assigns the torsion element $\frac{z}{p}$ whereas ψ_2 assigns the torsion element $\varphi_D(g_0)\left(\frac{z}{p}\right)$. It is easy to see that both these maps are injective and well defined on $\Gamma^D(p)\backslash\mathbb{H}^{\pm}$. To set notation, we label the sources of ψ_i as $\mathbb{H}_i^{\pm}$ for $i = 1, 2$.

Now, note that the monodromy group of $X^D(p)/X^D$ is $G_p := \mathbb{P}(\mathcal{O}_D \otimes \mathbb{F}_p)^*$ acting naturally on the p-torsion element.

Lemma 3. (a) *If $\mathcal{O}_D^*$ has an element of norm -1 and $-1 \notin (\mathbb{F}_p^*)^2$ then*

$$X^D(p)(\mathbb{C}) \cong \Gamma^D(p)\backslash\mathbb{H}^{\pm}, \qquad \mathcal{O}_D^*/\Gamma^D(p) \cong G_p$$

and the action of G_p is induced by the φ_D action of $\mathcal{O}_D^*$ on $\mathbb{H}^{\pm}$.

(b) *Else,*

$$X^D(p)(\mathbb{C}) \cong \Gamma^D(p)\backslash\mathbb{H}_1^{\pm} \cup \Gamma^D(p)\backslash\mathbb{H}_2^{\pm}$$

where the monodromy action is given as follows: $g_0 \in G_p$ acts as $(z, w) \to (w, z)$, while $g \in (\mathcal{O}_D \otimes \mathbb{F}_p)_1^*$ acts as $(z, w) \to (gz, g_0^{-1} g g_0 w)$, as induced by the $\varphi_D \times \varphi_D$ action of $\mathcal{O}_D^*$.

Proof. First, we show that the union of ψ_1 and ψ_2 is surjective. If we have an abelian variety with an $\mathcal{O}_D$-action (A_z, ι_z) with a non-zero p-torsion element v, then $v = hz$ for some $h \in (\mathcal{O}_D \otimes \mathbb{F}_p)^*$. Since $\mathrm{SL}_2(\mathbb{F}_p) \subset (\mathcal{O}_D \otimes \mathbb{F}_p)^*$ we have

$$\mathbb{P}(\mathcal{O}_D \otimes \mathbb{F}_p) = (\mathcal{O}_D \otimes \mathbb{F}_p)^* \cup g_0(\mathcal{O}_D \otimes \mathbb{F}_p)^*$$

where the tilde denotes reduction mod p. The surjectivity follows. Moreover, if $\mathcal{O}_D^*$ has an element of norm -1 and $-1 \notin (\mathbb{F}_p^*)^2$ then we likewise see that ψ_1 is surjective. In this case, the rest of (a) follows easily.

For (b), we must only point out that $g \cdot g_0 z = g_0 \cdot g_0^{-1} g g_0 z$, and that g_0^2 acts as a constant on p-torsion by construction. The proof follows similarly to above. □

2.4. **Heegner and anti-Heegner CM points.**

Suppose $z \in X^D(p)$ is a point with a non-trivial stabilizer in G_p. Then z corresponds to a pair $(A, \iota : \mathcal{O}_D \hookrightarrow E = \mathrm{End}(A))$ that has an automorphism besides ± 1.

Lemma 4. *For z with non-trivial stabilizer in G_p, A_z is isogenous to $E_i \times E_i$ or $E_\omega \times E_\omega$, where we write E_z for the elliptic curve $\mathbb{C}/\langle 1, z\rangle$.*

Proof. By assumption, there is an element $x \in E^*$ which commutes with all of $\mathcal{O}_D$. Now, $E \otimes \mathbb{Q}$ must be a central simple algebra of rank 8, and thus is a matrix algebra over a quadratic field $K = Z(E \otimes \mathbb{Q})$. Let $R = K \cap E$. Since D is rank 4 over $\mathbb{Q}$, D and K generate all of E over $\mathbb{Q}$. Thus, the commutant of D in $E \otimes \mathbb{Q}$ is the center of E, and thus the commutant of $\mathcal{O}_D$ in E is R. Since R is a quadratic ring, it follows that the order of x is either 2 or 3, and $R = \mathbb{Z}[i]$ or $R = \mathbb{Z}[\omega]$. We claim that $E \otimes \mathbb{Q} \cong M_2(K)$. Indeed, if A were simple and had a 4-dimensional CM field acting on it, then that would be its entire endomorphism algebra. Thus A is isogenous to $E_i \times E_i$ or $E_\omega \times E_\omega$, as desired. □

We let A_i and A_ω to be fixed abelian varieties isogenous to $E_i \times E_i$ and $E_\omega \times E_\omega$ respectively, together with an $\mathcal{O}_D$-action. Let R_i, R_ω be fixed subrings of $\mathcal{O}_D$ isomorphic to $\mathbb{Z}[i], \mathbb{Z}[\omega]$ respectively, assuming these exist. Then we have the more refined

Lemma 5. *For z with non-trivial stabilizer in G_p, A_z is $\mathcal{O}_D$-isogenous to A_i or A_ω.*

Proof. We assume A_i is isogenous to A_z, the other case being analogous. We can consider their complex points as $A_i(\mathbb{C}) = \mathbb{C}^2/L_1$ and $A_z(\mathbb{C}) = \mathbb{C}^2/L_2$ respectively, where L_1 and L_2 are commensurable. Let $V = L_1 \otimes \mathbb{Q} = L_2 \otimes \mathbb{Q}$.

