Counting points[*]

Avi Zeff

Our next goal is to understand reductions of Shimura varieties modulo p, and in particular to count points of Shimura varieties over finite fields. To do so we first review the theory of abelian varieties over finite fields, and then specialize to the case of Shimura varieties and state the Langlands-Rapoport conjecture; finally we apply this to get a point-counting formula.

1. Abelian varieties over finite fields

Our goal is to understand the category of abelian varieties (up to isogeny) over finite fields. This is a semisimple $\mathbb{Q}$-linear category; it turns out that any semisimple F-linear category C can be described by its set of simple objects, up to isomorphism, and their endomorphism algebras.

In more detail: let C be a semisimple F-linear category, i.e. an abelian category where each hom-set is a finite-dimensional F-vector space, composition is F-bilinear, and every object is a (finite) direct sum of simple objects, i.e. nonzero objects with no nonzero proper subobjects. If e is a simple object and $e \to e$ is a morphism, it must be an isomorphism since its kernel and image are subobjects of e. Therefore $\mathrm{End}(e)$ is a division algebra over F. If re is the direct sum of r copies of e, then $\mathrm{End}(re) \simeq M_r(\mathrm{End}(E))$. For another simple object e', by the same argument any nonzero map $e \to e'$ must be an isomorphism, i.e. $\mathrm{Hom}(e, e')$ is either 0 or $\mathrm{End}(e) \simeq E(e')$, depending whether e and e' are isomorphic. Therefore if $e_1, \ldots, e_n$ are distinct simple objects and $x = r_1e_1 + \cdots + r_ne_n$, $y = s_1e_1 + \cdots + s_ne_n$, then

$$\mathrm{Hom}(x, y) = \prod M_{s_i,r_i}(\mathrm{End}(e_i)).$$

In particular every object is the sum of simple objects and every hom set is the product of matrix algebras over endomorphism algebras of simple objects, which are division algebras. Therefore if we can understand the sets $\Sigma(C)$ of isomorphism classes of simple objects and $D(C)$ of endomorphism algebras of representative simple objects, then we understand C. These are called the numerical invariants of C; our first goal is to compute these invariants for the category of abelian varieties up to isogeny over a finite field.

First, we want to understand the simple objects, i.e. simple abelian varieties over $\mathbb{F}_q$. Recall that we have the Frobenius invariant $A \mapsto \pi_A \in \mathrm{End}^0(A)$, which is a Weil q-integer, i.e. an algebraic integer whose image under any embedding $\mathbb{Q}[\pi_A] \hookrightarrow \mathbb{C}$ has absolute value $q^{1/2}$. We have a natural notion of conjugacy by varying embeddings: two Weil q-integers π, π' are conjugate if there is an isomorphism $\mathbb{Q}[\pi] \to \mathbb{Q}[\pi']$ sending π to π'. It turns out that this is in a certain sense a complete invariant: by a theorem of Tate, $A \mapsto \pi_A$ is an injective map from the set of isomorphism classes of simple abelian varieties over $\mathbb{F}_q$ to the set of conjugacy classes of Weil q-integers. Further work of Honda shows that in fact this is a bijection. Writing $W_1(q)$ for the set of Weil q-integers in $\overline{\mathbb{Q}}$ and Γ for $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, this is a bijection $\Sigma(\mathsf{AV}^0(\mathbb{F}_q)) \to \Gamma\backslash W_1(q)$.

[*] These notes are based on chapters 15-17 of [1].

Now we understand the set of isomorphism classes of simple objects; what about their endomorphism algebras? Each $\mathrm{End}^0(A)$ is a division algebra over $\mathbb{Q}$, and in fact over its center F, which is generated by the Frobenius $F = \mathbb{Q}[\pi_A]$. These are classified by the short exact sequence

$$0 \to \mathrm{Br}(F) \to \bigoplus_v \mathrm{Br}(F_v) \xrightarrow{\sum_v \mathrm{inv}_v} \mathbb{Q}/\mathbb{Z} \to 0$$

by class field theory, where $\mathrm{Br}(F)$ is the Brauer group and v ranges over places of F. Thus it suffices to specify $\mathrm{inv}_v(\mathrm{End}^0(A))$ at every place v.

Since A is determined by π_A, we might hope to be able to give $\mathrm{End}^0(A)$ in terms of π_A. First, we can note that π_A is in the center of the endomorphism ring, and it turns out that it generates the center. To say more, as above this boils down to determining each $\mathrm{inv}_v(\mathrm{End}^0(A))$ in terms of π_A; and in fact this is possible: by the same work of Tate, it turns out that $\mathrm{inv}_v(\mathrm{End}^0(A))$ is equal to $\frac{1}{2}$ if v is real, $\frac{\mathrm{ord}_v(\pi_A)}{\mathrm{ord}_v(q)}[F_v : \mathbb{Q}_p]$ for $v|p$, and 0 otherwise, and $[D : F] = \left(\frac{2\dim A}{[F:\mathbb{Q}]}\right)^2$.

Now that we understand $\mathsf{AV}^0(\mathbb{F}_q)$ for every finite field $\mathbb{F}_q$, we'd like to be able to give a similar description for $\mathsf{AV}^0(\overline{\mathbb{F}_q})$.

Every abelian variety over $\overline{\mathbb{F}_q}$ has a model over some finite field, and isomorphisms over $\overline{\mathbb{F}_q}$ descend to some finite field. If A is an abelian variety over $\mathbb{F}_q$, we can also understand it as an abelian variety $A_{\mathbb{F}_{q^m}}$ over an extension $\mathbb{F}_{q^m}$, with the Frobenius given by the mth power: $\pi_{A_{\mathbb{F}_{q^m}}}= \pi_A^m$. Therefore for any place v the ratio $\frac{\mathrm{ord}_v(\pi_{A_{\mathbb{F}_{q^m}}})}{\mathrm{ord}_v(q^m)} = \frac{\mathrm{ord}_v(\pi_A)}{\mathrm{ord}_v(q)}$ is independent of m. In particular, if A is an abelian variety over $\overline{\mathbb{F}_q}$ and A_0 is a model over some finite field $\mathbb{F}_q$, then $s_A(v) := \frac{\mathrm{ord}_v(\pi_{A_0})}{ord_v(q)}$ does not depend on the model A or the chosen finite field $\mathbb{F}_q$.

Similarly, to understand Weil q-numbers as q varies, observe that $\pi \mapsto \pi^m$ gives a homomorphism $W_1(q) \to W_1(q^m)$ for every q and m, so we get a direct system of $W_1(q^m)$ and can define $W_1 = \varinjlim_m W_1(q^m)$. For any $\pi \in W_1$, we can find some q^m such that π has a representative π_m in $W_1(q^m)$; write $\mathbb{Q}[\pi]$ for the smallest field over $\mathbb{Q}$ generated by some representative π_m of π.

In the above case, π_{A_0} gives a representative for some $\pi_A \in W_1$, and for any embedding $\mathbb{Q}[\pi_{A_0}] \hookrightarrow \overline{\mathbb{Q}}$ the image of π_{A_0}, up to the action of Γ, is independent of the model A_0, and so we can view it as an embedding of $\mathbb{Q}[\pi]$. The result of Honda and Tate then shows that the same thing holds over $\overline{\mathbb{F}_q}$: $A \mapsto \pi_A$ defines a bijection $\Sigma(\mathsf{AV}^0(\overline{\mathbb{F}_q})) \to \Gamma\backslash W_1$, and when A is simple $\mathrm{End}^0(A)$ has center $F = \mathbb{Q}[\pi_A]$ with $\mathrm{inv}_v(\mathrm{End}^0(A))$ equal to $\frac{1}{2}$ for v real, $s_A(v)[F_v : \mathbb{Q}_p]$ for $v|p$, and 0 otherwise for v places of F.

Ultimately, we want to end up with some more general category subsuming abelian varieties over $\overline{\mathbb{F}_q}$, extending this property that the simple objects are in bijection with Galois orbits on some finitely generated $\mathbb{Z}$-module. (Eventually we should be able to find some natural subcategory whose numerical invariants coincide with those of $\mathsf{AV}^0(\overline{\mathbb{F}_q})$.)