Let E_1 and E_2 be the endomorphism rings of A_i and A_z respectively. Then we can identify $E_1 \otimes \mathbb{Q}$ with $E_2 \otimes \mathbb{Q}$ as the complex endomorphisms of $\mathbb{C}^2$ that preserve V. Now, let ι_1, ι_2 be the two embeddings of $\mathcal{O}_D$ into E corresponding to A_i, A_z respectively. Since $E_1 \otimes \mathbb{Q}$ is a central simple algebra, there exists an element e such that $e\iota_1 e^{-1} = \iota_2$. Considering a sufficiently large integer n such that $neL_1 \subset L_2$, we see that the map $z \to nez$ is an $\mathcal{O}_D$ isogeny from A_i to A. This completes the proof. □

Now, suppose that $z, w \in X_D(p)$ have a common stabilizer $g \in G_p$. Then by the discussion above, the centers of their rings of endomorphisms are $R = \mathbb{Z}[i]$ or $R = \mathbb{Z}[\omega]$, and there is a lift $g' \in (\mathcal{O}_D \otimes \mathbb{F}_p)^*$ and $x_z, x_w \in R^*$ such that $g'v_z = x_z v_z, g'v_w = x_w v_w$ for v the chosen p-torsion element. Note that g' must have reduced norm 1, and so is determined up to negation, and in particular we can ensure it has order 4 or 3. Once we pick such a g', it determines x_z, x_w. Then on the complex tangent space at 0 of A_z, x_z acts as one of two scalars: either $\{i, -i\}$ or $\{\omega, -\omega\}$.

Definition 2.5. Generalizing the notation of [KS98], if the eigenvalues of x_z and x_w acting on the tangent spaces at 0 are the same, we say that (z, w) is a

Heegner CM point. Else, the eigenvalues differ by conjugation and we say (z, w) is an anti-Heegner CM point.

We remark here that (z, w) is a Heegner CM point if and only if $(\overline{z}, w)$ is an anti-Heegner CM point.

2.6. **Diagonal quotient varieties.**

In this section we introduce the main objects of study, the diagonal quotient varieties $Z^D(p)$. The stack $\mathcal{Z}^D(p)$ is the stack of pairs of abelian surfaces (A_1, A_2) with $\mathcal{O}_D$-actions, together with a $\mathcal{O}_D$-isomorphisms up to scale $A_1[p] \xrightarrow{\cong} A_2[p]$. $\mathcal{Z}^D(p)$ is the stack quotient $[G\backslash \mathcal{X}^D(p) \times \mathcal{X}^D(p)]$, where $G = \Gamma^D/\Gamma^D(p)$ acts diagonally. We let $Z^D(p)$ be the coarse space associated to $\mathcal{Z}^D(p)$. Note that $Z^D(p)$ is simply the scheme quotient $G\backslash X^D(p) \times X^D(p)$, again with G acting diagonally. The variety $Z^D(p)$ is therefore a proper projective normal scheme defined over $\mathbb{Q}$ with cyclic quotient singularities of order 2 or 3. The diagonal quotient surface obtained as above from elliptic modular curves has been studied in detail by [Her91] and [KS98].

Remark 6. Recall that our definition of a full level structure is only up to scale, and thus the $\mathcal{O}_D$-isomorphisms parameterized by $\mathcal{Z}^D(p)$ are up to scale. For the problem we are considering, it would be more natural for the notion of isomorphism of level structures to coincide with that of $\mathcal{O}_D$-modules, but we prefer our approach as it avoids keeping track of the Weil pairing and notationally distinguishing different connected components of the Shimura curves. We note that there is a forgetful map from the unscaled level structure to the one we are considering, so in actuality we are proving a slightly stronger theorem.

The variety $Z^D(p)$ comes equipped with Hecke curves H_M for any M coprime to p. H_M is the image of the Hecke correspondence $T_M \to X^D(p) \times X^D(p)$ from Section 2.2 under the quotient map and parametrizes pairs of abelian surfaces (A_1, A_2) with an isomorphism up to scale $A_1[p] \xrightarrow{\cong} A_2[p]$ induced by a cyclic isogeny $A_1 \to A_2$ of degree M.

3. Geometry of the Heegner CM points

3.1. **Preliminaries.**

Throughout this section, the complex points of our modular curves X will be equipped with canonical metrics of constant sectional curvature -1 inherited from $\mathbb{H}^{\pm}$. Henceforth we will typically blur the notational distinction between X and $X(\mathbb{C})$. For $x \in X$ we let $B(x, r)$ be the set of all points in X within distance r of x.

Recall that the injectivity radius $\rho_X(x)$ of X at a point $x \in X$ is half the length of the shortest closed geodesic through x. Equivalently, it is the radius r of the largest isometrically embedded hyperbolic ball $B(x, r) \subset X$. The injectivity radius ρ_X is then the infimum of $\rho_X(x)$ over all $x \in X$, or equivalently half the length of the shortest closed geodesic in X. It was first observed by Buser-Sarnak in [BS94] that the injectivity radii of Shimura curves are large. For the convenience of the reader, we recall the proof:

Lemma 7. $\rho_{X^D(p)} \geq 2 \log p + O(1)$.

Proof. A closed geodesic through $x \in X^D(p)$ lifts to the unique geodesic arc between two lifts $z, \gamma z \in \mathbb{H}^{\pm}$ for some non-identity $\gamma \in \Gamma^D(p)$. Note that because $\Gamma^D(p)$ is cocompact, every element of $\Gamma^D(p)$ is semisimple, so $d(z, \gamma z) = d(Az, (\operatorname{tr}\gamma)Az)$, where $A \in \mathrm{SL}_2\,\mathbb{R}$ is the diagonalizing matrix. In particular, using the formula for distance in the upper half-plane, this means

$$\begin{aligned}\min_z d(z, \gamma z) &= \min_z d(z, az) \\ &= \min_z \operatorname{arcosh}\left(1 + \frac{(a-1)^2|z|^2}{2a(\Im z)^2}\right) \\ &\geq \operatorname{arcosh}\left(\operatorname{tr}(\gamma^2)/2\right)\end{aligned}$$

Again because Γ^D is cocompact, the only element with trace 2 is the identity, and thus the minimal value of $|\operatorname{tr}\gamma|$ for $1 \neq \gamma \in \Gamma^D(p)$ is $2 + p^2$ as $\operatorname{tr}(\gamma) \cong 1 \bmod p^2$. Thus, $\operatorname{tr}(\gamma^2)/2 = \operatorname{tr}(\gamma)^2/2 - 1 \geq p^4/2$, and the result follows. □

3.2. **Repulsion of Heegner CM points.**

The technical heart of this section is the following

Proposition 8. *For each $r > 0$, there exists $d > 0$ such that for all sufficiently large p, if $(x, y), (x', y')$ are distinct CM Heegner points in $X^D(p) \times X^D(p)$ with the same projections to $X^D \times X^D$ and $B(x, r) \cap B(x', r) \neq \varnothing$ and $B(y, r) \cap B(y', r) \neq \varnothing$, then one of $(x, y), (x', y')$ lies on some Hecke divisor T_k with $k < d$.*

Proof. We first reduce to the case where x, y project to the same point in X^D (and hence so do x', y'). By Lemma 5 we can find an isogeny of bounded degree B between the images of x, y in X^D. Since the Hecke correspondences are locally geodesic and the pullback of T_M along T_B is a union of T_k for $k < MB$, by pulling back along T_B in the first co-ordinate we can assume that x and y project to the same point in X^D, at the cost of scaling d by some bounded amount B.

Since (x, y), (x', y') lie in the same component of $X^D(p)$, there must be $i, j \in \{1, 2\}$ and lifts $(z, w)(z', w') \in \mathbb{H}_i^{\pm} \times \mathbb{H}_j^{\pm}$ such that (z, w) is in the $\mathcal{O}_{D,1}^* \times \mathcal{O}_{D,1}^*$ orbit of (z', w'). Replacing (x, y) by $(g_0 x, g_0 y)$ if necessary, we can assume that $i = 1$. Moreover, we can pick these lifts so that

$$d((z, w), (z', w')) = d((x, y), (x', y')) < 4r$$

and by acting diagonally by G_p we can pick z from a finite fixed set of points S independently of p. Set $w = gz$, $g \in G_p$.

Let $t \in \mathcal{O}_{D,1}^*$ be a stabilizer of z. Since w must also be stabilized by t, gtg^{-1} is either t or t^{-1} in G_p. This is true in $(\mathcal{O}_D \otimes \mathbb{F}_p)^*$ up to a sign *a priori*, but by comparing the traces gtg^{-1} must in fact be t or t^{-1} in $\mathcal{O}_D \otimes \mathbb{F}_p)^*$, for any lift of g. Since (z, w) is a Heegner CM point by assumption, it follows that $gt = tg$.

Next, fix the set $S_r \subset \mathcal{O}_{D,1}^*$ of elements γ such that $d(\gamma z, z) < r$. Then there must exist elements $h_z, h_w \in S_r$ such that $h_z z = z', w' = gh_w z = gh_w h_z^{-1} z'$. Hence, as in the above we must have that $gh_w h_z^{-1}$ commutes with $h_z t h_z^{-1}$, or equivalently $h_z g h_w$ commutes with t.

The two relations $gt = tg$ and $h_z g h_w t = t h_z g h_w$ can be interpreted as a set of linear equations in the coefficients of g, where we view g as a 2×2 matrix. The relation $gt = tg$ defines the field generated by g, so it has a 2-dimensional set of solutions. Thus, either the 2 relations define a line, or the second relation

is redundant. Note that the relation is redundant over $\mathbb{Q}$ if and only if it is redundant over all sufficiently large finite fields.

If the second relation is redundant over $\mathbb{Q}$, setting H to be the centralizer of t in D^*, we must have $h_z H h_w = H$, which implies

$$h_z H h_z^{-1} = (h_z H h_w)(h_z H h_w)^{-1} = H H^{-1} = H.$$

Likewise, $h_w H h_w^{-1} = H$. Now, we claim that the elements of the normalizer of H in D^* which have positive norm consist exactly of H. To prove this, note that its enough to check it after tensoring with $\mathbb{R}$, in which case D become $M_2(\mathbb{R})$ and H becomes an embedded $\mathbb{C}^*$ (unique up to conjugation). Thus, $h_z, h_w \in H$. Finally, note that since $H \cap \mathcal{O}_D$ is isomorphic to either $\mathbb{Z}[i]$ or $\mathbb{Z}[\omega]$ we must have h_z, h_w stabilizers of z, contradicting the assumption that our Heegner CM points were distinct.