One way to generate such categories is through representations of tori. For any torus T over a field F split by a (possibly infinite) Galois extension L/F with Galois group $\Gamma = \mathrm{Gal}(L/F)$, a representation of T on an F-vector space V is equivalent to giving an $X^*(T)$-grading $V(L) = \bigoplus_{\chi \in X^*(T)} V_\chi$ of $V(L)$ such that $\sigma V_\chi = V_\sigma$ for all $\sigma \in \Gamma$ and $\chi \in X^*(T)$, i.e. over L every representation decomposes into χ-isotypic components, and if the representation comes from one over F then they satisfy this Galois structure and so we can recover the original representation over F. In particular, if $L = \overline{F}$ we conclude that the category of representations $\mathrm{Rep}_F(T)$ of T over F is semisimple, with isomorphism classes of simple objects given by Galois orbits $\Gamma \backslash X^*(T)$. We can also compute the endomorphism algebras: for $\chi \in X^*(T)$, the corresponding representation V_χ has dimension given by the size of the Galois orbit of χ, and $\mathrm{End}(V_\chi) \simeq F(\chi)$, the subfield of $\overline{F}$ fixed by the subgroup of Γ fixing χ.

Since we can find a torus T with character group $X^*(T)$ isomorphic to any fixed finitely generated (continuous) $\mathbb{Z}[\Gamma]$-module M, we can find a semisimple F-linear category C with $\Sigma(C) = \Gamma \backslash M$ for any such M. However this is not completely satisfying for our case: often our endomorphism algebras for abelian varieties were noncommutative, and in this case all of the endomorphisms algebras $F(\chi)$ are commutative. To go further, we need a more general notion.

Fix a field F of characteristic 0, and let L/F be a Galois extension with Galois group Γ and G an algebraic group over F. For us, an extension of Γ by $G(L)$ is a short exact sequence

$$1 \to G(L) \to E \to \Gamma \to 1$$

compatible with the Galois action, i.e. if $e_\sigma \in E$ maps to $\sigma \in \Gamma$ then there is some $g \in G(L)$ such that

$$e_\sigma t e_\sigma^{-1} = g(\sigma t) g^{-1}$$

for all $t \in T(\overline{F})$. This defines an intertwining action by which we can define the split extension $E_G = G(L) \rtimes \Gamma$.

We say that an extension E is affine if its pullback to some open subgroup of Γ is split. This is equivalent to requiring that every σ lying in some open subgroup of Γ has a preimage $e_\sigma \in E$ such that $e_{\sigma\tau} = e_\sigma e_\tau$. In this case we say that G is the kernel of the extension.

If $G = T$ is commutative and we have an affine extension

$$1 \to T \to E \to \Gamma \to 1,$$

then we can restrict to some open subgroup of Γ in which $e_{\sigma\tau} = e_\sigma e_\tau$ up to an element of $T(L)$, which we denote by $d(\sigma, \tau)$; this gives a 2-cocycle $d : \Gamma \times \Gamma \to T(L)$, and the assumption that E is affine implies that we can choose d to be continuous. Therefore any affine extension E by a torus T gives a class $\mathrm{cl}(E) \in H^2(F, T)$.

In fact, this class gives an element of the Brauer group $\mathrm{Br}(F(\chi))$ for every $\chi \in X^*(T)$, with $F(\chi)$ the fixed field as above. To see this, first write $\mathrm{Br}(F(\chi))$ as $H^2(F(\chi), \mathbb{G}_\mathrm{m})$, its definition. By Shapiro's lemma, this is isomorphic to $H^2(F, \mathrm{Res}_{F(\chi)/F} \mathbb{G}_\mathrm{m})$. But we have a homomorphism of algebraic groups $T \to \mathrm{Res}_{F(\chi)/F} \mathbb{G}_\mathrm{m}$ dual to the morphism on character groups $X^*(\mathrm{Res}_{F(\chi)/F} \mathbb{G}_\mathrm{m}) = \mathbb{Z}[\Gamma/\Gamma(\chi)] \to X^*(T)$ defined by $\sum_\sigma n_\sigma \sigma \mapsto \sum_\sigma n_\sigma (\sigma\chi)$, and on cohomology this induces a homomorphism $H^2(F, T) \to H^2(F, \mathrm{Res}_{F(\chi)/F} \mathbb{G}_\mathrm{m}) \simeq \mathrm{Br}(F(\chi))$. Thus $\mathrm{cl}(E)$ gives an element of $\mathrm{Br}(F(\chi))$, and thus an equivalence class of central simple algebras over $F(\chi)$.

A homomorphism ϕ of affine extensions is a homomorphism of the corresponding exact sequences, i.e. a commutative diagram

$$\begin{tikzcd}
1 \arrow[r] & G_1(L) \arrow[r] \arrow[d] & E_1 \arrow[r] \arrow[d, "\phi"] & \Gamma \arrow[r] \arrow[d, equal] & 1 \\
1 \arrow[r] & G_2(L) \arrow[r] & E_2 \arrow[r] & \Gamma \arrow[r] & 1
\end{tikzcd}$$

such that the restriction of ϕ to $G_1(L)$ descends to a homomorphism $G_1(L) \to G_2(L)$ coming from a homomorphism of algebraic groups over L. We can even naturally define a 2-category structure: a morphism $\phi \to \phi'$ of homomorphisms $E_1 \to E_2$ of affine extensions is an element $g \in G_2(L)$ such that $\mathrm{ad}(g) \circ \phi = \phi'$, i.e. for every $e \in E_1$ we have $g\phi(e)g^{-1} = \phi'(e)$.

For an F-vector space V, we abbreviate the split extension by $\mathrm{GL}(V)$, $E_{\mathrm{GL}(V)}$, to E_V. A representation of an affine extension E is a homomorphism of affine extensions $E \to E_V$.

In the case where $E = E_G$ is a split extension, this is equivalent to giving a representation of G on V: the functor $\mathrm{Rep}(G) \to \mathrm{Rep}(E_G)$ is an equivalence of categories. This is proved using the fact that $H^1(\Gamma, \mathrm{GL}(V)) = 1$.

We now come to the main point: representations of affine extensions by tori give us the desired kind of generalization of representations of tori.

Proposition 1.1. *Let L/F be a Galois extension with Galois group Γ, T be a torus over F split by L, and E be an affine extension by T over L/F. Then $\mathrm{Rep}(E)$ is a semisimple F-linear category with $\Sigma(\mathrm{Rep}(E)) \simeq \Gamma\backslash X^*(T)$. Further, for V_χ the representation of E corresponding to $\chi \in X^*(T)$, the endomorphism algebra $\mathrm{End}(V_\chi)$ has center $F(\chi)$, and its class in $\mathrm{Br}(F(\chi))$ is the image of $\mathrm{cl}(E)$ under the homomorphism defined above.*

Proof. In the case where E is split, $\mathrm{cl}(E)$ is trivial and this is just the case of representations of tori discussed above. In general, a representation of E is a homomorphism of affine extensions $E \to E_V = E_{\mathrm{GL}(V)}$, which restricts to a homomorphism of algebraic groups $\phi|_T : T \to \mathrm{GL}(V)$, the category of which we know is semisimple. If $\phi|_T \simeq \phi_1 \oplus \phi_2$, since E_V is split it follows that ϕ composed with the inclusion $T \hookrightarrow E$ is a direct sum $\phi_1 \oplus \phi_2$ composed with Galois action on each factor separately, and so also decomposes as a direct sum, i.e. $\mathrm{Rep}(E)$ is semisimple, and $\phi \mapsto \phi|_T$ gives a map $\mathrm{Rep}(E) \to \mathrm{Rep}(T)$ which restricts to simple objects $\Sigma(\mathrm{Rep}(E)) \to \Sigma(\mathrm{Rep}(T)) \xrightarrow{\sim} \Gamma\backslash X^*(T)$. Every simple representation V_χ of T lifts to a representation of E by taking the pushout, and if two simple representations ϕ, ϕ' of E have the same image in $\Sigma(\mathrm{Rep}(T))$, i.e. the same restriction to T, both factor through the pushout and so by simplicity must be the same, i.e. this gives a bijection on simple objects. Finally, if V_χ corresponds to the character χ of T, the corresponding representation $\phi : E \to E_{V_\chi}$ of E is given by the pushout and has endomorphism algebra given by the subalgebra of $F(\chi)[\mathrm{GL}(V_\chi)]$ commuting with ϕ. Over a sufficiently large field this will split, and so its class in the Brauer group is determined by the Galois action, which is given by the Galois action on χ as described above; therefore it has the corresponding Brauer class, the image of $\mathrm{cl}(E)$ in $\mathrm{Br}(F(\chi))$. □

In fact we actually want something slightly more general: we want to allow the kernel to be not just tori but protori, i.e. limits (inverse limits) of tori. If $T = \varprojlim T_i$ over a field F, then $X^*(T) = \varinjlim X^*(T_i)$, and $T \mapsto X^*(T)$ defines an equivalence of categories between the category of protori and the category of torsion-free $\mathbb{Z}$-modules with a continuous action of $\Gamma = \mathrm{Gal}(\overline{F}/F)$, where $X^*(T)$ is torsion-free because it is a colimit (direct limit) of finitely generated free $\mathbb{Z}$-modules and therefore flat, and flat $\mathbb{Z}$-modules are torsion-free, and continuous means every element is fixed by an open subgroup. An affine extension with kernel T is a short exact sequence

$$1 \to T(\overline{F}) \to E \to \Gamma \to 1$$

whose pushout to each

$$1 \to T_i(\overline{F}) \to E_i \to \Gamma$$

by $T(\overline{F}) \to T_i(\overline{F})$ is an affine extension in our previous sense; a representation of such an affine extension is as above.

Suppose we have a commutative diagram of fields

$$\begin{tikzcd} L \arrow[r, hook] & L' \\ F \arrow[u, hook] \arrow[r, hook] & F' \arrow[u, hook] \end{tikzcd}$$

with $\Gamma = \mathrm{Gal}(L/F)$, $\Gamma' = \mathrm{Gal}(L'/F')$. An L/F-affine extension

$$1 \to G(L) \to E \to \Gamma \to 1$$

with kernel G over F yields an L'/F'-affine extension

$$1 \to G(L') \to E' \to \Gamma' \to 1$$

with kernel $G_{F'}$ by pulling back along the restriction map $\Gamma' \to \Gamma$ sending $\sigma \mapsto \sigma|_L$ and pushing out along $G(L) \to G(L')$.

For example, let $\mathbb{Q}_p^{\mathrm{unr}}$ be the maximal unramified extension of $\mathbb{Q}_p$, and let L_n be the unique subfield of $\mathbb{Q}_p^{\mathrm{unr}}$ with $[L_n : \mathbb{Q}_p] = n$, and $\Gamma_n = \mathrm{Gal}(L_n/\mathbb{Q}_p)$. For every $1 \leq i \leq n$ we can define a $\mathbb{Q}_p$-algebra $D_{i,n}$ as $L_n e_0 \oplus L_n e_1 \oplus \cdots \oplus L_n e_{n-1}$ as a $\mathbb{Q}_p$-vector space, with multiplication determined by $e_j c = \sigma^j c e_j$ for $c \in L$ and σ the Frobenius element in Γ_n, and $e_j e_l = e_{j+l}$ if $j + l \leq n - 1$ and $e_j e_l = \pi^i e_{j+l-n}$ for $j + l \geq n$, where π is a uniformizer. We can identify L with a subfield of $D_{i,n}$ by setting $e_0 = 1$, $e_1 = a$, and $e_j = a^j$ with $a^n = \pi^i$. Every central division algebra over $\mathbb{Q}_p$ is isomorphic to some $D_{i,n}$ for (i, n) relatively prime. We look in particular at $D_{1,n}$. By Proposition 1.1, if we can find an affine extension E by a torus whose image in the Brauer group in some extension agrees with that of $D_{1,n}$ (i.e. is $\frac{1}{n}$), then we can interpret $D_{1,n}$ as the endomorphism algebra of some representation of this extension.

Such an extension is given by

$$1 \to L_n^\times \to N(L_n^\times) \to \Gamma_n \to 1,$$

where $N(L_n^\times)$ is the normalizer of $L_n^\times$ in $D_{1,n}$ and is given by the disjoint union of $L_n^\times a^i$ over $0 \leq i \leq n-1$. This is an $L_n/\mathbb{Q}_p$-affine extension with kernel $\mathbb{G}_{\mathrm{m}}$. Pulling back by the restriction map $\Gamma \to \Gamma_n$ and pushing out by $L_n^\times \hookrightarrow \mathbb{Q}_p^{\mathrm{unr}\times}$ gives an affine extension

$$1 \to \mathbb{Q}_p^{\mathrm{unr}\times} \to D_n \to \mathrm{Gal}(\mathbb{Q}_p^{\mathrm{unr}}/\mathbb{Q}_p) \to 1$$

with kernel $\mathbb{G}_{\mathrm{m}}$. A representation $\rho : D_n \to E_V = \mathrm{GL}_{\mathbb{Q}_p^{\mathrm{unr}}}(V) \rtimes \mathrm{Gal}(\mathbb{Q}_p^{\mathrm{unr}}/\mathbb{Q}_p)$ of D_n is a vector space V over $\mathbb{Q}_p^{\mathrm{unr}}$ with a suitable action of D_n; the image of $(1, a)$ in D_n under ρ is some pair (F, τ), with F an automorphism of V commuting with the Galois action of τ. Since τ comes from the image of $(1, a)$, which acts on L_n by the Frobenius σ, its image in $\mathrm{Gal}(\mathbb{Q}_p^{\mathrm{unr}}/\mathbb{Q}_p)$ is precisely σ, so the structure on V is just a σ-linear automorphism F. Tensoring with $L = \widehat{\mathbb{Q}_p^{\mathrm{unr}}}$ gives an isocrystal, i.e. an L-vector space equipped with a σ-linear automorphism F. Over L, simple isocrystals are classified by rational numbers, by taking the local invariant of the endomorphism algebra, which will be a division algebra over L which are classified by rational numbers. In this case, our isocrystal can be decomposed as a sum of simple isocrystals E^{λ} with $\lambda \in \frac{1}{n}\mathbb{Z}$.

There is a canonical section of $N(L_n^{\times}) \to \Gamma_n$ by sending $\sigma^i \mapsto a^i$ for $0 \leq i \leq n-1$, which gives a canonical section of $D_n \to \Gamma$.

For varying n, we get a homomorphism $D_{nm} \to D_n$ whose restriction to the kernel $\mathbb{G}_{\mathrm{m}}$ is multiplication by m. Taking inverse limits gives a $\mathbb{Q}_p^{\mathrm{unr}}/\mathbb{Q}_p$-affine extension D with kernel $\mathbb{G} = \varprojlim \mathbb{G}_{\mathrm{m}}$, with $X^*(\mathbb{G}) = \varinjlim \frac{1}{n}\mathbb{Z}/\mathbb{Z} = \mathbb{Q}$. There is a natural functor from $\mathrm{Rep}(D)$ to the category of isocrystals, which is faithful and essentially surjective but not full. We call D the Dieudonné affine extension.

We now want to pick out a particular affine extension. Let $W(p^n)$ be the subgroup of $\overline{\mathbb{Q}}^{\times}$ generated by the Weil p^n-integers $W_1(p^n)$, and let $W = \varinjlim_n W(p^n)$. This is a free $\mathbb{Z}$-module of infinite rank and a continuous action of $\Gamma = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. As above, for $\pi \in W$ write $\mathbb{Q}[\pi]$ for the smallest field generated by a representative π_n of π; if π is represented by π_n and for some embedding into $\mathbb{C}$ it has absolute value $p^{nm/2}$, we say that it has weight $\mathrm{wt}(\pi) = m$, and for any prime v of $\mathbb{Q}[\pi]$ above p we write

$$s_\pi(v) = \frac{\mathrm{ord}_v(\pi_n)}{\mathrm{ord}_v(p^n)},$$

which as above is independent of the representative π_n.

Theorem 1.2. *Let P be the protorus over $\mathbb{Q}$ with $X^*(P) = W$. Then there exists an affine extension*

$$1 \to P(\overline{\mathbb{Q}}) \to \mathfrak{P} \to \Gamma \to 1,$$

unique up to isomorphism, whose representation category satisfies $\Sigma(\mathrm{Rep}(\mathfrak{P})) \simeq \Gamma\backslash W$, and for each $\pi \in W$ any representation V_π corresponding to π has endomorphism algebra D with center $\mathbb{Q}[\pi]$ and $\mathrm{inv}_v(D)$ is $\frac{\mathrm{wt}\,\pi}{2}$ if v is real, $s_\pi(v)[\mathbb{Q}[\pi]_v : \mathbb{Q}_p]$ if $v|p$, and 0 otherwise.

Proof. The description of the invariants of D defines a class $c(\pi)$ in $\mathrm{Br}(\mathbb{Q}[\pi])$. Thus to prove the result it suffices to show that there exists a unique class in $H^2(\mathbb{Q}, P)$ mapping to $c(\pi)$ for every $\pi \in \Gamma\backslash W$; then this class defines an affine extension $\mathfrak{P}$ with kernel P, whose simple representations are given by Galois orbits of the simple representations of P, which by definition is $\Gamma\backslash W$.