Thus, the two relations must not be redundant, and we end up with a single projective solution g which we can take to be in $\mathcal{O}_D$. Thus z and w lie on T_M where $M = \mathrm{N}(g)$. Setting d to be bigger then all (finitely many) M arising in this way gives the result. □

4. CM incidence estimates

For any hyperbolic curve X and any curve $C \subset X \times X$, work of Hwang and To shows that the multiplicity of C at a point $x \in X \times X$ is bounded in terms of the volume of C in a geodesic ball centered at x, and similarly its intersection with a totally geodesic curve $H \subset X \times X$ is bounded by the volume of C in a geodesic tubular neighborhood of H. Note that the volume of a curve can be interpreted as the degree of the restriction of the canonical divisor $K_{X\times X}$ to C, and therefore as the intersection number $C \cdot K_{X\times X}$. This allows one to deduce algebro-geometric results from the hyperbolic "spread-outedness" of X.

Theorem 9. *For any curve $C \subset X \times X$, we have*

(a) [HT02, Theorem 1] *For any point $x \in X$, let $B_r = B(x, r)$ for $r < \rho_X(x)$. Then*

$$\mathrm{vol}(C \cap B) \geq 4\pi \sinh^2\left(\frac{r}{2}\right) \mathrm{mult}_x(C)$$

(b) [HT12, Theorem 1] *Let $\Delta \subset X \times X$ and $W_r = \{(z, w) \in X \times X | d(z, w) < r\}$ for $r < \rho_X$. Then*

$$\mathrm{vol}(C \cap W_r) \geq 8\pi \sinh^2\left(\frac{r}{4}\right) (C \cdot \Delta)$$

Recall from Section 2.2 that for any m there are two measure preserving maps $\mu, \nu : T_m \to X^D(p)$ in the sense that push-forward of multisets preserves volume and pull-back of multi-sets multiplies volume by the degree. The maps $X^D(p) \times T_m \to X^D(p) \times X^D(p)$ given by $\varphi = \mathrm{id} \times \mu$ and $\psi = \mathrm{id} \times \nu$ then have the same property. Letting $W_r^m = \psi_* \varphi^* W_r$, we can generalize part (b) of Theorem 9 to Hecke curves:

Corollary 10. *For any Hecke curve $T_m \subset X^D(p) \times X^D(p)$, and $r < \rho_{X^D(p)}$*

$$\mathrm{vol}(C \cap W_r^m) \geq 8\pi \sinh^2\left(\frac{r}{4}\right) (C \cdot T_m)$$

Proof.

$$\begin{aligned}\mathrm{vol}(C \cap W_r^m) &= \mathrm{vol}(\varphi_* \psi^* C \cap W_r) \\ &\geq \sinh^2\left(\frac{r}{4}\right)(\varphi_* \psi^* C \cdot \Delta) \\ &= \sinh^2\left(\frac{r}{4}\right)(C \cdot \psi_* \varphi^* \Delta) \\ &= \sinh^2\left(\frac{r}{4}\right)(C \cdot T_m)\end{aligned}$$

□

Both statements in Theorem 9 are optimal in the sense that the bound is realized by a union of translates of the image of the graph of $-z : \mathbb{D} \to \mathbb{D}$. For the convenience of the reader, we summarize the proof of part (b) above. Recall the following

Definition. For $\varphi(z)$ a plurisubharmonic function on a neighborhood of a point x in some complex manifold M, the Lelong number of φ at x is

$$\nu(\varphi, x) := \liminf_{z \to x} \frac{\varphi(z)}{\log|z - x|}.$$

For example, if $V \subset M$ is a divisor cut out locally by f, and $\varphi(z) = \log|f(z)|$, then $\nu(\varphi, x) = \mathrm{mult}_x(f)$.

Hwang and To define a plurisubharmonic function $F : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ that is diagonally-invariant under the full isometry group $\mathrm{PSL}_2 \mathbb{R}$ such that $0 \leq \omega_F = i\partial\overline{\partial} F \leq \omega_{std}$, as well as diagonally-invariant functions $f_\epsilon : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ that

(1) are plurisubharmonic off the diagonal $\Delta_{\mathbb{D}} \subset \mathbb{D} \times \mathbb{D}$;
(2) agree with F outside of $B(\Delta_{\mathbb{D}}, r)$;
(3) have a logarithmic pole along the diagonal, and for any $\xi \in \Delta_{\mathbb{D}}$,

$$\liminf_{\epsilon \to 0} \nu(f_\epsilon, \xi) = 8 \sinh^2(r/4)$$

As f_ϵ and F descend to functions on $X \times X$, it then follows that for any curve $C \subset X \times X$,

$$\mathrm{vol}(C \cap B(\Delta, r)) \geq \int_{C \cap B(\Delta, r)} \omega_F = \int_{C \cap B(\Delta, r)} \omega_{f_\epsilon} \geq \pi \sum_{\xi \in C \cap \Delta} \nu(f_\epsilon, \xi) \, \mathrm{mult}_\xi C$$

where we've used the diagonal invariance to descend the forms to $X \times X$. The equality follows from Stoke's theorem and the second inequality from the fact that, for $[C]$ denoting the current of integration along C,

$$\nu([C] \wedge \omega_{f_\epsilon}, \xi) \geq \nu([C], \xi) \nu(f_\epsilon, \xi)$$

(*cf* [HT02], Proposition 2.2.1(a)).