To see that there exists such a class, we want to find a unique preimage in $H^2(\mathbb{Q}, P)$ of $\prod_{\pi\in\Gamma\backslash W} c(\pi)$ in $\prod_{\pi\in\Gamma\backslash W} \mathrm{Br}(\mathbb{Q}[\pi])$. We can expand by class field theory: for any finite extension L/K of number fields and algebraic protorus T over K, we have a short exact sequence

$$0 \to L \to \mathbb{A}_L \to L\backslash\mathbb{A}_L \to 1$$

and therefore an exact sequence

$$1 \to T(L) \to T(\mathbb{A}_L) \to T(L \backslash \mathbb{A}_L).$$

When $H^1(T, L) = 0$, which will occur for both $K = \mathbb{Q}$, $T = P$ and $K = \mathbb{Q}[\pi]$, $T = \mathbb{G}_{\mathrm{m}}$, this in fact extends to a short exact sequence, i.e. the last map is surjective. In this case taking Galois cohomology for $G = \mathrm{Gal}(L/K)$ gives an exact sequence

$$0 \to H^2(G, T(L)) \to H^2(G, T(\mathbb{A}_L)) = \bigoplus_v H^2(\mathrm{Gal}(L_v/K_{N(v)}), T(L_v)) \to \cdots ,$$

and we can take the limit over L to get infinite extensions. In our cases, using the map $H^2(\mathbb{Q}, P) \to H^2(\mathbb{Q}[\pi], \mathbb{G}_{\mathrm{m}})$ we get a commutative diagram with exact rows

$$\begin{tikzcd}
0 \arrow[r] & H^2(\mathbb{Q}, P) \arrow[r] \arrow[d] & \bigoplus_v H^2(\mathbb{Q}_v, P_v) \arrow[d] & & \\
0 \arrow[r] & H^2(\mathbb{Q}[\pi], \mathbb{G}_{\mathrm{m}}) \arrow[r] & \bigoplus_u H^2(\mathbb{Q}[\pi]_u, \mathbb{G}_{\mathrm{m}}) \arrow[r] & \mathbb{Q}/\mathbb{Z} \arrow[r] & 0
\end{tikzcd}.$$

The image of $H^2(\mathbb{Q}, P)$ in $\bigoplus_v H^2(\mathbb{Q}_v, P_v)$ is exactly the set of classes whose image in $\mathbb{Q}/\mathbb{Z}$ vanish, which occurs for all our classes $c(\pi)$, so it suffices to check that each $c(\pi)_v$ lifts. For v real, P_v is just determined by the weight so clearly this is possible uniquely; for $v \nmid p$ there is nothing to say, so it suffices to check the conditions over p. If we view $s_\pi(v)$ as a function of π for a fixed v, we just need to know that it is well-defined independent of the model, which we know from the discussion above; then picking any representative, this is the order of the Galois orbit of π_n up to rescaling and therefore is the image under the π-map of some class in $H^2(\mathbb{Q}_p, P_p)$ which does not depend on π. Combining all the places together gives the result. □

This representation category $\mathrm{Rep}(\mathfrak{P})$ has numerical data strongly reminiscent of that of $\mathsf{AV}^0(\overline{\mathbb{F}_q})$. In particular, the isomorphism classes of simple objects, which we can identify with $\Gamma \backslash W$, has a subset given by the orbits $\Gamma \backslash W_1$ of π coming from honest Weil q-numbers for some q. This defines a full subcategory of $\mathrm{Rep}(\mathfrak{P})$ consisting of objects whose simple summands correspond to $\pi \in \Gamma \backslash W_1 \subset \Gamma \backslash W$, which we call the category of fake abelian varieties; notice that it is a semisimple $\mathbb{Q}$-linear category with the same numerical data as $\mathsf{AV}^0(\overline{\mathbb{F}_q})$. More generally, we call representations of $\mathfrak{P}$ fake motives over $\overline{\mathbb{F}_q}$, corresponding to how abelian varieties over $\overline{\mathbb{F}_q}$ generate the category of motives.

For each prime $\ell \neq p$, we can define a local form $\mathfrak{P}(\ell)$ of $\mathfrak{P}(\ell)$ as follows. Let w_ℓ be a prime of $\overline{\mathbb{Q}}$ over ℓ, and write $\overline{\mathbb{Q}}_\ell$ for the image of $\overline{\mathbb{Q}_\ell}$ in the completion of $\overline{\mathbb{Q}}$ at w_ℓ. We get a closed subgroup $\Gamma_\ell = \mathrm{Gal}(\overline{\mathbb{Q}}_\ell/\mathbb{Q}_\ell)$ of $\Gamma = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ corresponding to the commutative diagram of fields

$$\begin{tikzcd}
\overline{\mathbb{Q}} \arrow[r, hook] & \overline{\mathbb{Q}}_\ell \\
\mathbb{Q} \arrow[r, hook] \arrow[u] & \mathbb{Q}_\ell \arrow[u]
\end{tikzcd}.$$

We can obtain from $\mathfrak{P}$ a $\overline{\mathbb{Q}}_\ell/\mathbb{Q}_\ell$-affine extension $\mathfrak{P}(\ell)$ by pulling back by the restriction map $\Gamma_\ell \to \Gamma$ and pushing out along $P(\overline{\mathbb{Q}}) \to P(\overline{\mathbb{Q}}_\ell)$.

Proposition 1.3. *There is a continuous section $\zeta_\ell : \Gamma_\ell \to \mathfrak{P}(\ell)$ in the short exact sequence*

$$1 \to P(\overline{\mathbb{Q}}_\ell) \to \mathfrak{P}(\ell) \to \Gamma_\ell \to 1$$

defining $\mathfrak{P}(\ell)$.

Proof. This is the claim that the cohomology class of $\mathfrak{P}$ in $H^2(\mathbb{Q}, P)$ maps to zero in $H^2(\mathbb{Q}_\ell, P)$. From the commutative diagram in the proof of Theorem 1.2, this is true when it has trivial image in every place v of $\mathbb{Q}[\pi]$ over ℓ for every $\pi \in \Gamma \backslash W$; this image is just $c(\pi)_v = \mathrm{inv}_v(D)$ for D the corresponding endomorphism algebra to π, which is always 0 for $v \nmid p$, so if $v | \ell \neq p$ the result is immediate. □

Let $\rho : \mathfrak{P} \to E_V$ be a fake motive. Pulling back along $\Gamma_\ell \to \Gamma$ and pushing out along ℓ-completion applied to the restriction of ρ to P, i.e. the commutative diagram

$$\begin{tikzcd}
P(\overline{\mathbb{Q}}) \arrow[r] \arrow[d] & P(\overline{\mathbb{Q}}_\ell) \arrow[d] \\
\mathrm{GL}(V(\overline{\mathbb{Q}})) \arrow[r] & \mathrm{GL}(V(\overline{\mathbb{Q}}_\ell))
\end{tikzcd},$$

gives a homomorphism of affine extensions $\rho(\ell) : \mathfrak{P}(\ell) \to E_{V_\ell} = \mathrm{GL}(V(\overline{\mathbb{Q}}_\ell)) \rtimes \Gamma_\ell$, which is just a representation of $\mathfrak{P}(\ell)$ on $V(\mathbb{Q}_\ell)$.

Fixing a homomorphism $\zeta_\ell : \Gamma_\ell \to \mathfrak{P}(\ell)$ as in Proposition 1.3, we can take the composition $\Gamma_\ell \xrightarrow{\zeta_\ell} \mathfrak{P}(\ell) \xrightarrow{\rho(\ell)} \mathrm{GL}(V(\overline{\mathbb{Q}}_\ell)) \rtimes \Gamma_\ell$. Projecting onto the second factor gives the identity since ζ_ℓ is a section; write (e_σ, σ) for the image of $\sigma \in \Gamma_\ell$. Since each of ζ_ℓ and $\rho(\ell)$ are homomorphisms, we have $(e_{\sigma\tau}, \sigma\tau) = (e_\sigma, \sigma) \cdot (e_\tau, \tau) = (e_\sigma \sigma e_\tau, \sigma\tau)$ and so the e_σ satisfy $e_\sigma \circ \sigma e_\tau = e_{\sigma\tau}$ for $\sigma, \tau \in \Gamma_\ell$, and so we get a continuous action of Γ_ℓ on $V(\overline{\mathbb{Q}}_\ell)$ by $\sigma \cdot v = e_\sigma(\sigma v)$. Therefore $V_\ell(\rho) = V(\overline{\mathbb{Q}}_\ell)^{\Gamma_\ell}$ for this action defines a $\mathbb{Q}_\ell$-structure on $V(\overline{\mathbb{Q}}_\ell)$, giving us a functor $\rho \mapsto V_\ell(\rho)$ from the category of fake motives over $\overline{\mathbb{F}_q}$ to vector spaces over $\mathbb{Q}_\ell$.

Working similarly, one can define a functor sending ρ to a free module $V_f^p(\rho)$ over $\mathbb{A}_f^p$, the adeles away from ∞ and p, such that $V_\ell(\rho) = V_f^p(\rho) \otimes_{\mathbb{A}_f^p} \mathbb{Q}_\ell$ for all $\ell \neq p, \infty$.

We would like to expand to the prime p, which the above methods omit. To do so, what we want to assign to each fake motive is an isocrystal. Choose a prime w_p of $\overline{\mathbb{Q}}$ over p, and similarly write $\overline{\mathbb{Q}}_p$ for the image of $\overline{\mathbb{Q}_p}$ in the completion of $\overline{\mathbb{Q}}$ at w_p; also write $\mathbb{Q}_p^{\mathrm{unr}}$ for its image in this completion. Then $\Gamma_p = \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ is a closed subgroup of $\Gamma = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and $\Gamma_p^{\mathrm{unr}} = \mathrm{Gal}(\mathbb{Q}_p^{\mathrm{unr}}/\mathbb{Q}_p)$ is a quotient of Γ_p. One can similarly define $\mathfrak{P}(p)$ and $\mathfrak{P}(p)^{\mathrm{unr}}$ by the same procedure as for ℓ.

Proposition 1.4. *The affine extension $\mathfrak{P}(p)$ arises by pullback and pushout from $\mathfrak{P}(p)^{unr}$. Further, there is a homomorphism of $\mathbb{Q}_p^{unr}/\mathbb{Q}_p$-affine extensions $D \to \mathfrak{P}(p)^{unr}$ whose restriction to the kernels $\mathbb{G} \to P_{\mathbb{Q}_p}$ corresponds to the map on characters $W \to \mathbb{Q}$ sending $\pi \mapsto s_\pi(w_p)$.*

Proof. The first statement reduces to the statement that the image of the cohomology class of $\mathfrak{P}$ in $H^2(\Gamma_p, P(\overline{\mathbb{Q}}_p))$ lifts to $H^2(\Gamma, P(\mathbb{Q}_p^{\mathrm{unr}}))$, which follows (I think) from Shapiro's lemma. The second statement is more straightforward: the map of extensions corresponds to a map on cohomology, which arises from the map of characters, and so we just need to know that the Galois actions match up, which is clear from the invariance of the s_π. □

A fake motive $\rho : \mathfrak{P} \to E_V$ gives rise to a representation of $\mathfrak{P}(p)$ just as for ℓ; doing the pullback-pushforward process (in reverse) gives a representation of $\mathfrak{P}(p)^{\mathrm{unr}}$. Composing with the homomorphism $D \to \mathfrak{P}(p)^{\mathrm{unr}}$ gives a representation of D, which we know is the same thing as an isocrystal $D(\rho)$.

Given an abelian variety of CM type over $\overline{\mathbb{Q}}$, we saw that it has good reduction and therefore defines an abelian variety over $\overline{\mathbb{F}_q}$. In fact it also defines a fake abelian variety, i.e. a representation of $\mathfrak{P}$ with simple summands corresponding to Weil q-integers for some q. For any abelian variety A of CM type (E, Φ) over $\overline{\mathbb{Q}}$, let $T = \mathrm{Res}_{E/\mathbb{Q}} \mathbb{G}_{\mathrm{m}}$; then Φ defines a cocharacter $\mu_\Phi : \mathbb{G}_{\mathrm{m}} \to T$. Thus we are in the following more general situation, which will be useful later:

Proposition 1.5. *Let T be a torus over $\mathbb{Q}$ split by a CM field, and let μ be a cocharacter of T such that $\mu + \overline{\mu}$ is defined over $\mathbb{Q}$. Then there is a homomorphism $\phi_\mu : \mathfrak{P} \to E_T$ well-defined up to isomorphism.*

Proof. The data of T and μ corresponds to an abelian variety A with T acting on $H_1(A, \mathbb{Q})$. By the equivalence of categories between $\mathsf{AV}^0(\overline{\mathbb{F}_q})$ and fake abelian varieties, there is some $\phi_\mu : \mathfrak{P} \to E_V$ corresponding to A, which factors through the action of T. □

Now in our case T acts naturally on $V = H_1(A, \mathbb{Q})$, and composing with $\mathfrak{P} \to E_T$ gives a fake abelian variety $\rho : \mathfrak{P} \to E_V$ such that $V_\ell(\rho) = (H_1(A, \mathbb{Q}) \otimes \overline{\mathbb{Q}_\ell})^{\Gamma_\ell} = H_1(A, \mathbb{Q}_\ell)$ and $D(\rho)$, the isocrystal corresponding to composing ρ with the map $D \to \mathfrak{P}$, is isomorphic to the Dieudonné module of the reduction of A modulo p. This essentially follows from the fact that the invariants of $\mathrm{End}(\overline{A})$ (or equivalently the endomorphism algebra of the corresponding fake motive) are (uniquely) compatible with the description of the Frobenius of $\overline{A}$ from the Taniyama-Shimura formula.

2. Good reduction of Shimura varieties

Since we've seen that a Shimura variety $\mathrm{Sh}_K = \mathrm{Sh}_K(G, X)$ has a unique canonical model over its reflex $E = E(G, X)$, we now identify Sh_K with its canonical model and speak of it as a variety over E.

When the Shimura variety has a moduli interpretation over $\mathbb{C}$, e.g. as a moduli space of abelian varieties with some additional structure, then this description descends to $\overline{\mathbb{Q}}$ since the data themselves do. For example, if Sh_K is the Siegel modular variety attached to a symplectic space (V, ψ), the $\overline{\mathbb{Q}}$-points $\mathrm{Sh}_K(\overline{\mathbb{Q}})$ classify isomorphism classes of triples $(A, s, \eta K)$ where A is an abelian variety over $\overline{\mathbb{Q}}$, s is a rational multiple of a divisor up to equivalence (thus corresponding by the cycle class map to a Hodge tensor in $H^2(A, \mathbb{Q})$ and so a polarization over $\mathbb{C}$) or more precisely an element of $\mathrm{NS}(A) \otimes \mathbb{Q}$ containing a $\mathbb{Q}^\times$-multiple of an ample divisor, where $\mathrm{NS}(A)$ is the Nèron-Severi group of A, i.e. divisors modulo algebraic equivalence, and ηK is a K-orbit of isomorphisms $V(\mathbb{A}_f) \to V_f(A)$ sending ψ to an $\mathbb{A}_f^\times$-multiple of (the pairing corresponding to) s.

For more general Shimura varieties, where the proof of the existence of the canonical model does not pass through the moduli interpretation, the situation is more complicated and no good description of $\mathrm{Sh}_K(\overline{\mathbb{Q}})$ is known.

We would like to be able to describe $\mathrm{Sh}_K(L)$ for fields all the way down to E, at least when $\mathrm{Sh}_K(\mathbb{C})$ has a moduli interpretation (for example when (G, X) is of abelian type). However first we encounter another problem: for A an abelian variety over $\overline{\mathbb{Q}}$, suppose that we know that σA is abstractly isomorphic to A for every $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/E)$. Does it follow that A is defined over E?

If we choose for each σ an isomorphism $f_\sigma : \sigma A \to A$, for these to form a descent datum we need them to satisfy the cocycle condition $f_\sigma \circ \sigma f_\tau = f_{\sigma\tau}$. An obstruction to the existence of such a cocycle lies in the second cohomology set $H^2(\mathrm{Gal}(\overline{\mathbb{Q}}/E), \mathrm{Aut}(A))$.

To avoid this issue, we could hope that $\mathrm{Aut}(A)$ is trivial, at least if we require the automorphisms to preserve the extra data of A classified by Sh_K, if K is sufficiently small. If the additional axiom (5) holds, i.e. the center $Z(\mathbb{Q})$ is discrete in $Z(\mathbb{A}_f)$, then this is true: this gives a sort of rigidity condition. In general however this may fail. For example, in the Siegel case the center is $\mathbb{G}_\mathrm{m}$ and so this axiom holds, and so (for K sufficiently small) for any field L containing $E = \mathbb{Q}$ we get a moduli description of $\mathrm{Sh}_K(L)$ as classifying triples $(A, s, \eta K)$ with the same description as for $\overline{\mathbb{Q}}$, replacing $\overline{\mathbb{Q}}$ by L. On the other hand, for Hilbert modular varieties, where $G = \mathrm{Res}_{F/\mathbb{Q}} M_2(F)$ for some totally real field F and so its center is $\mathrm{Res}_{F/\mathbb{Q}} \mathbb{G}_\mathrm{m}$, and so axiom (5) fails: $F^\times$ is not discrete in $\mathbb{A}_{F,f}^\times$ by strong approximation. In this case we can describe $\mathrm{Sh}_K(\overline{\mathbb{Q}})$ as above, but all we can say for number fields is $\mathrm{Sh}_K(L) = \mathrm{Sh}_K(\overline{\mathbb{Q}})^{\mathrm{Gal}(\overline{\mathbb{Q}}/L)}$.