We will need a result comparing the volume within a radius R to that within a smaller radius r of the *conjugate* diagonal in order to handle the anti-Heegner CM points. For a curve X, its conjugate $\overline{X}$ is the same curve with the negated complex structure, and the pointwise diagonal $\overline{\Delta} \subset X \times \overline{X}$ is called the conjugate diagonal.

Proposition 11. *For X a compact hyperbolic complex curve, any complex curve $C \subset X \times \overline{X}$ that is not the conjugate diagonal, and any $\rho_X > R > r > 0$,*

$$\mathrm{vol}(C \cap B(\overline{\Delta}, R)) \geq \frac{\sinh(R/2)}{\sinh(r/2)} \mathrm{vol}(C \cap B(\overline{\Delta}, r))$$

Furthermore, the bound is optimal in the following sense: suppose X is isomorphic to $\overline{X}$ via a map $z \to \overline{z}$. (For instance, X is defined over $\mathbb{R}$). Then the graph $(z, \overline{z})$ achieves the bound.

Proof. The proof is very similar to Proposition **??**. Suppose $X = \Gamma\backslash\mathbb{D}$, so that $\overline{X} = \overline{\Gamma}\backslash\mathbb{D}$, and consider the function ψ on $\mathbb{D} \times \mathbb{D}$ given by

$$\psi(z, w) = \tanh^2(d_{\mathbb{D}}(z, \overline{w})/2) = \left| \frac{\overline{w} - z}{1 - zw} \right|^2.$$

ψ is invariant under the diagonal action of $\mathrm{SL}_2 \mathbb{R}$. For any function $f : \mathbb{R} \to \mathbb{R}$, we compute that at $(0, w)$, the potential $F(z, w) = f(\psi)$ yields a form

$$\begin{aligned} \omega_F &= i\partial\overline{\partial} F \\ &= f'(\psi) \begin{pmatrix} (1 - |w|^2)^2 & w^2 \\ \overline{w}^2 & 1 \end{pmatrix} + f''(\psi) \begin{pmatrix} |w|^2(1 - |w|^2)^2 & -w^2(1 - |w|^2) \\ -\overline{w}^2(1 - |w|^2) & |w|^2 \end{pmatrix} \end{aligned}$$

Taking $s(\psi) = -\log(1 - \psi)$ and $S(z, w) = s(\psi(z, w))$, for instance, we have by direct computation

$$\omega_S = \begin{pmatrix} 1 & 0 \\ 0 & (1 - |w|^2)^{-2} \end{pmatrix} = \frac{\omega_{std}}{2}$$

where ω_{std} is the standard form on $\mathbb{D} \times \mathbb{D}$. Let $C = s(\tanh^2(R/2))$, $c = s(\tanh^2(r/2))$, and define a continuous function $f : [0, 1] \to \mathbb{R}$ on the interval $\left[\tanh^2(r/2), \tanh^2(R/2)\right]$ by $f(\psi) = h(s(\psi))$, where

$$h'(s) = \frac{1 - \sqrt{\frac{e^c - 1}{e^s - 1}}}{1 - \sqrt{\frac{e^c - 1}{e^C - 1}}}$$

Take f to be constant on $[0, \tanh^2(r/2)]$, and linear of slope 1 on $[\tanh^2(R/2), 1]$. One can easily compute that the resulting ω_F is positive, and dominated by

$$(h'(s) + 2(1 - e^{-s})h''(s))\frac{\omega_{std}}{2} = \left(1 - \sqrt{\frac{e^c - 1}{e^C - 1}}\right)^{-1} \frac{\omega_{std}}{2}$$

on (the interior of) $B(\overline{\Delta}_{\mathbb{D}}, R) - B(\overline{\Delta}_{\mathbb{D}}, r)$. Further we have that

$$\omega_F|_{B(\overline{\Delta}_{\mathbb{D}}, r)} = 0 \quad \text{and} \quad \omega_F|_{B(\overline{\Delta}_{\mathbb{D}}, \rho_X) - B(\overline{\Delta}_{\mathbb{D}}, R)} = \frac{\omega_{std}}{2}$$

Smoothing F out by the same trick as in the proof of Proposition **??** and descending these forms down to $X \times \overline{X}$, we have that

$$\begin{aligned}\mathrm{vol}(C \cap B(\overline{\Delta}, R)) &= \int_{C \cap B(\overline{\Delta},R)} \omega_{std} \\ &= 2 \int_{C \cap B(\overline{\Delta},R)} \omega_F \\ &\leq \left(1 - \sqrt{\frac{e^c - 1}{e^C - 1}}\right)^{-1} \left(\mathrm{vol}(C \cap B(\overline{\Delta}, R)) - \mathrm{vol}(C \cap B(\overline{\Delta}, r))\right)\end{aligned}$$

yielding the statement, as $e^c - 1 = \sinh^2(r/2)$, and likewise for C and R. □

Let CM^+ be the set of Heegner CM points on $X^D(p) \times X^D(p)$, and CM^- the set of anti-Heegner CM points. To get a upper bound for the genus of C we will need an estimate for the total multiplicities $\mathrm{mult}_{\mathrm{CM}^\pm}(C) = \sum_{x \in \{\mathrm{CM}^\pm\}} \mathrm{mult}_x(C)$.

Proposition 12. *For any non-Hecke curve $C \subset X^D(p) \times X^D(p)$, we have*

$$\mathrm{mult}_{\mathrm{CM}^+}(C) = o(C \cdot K_{X \times X})$$

as $p \to \infty$.

Proof. Fix some $R > 0$. For $d > 0$, partition CM^+ into two sets

$$T := \mathrm{CM}^+ \cap \cup_{m<d} T_m$$

and $S := \mathrm{CM}^+ \backslash T$. By Proposition 8, if d is large enough in relation to R, the balls $B(z, R)$ are disjoint as z varies over S. By Theorem 9, Lemma 7 and Corollary 10 we then have that

$$\begin{aligned}\mathrm{mult}_{\mathrm{CM}^+}(C) &= \sum_{x \in S} \mathrm{mult}_x(C) + \sum_{x \in T} \mathrm{mult}_x(T) \\ &\ll \sinh^{-2}(R/2)\, \mathrm{vol}\left(C \cap \cup_{x \in S} B(x, r)\right) + \sum_{m<d} (C.T_d) \\ &\ll \sinh^{-2}(R/2)\, \mathrm{vol}(C) + \sinh^{-2}(p/2) \sum_{m<d} \deg T_m\, \mathrm{vol}(C) \\ &\ll (C.K_{X \times X})(\sinh^{-2}(R) + d^3 \sinh^{-2}(p/2))\end{aligned}$$

As $p \to \infty$, we see that $\mathrm{mult}_{\mathrm{CM}^+}(C) \ll (C \cdot K_{X \times X})(\sinh^{-2}(R) + o(1))$. Since R can be chosen arbitrarily large, the claim follows.

□

Proposition 13. *For any non-Hecke curve $C \subset X^D(p) \times X^D(p)$, we have*

$$\mathrm{mult}_{\mathrm{CM}^-}(C) = o(C \cdot K_{X \times X})$$

as $p \to \infty$.

Proof. Fix some $R > 0$. We would like to perform the same trick for the anti-Heegner CM points, and for appropriately chosen d we again partition the points of CM^- into

$$T = \mathrm{CM}^- \cap \cup_{m<d} \overline{T}_m \quad \text{and} \quad S = \mathrm{CM}^- - T$$

where $\bar{\cdot}$ denotes complex conjugation on the second factor. It will still be the case that balls of radius R around points of S (with d chosen sufficiently large) are disjoint, but the multiplicity of a curve C along $\overline{T}_m$ does not quite make sense, so we adjust the argument slightly:

$$\begin{aligned}\mathrm{mult}_{\mathrm{CM}^-}(C) &= \sum_{x\in S}\mathrm{mult}_x(C)+\sum_{x\in T}\mathrm{mult}_x(C)\\ &\ll \sinh^{-2}(R/2)\,\mathrm{vol}\left(C\cap\cup_{x\in S}B(x,r)\right)+\sum_{m<d}\mathrm{mult}_{\mathrm{CM}^-\cap\overline{T}_m}(C)\\ &\ll \sinh^{-2}(R/2)\,\mathrm{vol}(C)+\sum_{m<d}\deg T_m\cdot\mathrm{mult}_{\mathrm{CM}^-\cap\overline{\Delta}}(T_m^*C)\end{aligned}\tag{1}$$

Since R can be taken arbitrarily large, it suffices to show that for a *fixed* m,

$$\mathrm{mult}_{\mathrm{CM}^-\cap\overline{\Delta}}(T_m^*C)=o(\mathrm{vol}(C))$$

Because the injectivity radius is $2\log p$, there are $O_R(1)$ many overlaps of balls of radius R centered at anti-Heegner CM points on the conjugate diagonal, so by Theorem 9 we have

$$\begin{aligned}\mathrm{mult}_{\mathrm{CM}^-\cap\overline{\Delta}}(T_m^*C) &\ll \sinh^{-2}(R/2)\sum_{x\in\mathrm{CM}^-\cap\overline{\Delta}}\mathrm{vol}(T_m^*\cap B(x,R))\\ &\ll O_R(1)\cdot\sinh^{-2}(R/2)\,\mathrm{vol}(T_m^*C\cap B(\Delta,R))\\ &\ll O_R(1)\cdot\sinh^{-1}(\log p)\,\mathrm{vol}(T_m^*C)\end{aligned}$$

where we've used Proposition 11 (and Lemma 7) in the last step. Since $\mathrm{vol}(T_m^*C)=\deg T_m\,\mathrm{vol}(C)$, the Proposition is proven. □

Therefore, writing

$$\mathrm{mult}_{\mathrm{CM}}C=\mathrm{mult}_{\mathrm{CM}^+}C+\mathrm{mult}_{\mathrm{CM}^-}C$$

we have

Corollary 14. *For any non-Hecke curve $C\subset X^D(p)\times X^D(p)$, we have*

$$\mathrm{mult}_{\mathrm{CM}}(C)=o(C\cdot K_{X\times X})$$

as $p\to\infty$.