Ultimately we'd like to be able to talk about points of Shimura varieties not just over fields of characteristic 0 but also characteristic p, in particular finite fields by reduction from the case of number fields. To get a good notion of reduction of Shimura varieties, we need to make sure that our subgroups K are compatible with the extension-reduction process:

Definition 2.1. Let G be a reductive group over $\mathbb{Q}$ (or even over $\mathbb{Q}_p$). A subgroup $K \subset G(\mathbb{Q}_p)$ is called hyperspecial if there exists a flat group scheme $\mathcal{G}$ over $\mathbb{Z}_p$ such that $\mathcal{G}_{\mathbb{Q}_p} = G$ (i.e. $\mathcal{G}$ is an extension of G to $\mathbb{Z}_p$), $\mathcal{G}_{\mathbb{F}_p}$ is a connected reductive group, which automatically has the same dimension as G by flatness (this is the good reduction condition), and $G(\mathbb{Z}_p) = K$ (i.e. K is compatible with this process).

For example, let $G = \mathrm{GSp}(V, \psi)$, as in the Siegel case. Fix a $\mathbb{Z}_p$-lattice in $V(\mathbb{Q}_p)$, and let K_p be the stabilizer of Λ. Then K_p is hyperspecial if the restriction of ψ to $\Lambda \times \Lambda \subset V \times V$ takes values in $\mathbb{Z}_p$ and is perfect, i.e. induces an isomorphism $\Lambda \to \Lambda^\vee$. The reduction modulo p gives a nondegenerate alternating pairing $\Lambda/p\Lambda \times \Lambda/p\Lambda \to \mathbb{F}_p$, with $\mathcal{G}_{\mathbb{F}_p}$ the group of symplectic similitudes of this pairing.

In the PEL case, where we have a semisimple $\mathbb{Q}$-algebra with involution B, in order for there to exist a hyperspecial subgroup B must be unramified over p, i.e. $B \otimes_{\mathbb{Q}} \mathbb{Q}_p$ must be a product of matrix algebras over unramified extensions of $\mathbb{Q}_p$. In this case there is a similar description to the Siegel case.

By results of Tits, it is known that there is a hyperspecial subgroup in $G(\mathbb{Q}_p)$ if and only if G is unramified over $\mathbb{Q}_p$, i.e. quasisplit over $\mathbb{Q}_p$ and split over an unramified extension.

For the rest of this section we fix a hyperspecial subgroup $K \subset G(\mathbb{Q}_p)$; for K^p a compact open subgroup of $G(\mathbb{A}_f^p)$, we write $K = K^p K_p$ and write $\mathrm{Sh}_p(G, X)$ for the limit of $\mathrm{Sh}_K(G, X)$ with K of this form, i.e. $\mathrm{Sh}_p(G, X) = \varprojlim_{K^p} \mathrm{Sh}_{K^p K_p}(G, X)$. This Shimura variety $\mathrm{Sh}_p(G, X)$ carries an action of $G(\mathbb{A}_f^p)$.

There are two essential reasons why a Shimura variety may fail to have good reduction at a prime dividing p: either the reductive group in question may be ramified at p, or p may divide the level (i.e. K is badly behaved at p). For example, a Shimura curve is defined by a quaternion algebra B over $\mathbb{Q}$; if p divides the discriminant of B, then the Shimura curve will have bad reduction at p, or the modular curve $\Gamma_0(N)\backslash\mathcal{H}$ will have bad reduction at $p|N$.

Let $E = E(G, X)$. We say that $\mathrm{Sh}_p(G, X)$ has good reduction at a prime $\mathfrak{p}$ of E if the inverse system $\mathrm{Sh}_p(G, X) = (\mathrm{Sh}_{K^pK_p}(G, X))_{K^p}$ extends to an inverse system of flat schemes $\mathcal{S}_p = (\mathcal{S}_{K^p})_{K^p}$ over the ring of integers $\widehat{\mathcal{O}}_\mathfrak{p}$ of the completion $E_\mathfrak{p}$ of E at $\mathfrak{p}$ with the action of $G(\mathbb{A}_f^p)$ on Sh_p extending to an action on $\mathcal{S}_p$, whose reduction modulo $\mathfrak{p}$ is an inverse system of varieties $\overline{\mathrm{Sh}}_p(G, X) = (\overline{\mathrm{Sh}}_{K^pK_p}(G, X))_{K^p}$ over the residue field $k(\mathfrak{p})$ of $\widehat{\mathcal{O}}_\mathfrak{p}$, such that for $K^p \supset K'^p$ sufficiently small the corresponding map

$$\overline{\mathrm{Sh}}_{K'^pK_p} \to \overline{\mathrm{Sh}}_{K^pK_p}$$

is a finite étale map of smooth varieties.

Generally, a variety over $E_\mathfrak{p}$ may fail to have good reduction to a smooth variety over $k(\mathfrak{p})$ (e.g. for elliptic curves) and if it does the reduction is generally not unique, with no obvious way to distinguish between them; for example, given one reduction we can blow up at a smooth subvariety of the closed fiber to obtain another. We say that a variety V with good reduction corresponding to an extension $\mathcal{V}$ over $E_\mathfrak{p}$ has canonical good reduction at $\mathfrak{p}$ if for any formally smooth scheme T over $\widehat{\mathcal{O}}_\mathfrak{p}$, the natural map

$$\mathrm{Hom}_{\widehat{\mathcal{O}}_\mathfrak{p}}(T, \mathcal{V}) \to \mathrm{Hom}_{E_\mathfrak{p}}(T_{E_\mathfrak{p}}, V)$$

is an isomorphism. We can apply this in particular with $V = \mathrm{Sh}_p$ and $\mathcal{V} = \mathcal{S}_p$; this is a limit of schemes étale over a smooth scheme and so formally smooth, and so this characterizes the model $\mathcal{S}_p$ uniquely up to unique isomorphism by the Yoneda lemma.

The following theorem is due to Mumford in the Siegel case, various authors independently including Kottwitz in the PEL case, and Vasiu and Kisin in the Hodge case, from which the abelian case follows. The existence of good reduction for hyperspecial subgroups was first conjectured by Langlands; the notion of the canonical reduction was conjectured by Milne.

Theorem 2.2. *Let (G, X) be a Shimura datum of abelian type, and p be any prime other than a finite set of primes depending only on (G, X), K_p a fixed hyperspecial subgroup of $G(\mathbb{Q}_p)$, and $\mathrm{Sh}_p(G, X)$ the corresponding inverse system of varieties over $E = E(G, X)$. Then $\mathrm{Sh}_p(G, X)$ has canonical good reduction at every prime $\mathfrak{p}$ of E dividing p.*

3. The Langlands-Rapoport conjecture

We next want to give a description of the points of reductions of Shimura varieties. This is conjecturally possible by work of Langlands and Rapoport.

We start by analogy in the complex case. Fix a Shimura datum (G, X) satisfying the additional axioms (4), (5), and (6). If we take the limit over K^p of complex points we get as schemes over $\mathbb{C}$

$$\mathrm{Sh}_p(\mathbb{C}) = \mathrm{Sh}(\mathbb{C})/K_p = \lim_{K^p} \mathrm{Sh}_{K^pK_p}(\mathbb{C}).$$

For each $x \in X$, let $I(x) \subset G(\mathbb{Q})$ be the stabilizer of x, and write $X^p(x) = G(\mathbb{A}_f^p)$, $X_p(x) = G(\mathbb{Q}_p)/K_p$, and $S(x) = I(x)\backslash X^p(x) \times X_p(x)$ (the sets $X^p(x)$ and $X_p(x)$ clearly do not depend on x, but this will be notationally convenient for the analogy). It is easy to see that there is a bijection

$$\bigsqcup_{x \in G(\mathbb{Q})\backslash X} S(x) \to \mathrm{Sh}_p(\mathbb{C})$$

by expanding:

$$\mathrm{Sh}_p(\mathbb{C}) = G(\mathbb{Q})\backslash X \times G(\mathbb{A}_f)/K_p = G(\mathbb{Q})\backslash X \times X^p \times X_p$$

(since we have the additional axioms), from which the decomposition is immediate. This has a modular interpretation: for example, for (G, X) of Hodge type, the set $S(x)$ classifies the isomorphism classes of triples $(A, (s_i)\eta K)$ with $(A, (s_i))$ isomorphic to a fixed abelian variety with tensors.

The idea of Langlands and Rapoport is that $\overline{\mathrm{Sh}}_p(\overline{\mathbb{F}_p})$ should have a similar description, where we replace the indexing set $G(\mathbb{Q})\backslash X$, which in the complex case classifies complex abelian varieties with tensors, with a set of isomorphism classes of homomorphisms $\phi : \mathfrak{P} \to E_G$. Note that for any faithful representation $G \hookrightarrow \mathrm{GL}(V)$ this gives a homomorphism $\mathfrak{P} \to G \rtimes \Gamma \to \mathrm{GL}(V) \rtimes \Gamma = E_V$, i.e. a fake motive; this is also equipped with additional data, namely the representation $G \hookrightarrow \mathrm{GL}(V)$ corresponds to tensors t_i for V such that G is the subgroup of $\mathrm{GL}(V)$ fixing the t_i. Thus we can think of ϕ as a fake motive with tensors, modulo representations of G, analogous to the data classified by $G(\mathbb{Q})\backslash X$. We are interested in isomorphism classes of homomorphisms $\mathfrak{P} \to E_G$; recall that an isomorphism of such homomorphisms ϕ, ϕ' is given by $g \in G(\overline{\mathbb{Q}})$ such that $\phi'(a) = g\phi(a)g^{-1}$ for all $a \in \mathfrak{P}$.

For a fixed $\phi : \mathfrak{P} \to E_G$, we now want to define a set $S(\phi)$ analogous to $S(x)$ above. We first define the analogue $I(\phi)$ of $I(x)$: this is the stabilizer of ϕ, i.e. the subgroup of $G(\overline{\mathbb{Q}})$ consisting of g such that $\mathrm{ad}(g) \circ \phi = \phi$, i.e. $\phi(a) = g\phi(a)g^{-1}$ for every $a \in \mathfrak{P}$.

Next, we define $X^p(\phi)$. For each prime $\ell \neq p, \infty$, as in section 1 choose a prime w_ℓ of $\overline{\mathbb{Q}}$ over ℓ and define $\overline{\mathbb{Q}}_\ell$ and $\Gamma_\ell \subset \Gamma$ as there. We can view Γ_ℓ as a $\overline{\mathbb{Q}}_\ell/\mathbb{Q}_\ell$-affine extension with trivial kernel, i.e.

$$1 \to 1 \to \Gamma_\ell \to \Gamma_\ell \to 1,$$

and we get a canonical morphism of affine extensions $\xi_\ell : \Gamma_\ell \to E_G(\ell) = G(\overline{\mathbb{Q}}_\ell) \rtimes \Gamma_\ell$ sending $\sigma \mapsto (1, \sigma)$. On the other hand ϕ gives a homomorphism $\phi(\ell) : \mathfrak{P}(\ell) \to E_G(\ell)$, and we have a morphism $\zeta_\ell : \Gamma_\ell \to \mathfrak{P}(\ell)$ from section 1 with which we can compose to get a second homomorphism $\phi(\ell) \circ \zeta_\ell : \Gamma_\ell \to \mathfrak{P}(\ell)$. We can then look at 2-morphisms between these homomorphisms, all of which are isomorphisms, and so we define

$$X_\ell(\phi) = \mathrm{Iso}(\xi_\ell, \phi(\ell) \circ \zeta_\ell),$$

which carries an action of $G(\mathbb{Q}_\ell)$ on the right and of $I(\phi)$ on the left. Elements of this set correspond to the data of the level structures η: in particular, picking a faithful representation $\rho : G \hookrightarrow \mathrm{GL}(V)$ composing with ξ_ℓ and $\phi(\ell) \circ \zeta_\ell$ gives $X_\ell(\phi)$ as a subset of $\mathrm{Iso}(V(\mathbb{Q}_\ell), V_\ell(\rho \circ \phi))$.

By choosing the ζ_ℓ judiciously (as to define $V_f^p(\rho)$ in section 1), we obtain compact subspaces and so we can take the restricted product $X^p(\phi)$ over all $\ell \neq p, \infty$. If it is nonempty, this gives a principal homogeneous space for $G(\mathbb{A}_f^p)$ acting on the right.

At p, we similarly choose a prime w_p of $\overline{\mathbb{Q}}$ over p and use the notation from section 1. From Proposition 1.4, we get a map $D \to \mathfrak{P}(p)^{\mathrm{unr}}$, and composing with $\phi(p)^{\mathrm{unr}}$ gives a map $D \to G(\mathbb{Q}_p^{\mathrm{unr}}) \rtimes \Gamma_p^{\mathrm{unr}}$. This composite factors through D_n for some n. The Frobenius $\sigma \in \Gamma_p^{\mathrm{unr}}$, sending $x \mapsto x^p$ on the residue field, lifts uniquely to D; let (b, σ) be its image in $G(\mathbb{Q}_p^{\mathrm{unr}}) \rtimes \Gamma_p^{\mathrm{unr}}$. By the definition of the semidirect product, projecting onto the first factor gives an element $b = b(\phi)$ which is well-defined up to σ-conjugation, i.e. any different choice of b is given by $g^{-1}b\sigma g$ for some $g \in G(\mathbb{Q}_p^{\mathrm{unr}})$. If $\rho : G \to \mathrm{GL}(V)$ is a faithful representation, then the isocrystal $D(\rho \circ \phi)$ is the vector space $V(L)$, where $L = \widehat{\mathbb{Q}_p^{\mathrm{unr}}}$, with the σ-linear map F given by $b\sigma$, i.e. $v \mapsto b\sigma v$.

Recall that the Shimura datum (G, X) defines a $G(\overline{\mathbb{Q}})$-conjugacy class $c(X)$ of cocharacters of $G_{\overline{\mathbb{Q}}}$. We can transfer this to a conjugacy class of cocharacters of $G_{\overline{\mathbb{Q}}_p}$; since we assume G contains a hyperspecial group, it splits over $\mathbb{Q}_p^{\mathrm{unr}}$ and so this class contains an element μ defined over $\mathbb{Q}_p^{\mathrm{unr}}$. Let

$$C_p = G(\mathcal{O}_L) \cdot \mu(p) \cdot G(\mathcal{O}_L),$$

where $\mathcal{O}_L = W(\overline{\mathbb{F}_p})$ is the ring of integers of L and $G(\mathcal{O}_L)$ should be interpreted as $\mathcal{G}(\mathcal{O}_L)$ for an extension $\mathcal{G}$ of G as in the definition of a hyperspecial subgroup K_p. We can then define

$$X_p(\phi) = \{g \in G(L)/G(\mathcal{O}_L) | g^{-1}b(\phi)g \in C_p\}.$$

The automorphisms $I(\phi)$ of ϕ act naturally on this set, since they are the $g \in G(\overline{\mathbb{Q}})$ such that $g\phi g^{-1} = \phi$ and so act on $X_p(\phi)$ simply by multiplication. We also have a Frobenius action: for $g \in X_p(\phi)$, define

$$\Phi(g) = b(\phi)\sigma b(\phi)\sigma^2 \cdots \sigma^{m-1}b(\phi)\sigma^m g,$$

where $m = [E_{\mathfrak{p}} : \mathbb{Q}_p]$. For example, if $m = 1$, so $\Phi(g) = b\sigma g$, we have $\Phi(g)^{-1}b\Phi(g) = g^{-1}\sigma^{-1}b^{-1}bb\sigma g = g^{-1}\sigma^{-1}b\sigma g$, which is in C_p because σ fixes $X_p(\phi)$.

We can now define $S(\phi)$, formally parallel to the complex case:

$$S(\phi) = I(\phi)\backslash X^p(\phi) \times X_p(\phi),$$

with the action of $I(\phi)$ on both factors as described and an action of $G(\mathbb{A}_f^p)$ on the right through the action on $X^p(\phi)$, as well as the action of Φ via its action on $X_p(\phi)$.

We want to restrict to certain "admissible" ϕ. This translates to a local condition at each place together with a global condition.