5. Low degree curves

Let $F=X^D\times\mathrm{pt}+\mathrm{pt}\times X^D$ be the sum of the fiber divisors on $X^D\times X^D$, and similarly $F_p=X^D(p)\times\mathrm{pt}+\mathrm{pt}\times X^D(p)$. Likewise, let F_Z be the pullback of F to $Z^D(p)$ via the quotient map $q:Z^D(p)\to X^D\times X^D$. Note that for any map from a curve $B\to Z^D(p)$, $B.F_Z$ is simply the sum of the degrees of the two maps $B\to X^D$.

Proposition 15. *For any $k>0$, there is an $N>0$ such that any map from a smooth curve $B\to Z^D(p)$ that does not factor through a Hecke curve has $B\cdot F_Z>k$ as long as $p>N$.*

Remark 16. For the following we work in the category of orbifold curves. X^D is naturally an orbifold curve whose orbifold points are the abelian surfaces with extra automorphisms. $X^D(p)$ likewise is naturally an orbifold curve, but for $p \gg 0$ its orbifold structure is trivial. Note that in this language the map $\pi : X^D(p) \to X^D$ is étale.

Proof. We first observe that for a fixed degree d, the image $f_*\pi_1(B)$ of the fundamental group under a map from a smooth orbifold curve $f : B \to X^D$ of degree d only depends on the ramification profile and the monodromy around the branch points. If we further assume the orbifold points of B lie over those of X^D, then there are finitely many choices for this data, and therefore only finitely many possible maps $f_* : \pi_1(B) \to \pi_1(X^D)$, up to conjugacy. It follows that for maps from orbifold curves $B \to X^D \times X^D$ of bounded bidegree $B.F$ for which the orbifold points of B map to those of $X^D \times X^D$, there are only finitely many possible images of the fundamental group, again up to conjugacy.

Now, given a $k > 0$ as in the Proposition, take $B \to Z^D(p)$ a smooth curve with $B.F_Z < k$, and let $B' \subset X^D \times X^D$ be its projection, with map $\alpha : B \to B'$. Note that B' is naturally an orbifold curve. Consider the image $\Pi \subset \pi_1(X^D \times X^D) = \Gamma^D \times \Gamma^D$ of the fundamental group of B' in that of $X^D \times X^D$. Note that $\Gamma^D = G(\mathbb{Z})$, where G is the algebraic group defined so that for a ring R, $G(R) = (\mathcal{O}_D \otimes R)^*$. For a connected component $X^D(p)_0$ of $X^D(p)$ we have

$$\pi_1(X^D(p)_0) = \Gamma^D(p) = \ker(G(\mathbb{Z}) \to G(\mathbb{F}_p))$$

We can view the map $B' \to X^D \times X^D$ as a variation of Hodge structures, in which case a theorem of Andre-Deligne [And92, Theorem 1] implies that Π is Zariski dense in the Mumford-Tate group $G \times G$ unless B factors through a Hecke curve. By a theorem of Nori [Nor87, Theorem 5.1], every Zariski dense subgroup of G surjects onto $G(\mathbb{F}_p)$ for $p \gg 0$. Since by the above argument the number of these subgroups is finite up to conjugacy, N can be chosen large enough so that we may assume Π surjects onto $G(\mathbb{F}_p) \times G(\mathbb{F}_p)$.

To finish, α has degree $|G(\mathbb{F}_p)|$ since the inverse image of B' in $Z^D(p)$ is irreducible. The composition $B \to Z^D(p) \to X^D$ on the one hand factors through B' but on the other hand has bounded degree, and we have a contradiction for p large enough.

□

Remark 17. In the above proof, one can avoid discussing orbifold points by manually removing the finitely many points in X^D corresponding to abelian surfaces with extra automorphisms.

Note that this immediately allows us to conclude

Corollary 18. *Suppose X^D has genus $g(X^D) > 1$. Then for any $k > 0$, there is an $N > 0$ such that any map from a smooth curve $B \to Z^D(p)$ of genus $g(B) < k$ must factor through a Hecke curve, provided $p > N$.*

Proof. One of the projections $B \to X^D$ has degree at least $\frac{1}{2}(B \cdot F_Z)$. Now use Riemann–Hurwitz and the Proposition. □

6. Proof of Main Theorem

We now prove Theorem 1. For a non-Hecke curve $B \to Z^D(p)$, let C be the normalization of a component of the preimage of B. The key point is that Proposition 15 bounds the genus of C from below, while Corollary 14 bounds the genus from above in an asymptotically smaller way.

Theorem 19. *For any $k > 0$, there exists an $N > 0$ such that any smooth curve $B \to Z^D(p)$ of genus $g(B) < k$ must factor through a Hecke curve, provided $p > N$.*

Proof. Suppose B does not factor through a Hecke divisor. Let $C \to X^D(p) \times X^D(p)$ be the normalization of a connected component of the preimage of B in $X^D(p) \times X^D(p)$, and let $\alpha : C \to B$ be the map to B. Fix a connected component $X^D(p)_0$ of $X^D(p)$, and identify all other components with $X^D(p)_0$. C then lands in a connected component identified with $X^D(p)_0 \times X^D(p)_0$. Note that

$$K_{X^D(p)_0 \times X^D(p)_0} = (2g(X^D(p)_0) - 2)F_{X^D(p)_0 \times X^D(p)_0}$$

where $F_{X^D(p)_0 \times X^D(p)_0}$ is the sum of the fibers, *i.e.* $X^D(p)_0 \times \mathrm{pt}$ and its flip. If $\pi : C \to X^D(p)_0$ is the projection onto a factor with largest degree, then on the one hand Riemann-Hurwitz applied to π yields

$$g(C) \geq \frac{1}{2}(C \cdot F_{X^D(p)_0 \times X^D(p)_0})(g(X^D(p)_0) - 1)) = \frac{1}{4}(C \cdot K_{X^D(p)_0 \times X^D(p)_0}) \tag{2}$$

On the other hand, by Corollary 14, as $p \to \infty$,

$$\mathrm{mult}_{\mathrm{CM}}(C) = o(C \cdot K_{X^D(p)_0 \times X^D(p)_0})$$

Thus, applying Riemann-Hurwitz to α,

$$\begin{aligned} g(C) &\leq 1 + (\deg \alpha)(g(B) - 1) + \frac{1}{2}\mathrm{mult}_{\mathrm{CM}}(C) \\ &\leq 1 + (\deg \alpha)(g(B) - 1) + o(C \cdot K_{X^D(p)_0 \times X^D(p)_0}) \end{aligned} \tag{3}$$

By Proposition 15 we know $(C \cdot K_{X^D(p)_0^2})/\deg \alpha = B \cdot F_Z \to \infty$ as $p \to \infty$, so after dividing equations (2) and (3) by $\deg \alpha$, for p large enough we obtain the result. □

Remark 20. As mentioned in the introduction, the techniques of [BT14] can be used to show that the constant N in the theorem can be taken to only depend on the gonality of B. Briefly, if B is d-gonal, then we obtain a map $\mathbb{P}^1 \to \mathrm{Sym}^d Z(p)$. By proving repulsion results akin to those of Section 3 for the diagonals in $(X^D(p) \times X^D(p))^d$ (see [BT14, Proposition 18]), we obtain multiplicity estimates for the pull back C of B to $(X^D(p) \times X^D(p))^d$ along those diagonals (see [BT14, Proposition 30]). The map $C \to B$ ramifies only when C passes through diagonals or CM points, and by an argument similar to the proof above (see also [BT14, Proposition 31]) it follows that $\mathrm{Sym}^d Z(p)$ has no rational curves for large enough p.

Remark 21. For each fixed (large enough) p, it is easy to deduce the same result over $\overline{\mathbb{F}}_\ell$ for sufficiently large ℓ by a standard argument.

References

[And92] Y. André. Mumford-Tate groups of mixed Hodge structures and the theorem of the fixed part. *Compositio Math.*, 82(1):1–24, 1992.

[Bil] N. Billerey. *Private communication.*

[BS94] P. Buser and P. Sarnak. On the period matrix of a Riemann surface of large genus. *Invent. Math.*, 117(1):27–56, 1994. With an appendix by J. H. Conway and N. J. A. Sloane.

[BT14] B. Bakker and J. Tsimerman. p-torsion monodromy representations of elliptic curves over geometric function fields. arXiv:1403.7168, 2014.

[Cla] P. L. Clark. *Lectures on Shimura curves 9.* http://math.uga.edu/~pete/SC9-Orders.pdf.

[Cla03] P. L. Clark. *Rational points on Atkin-Lehner quotients of Shimura curves.* PhD thesis, Harvard University Cambridge, Massachusetts, 2003.

[Elk98] N. D. Elkies. Shimura curve computations. In *Algorithmic Number Theory*, pages 1–47. Springer, 1998.

[Her91] C. F. Hermann. Modulflächen quadratischer Diskriminante. *Manuscripta Math.*, 72(1):95–110, 1991.

[HT02] J. Hwang and W. To. Volumes of complex analytic subvarieties of Hermitian symmetric spaces. *American Journal of Mathematics*, 124(6):1221–1246, 2002.

[HT12] J. Hwang and W. To. Injectivity radius and gonality of a compact Riemann surface. *American Journal of Mathematics*, 134(1):259–283, 2012.

[KS98] E. Kani and W. Schanz. Modular diagonal quotient surfaces. *Mathematische Zeitschrift*, 227(2):337–366, 1998.

[MG78] B. Mazur and Appendix by D. Goldfeld. Rational isogenies of prime degree. *Inventiones mathematicae*, 44(2):129–162, 1978.

[Mil97] J. S. Milne. *Class Field Theory.* http://www.math.lsa.umich.edu/jmilne, 1997.

[Nor87] M. V. Nori. On subgroups of $\mathrm{GL}_n(\mathbf{F}_p)$. *Invent. Math.*, 88(2):257–275, 1987.

[Voi09] John Voight. Shimura curves of genus at most two. *Math. Comp.*, 78(266):1155–1172, 2009.

B. Bakker: Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012

E-mail address: bakker@cims.nyu.edu

J. Tsimerman: Mathematics Department, Harvard University, 1 Oxford Street, Cambridge, MA,02138

E-mail address: jacobt@math.harvard.edu