We first treat the condition at infinity. Let E_∞ be the extension

$$1 \to C^\times \to E_\infty \to \Gamma_\infty \to 1$$

given by the quaternion algebra $\mathbb{H} = \mathbb{C}^\times \sqcup \mathbb{C}^\times j$, where $\Gamma_\infty = \mathrm{Gal}(\mathbb{C}/\mathbb{R})$, and regard it as an affine extension with kernel $\mathbb{G}_{\mathrm{m}}$. Pulling back and pushing out as usual gives a $\mathbb{C}/\mathbb{R}$-affine extension

$$1 \to P(\mathbb{C}) \to \mathfrak{P}(\infty) \to \Gamma_\infty \to 1.$$

One can check that these extensions have corresponding cohomology classes, which implies that there is a homomorphism $\zeta_\infty : E_\infty \to \mathfrak{P}(\infty)$ whose restriction to the kernels $\mathbb{G}_{\mathrm{m}} \to P_{\mathbb{C}}$ corresponds to the map on characters $\pi \mapsto \mathrm{wt}(\pi)$.

Write ι for complex conjugation, the nontrivial element of Γ_∞.

Proposition 3.1. *For any $x \in X$, the formulas*

$$\xi_x(z) = (w_X(z), 1), \qquad \xi_x(j) = (\mu_x(-1)^{-1}, \iota)$$

define a homomorphism $E_\infty \to E_G(\infty)$. Changing x does not change the isomorphism class of the homomorphism.

Proof. Since $E_\infty = \mathbb{C}^\times \sqcup \mathbb{C}^\times j$, these formula are independent and together define a homomorphism as claimed, so the only thing to change is that the isomorphism class is independent of x. Only the second depends on x and only through μ_x; all μ_x are conjugate, so this is immediate. □

Write ξ_X for the isomorphism class of homomorphisms $E_\infty \to E_G(\infty)$ of Proposition 3.1. On the other hand we can construct a morphism $E_\infty \to E_G(\infty)$ by composing $\zeta_\infty : E_\infty \to \mathfrak{P}(\infty)$ with $\phi(\infty) : \mathfrak{P}(\infty) \to E_G(\infty)$. The condition at infinity is that this composition should give an element of ξ_X.

The other local conditions are simpler: the condition at ℓ is that $X_\ell(\phi)$ should be nonempty, and the condition at p is also that $X_p(\phi)$ be nonempty (though these are defined differently).

Finally, we also want ϕ to satisfy a global condition: let $\nu : G \to T$ be the quotient of G by its derived group G^{der}. From X we get a conjugacy class of cocharacters of $G_{\mathbb{C}}$, which descends to a well-defined cocharacter μ of T. By our assumptions on (G, X), T and μ satisfy the conditions of Proposition 1.5, i.e. $\mu + \overline{\mu}$ is defined over $\mathbb{Q}$ and T is split by a CM field; thus there is a homomorphism $\phi_\mu : \mathfrak{P} \to E_T$. On the other hand composing with ν gives another homomorphism $\mathfrak{P} \xrightarrow{\phi} E_G \xrightarrow{\nu} E_T$; the global condition is that these be isomorphic, i.e. $\nu \circ \phi \simeq \phi_\mu$. If ϕ satisfies this condition and all the local conditions, we say that it is admissible.

We can now define the Langlands-Rapoport set $\mathrm{LR}(G, X)$: this is the disjoint union of $S(\phi)$ over all isomorphism classes of admissible homomorphisms $\phi : \mathfrak{P} \to E_G$. This set carries commuting actions of $G(\mathbb{A}_f^p)$ and Φ coming from the actions on each $S(\phi)$.

Conjecture 3.2 (Langlands-Rapoport). *Let (G, X) be a Shimura datum satisfying the additional axioms (4), (5), (6) such that G^{der} is simply connected, and let K_p be a hyperspecial subgroup of $G(\mathbb{Q}_p)$. Let $\mathfrak{p}$ be a prime of $E(G, X)$ over p such that Sh_p has canonical good reduction at $\mathfrak{p}$. Then there is a bijection of sets*

$$\mathrm{LR}(G, X) \to \overline{\mathrm{Sh}}_p(G, X)(\overline{\mathbb{F}_p})$$

compatible with the actions of $G(\mathbb{A}_f^p)$ and Φ.

It is possible to make more general conjectures, more complicated to state but not essentially deeper, without the axioms (4), (5), and (6). One can also remove the assumption that G^{der} is simply connected, at the cost of replacing the notion of admissible homomorphisms with a more complicated one, called special homomorphisms. However it is known (due to Milne it seems) that the conjecture for G^{der} simply connected implies the general case. One can also generalize to zero-dimensional Shimura varieties, in which case we should add that the bijection should commute with the induced map to the Shimura variety of the torus T.

Conjecture 3.2 follows in the case of PEL Shimura varieties corresponding to quaternion algebras over totally real fields from work of Honda and Tate. Even the case of general PEL Shimura varieties appears to be out of reach, however. Milne has shown that the conjecture at least for Shimura varieties of Hodge type follows from a sufficiently good theory of motives.

4. Counting points

The description of Conjecture 3.2 is far from explicit, and it is not clear that one could use it to actually compute anything. In this section we will derive from it a formula for the number of $\mathbb{F}_q$-points of $\overline{\mathrm{Sh}}_p$.

Let (G, X) be a Shimura datum satisfying the additional axioms (4), (5), and (6), and fix a hyperspecial subgroup $K_p \subset G(\mathbb{Q}_p)$. We assume that G^{der} is simply connected and $\mathrm{Sh}_p(G, X)$ has canonical good reduction at a prime $\mathfrak{p}$ of $E = E(G, X)$ over p (for example, when (G, X) is of abelian type and unramified at p by Theorem 2.2). We otherwise use the same notation as previous sections, such as L_n for the unramified extension of $\mathbb{Q}_p$ of degree n. Let $\mathbb{F}_q$ be a finite field containing the residue field $k(\mathfrak{p})$ of $E_{\mathfrak{p}}$.

To say anything over $\mathbb{F}_q$, we first need a version of Conjecture 3.2 descending to finite fields from the algebraic closure. To get this, we just need to equip the left-hand side with a Galois action so we can take invariants, i.e. we should replace the data composing $\mathrm{LR}(G, X)$ with data incorporating a Frobenius automorphism ϵ acting on ϕ, i.e. an element of $I(\phi)(\mathbb{Q})$. We can then replace all of the data by corresponding data with suitable ϵ-actions: we define $I(\phi, \epsilon)$ to be the automorphism group of the pair (ϕ, ϵ), i.e. the centralizer of ϵ in $I(\phi)$; $X^p(\phi, \epsilon)$ is the subset of $X^p(\phi)$ consisting of elements fixed by ϵ, and $X_p(\phi, \epsilon)$ is the subset of $X_p(\phi)$ such that ϵ acts by Φ^r where $r = [\mathbb{F}_q : k(\mathfrak{p})]$. It suffices to maintain the admissibility conditions on ϕ; since we want $X_p(\phi, \epsilon)$ to again be nonempty, we can equivalently require that $X_p(\phi)$ have some element on which ϵ acts by Φ^r, in which case (ϕ, ϵ) is said to be an admissible pair. If we define $\mathrm{LR}_{\mathbb{F}_q}(G, X)$ to be the disjoint union of the quotient $S(\phi, \epsilon) = I(\phi, \epsilon)\backslash X^p(\phi, \epsilon) \times X_p(\phi, \epsilon)$ over isomorphism classes of pairs (ϕ, ϵ) with ϕ admissible and $\epsilon \in I(\phi)$, we again conjecture a bijection $\mathrm{LR}_{\mathbb{F}_q}(G, X) \to \overline{\mathrm{Sh}}_p(G, X)(\mathbb{F}_q)$; this conjecture turns out to follow from Conjecture 3.2, so it suffices to assume that one.

We could also get a formula for points of $\overline{\mathrm{Sh}}_{K^p K_p}(\mathbb{F}_q)$ by quotienting each $S(\phi, \epsilon)$ by K^p; working at a fixed level K^p is sufficient since we just take the limit at the end, and often convenient, so we do so (and generally omit it from the notation).

Fixing an admissible pair (ϕ, ϵ), since $\epsilon \in I(\phi)(\mathbb{Q}) \subset G(\mathbb{Q})$, it has an image $\gamma \in G(\mathbb{A}_f^p)$ under the map $G(\mathbb{Q}) \to G(\mathbb{A}_f) \to G(\mathbb{A}_f^p)$. The set $X^p(\phi, \epsilon)$ is the restricted product of isomorphisms $\xi_\ell \to \phi(\ell) \circ \zeta_\ell$ over primes $\ell \neq p$ fixing ϵ or equivalently γ_ℓ. Such an isomorphism of homomorphisms of affine extensions $\Gamma_\ell \to E_G(\ell)$ is by definition an element of $G(\mathbb{Q}_\ell)$ such that $\mathrm{ad}(g) \circ \xi_\ell = \phi(\ell) \circ \zeta_\ell$. Working at level K^p, this is true up to the action of K^p for any g, so $X_\ell(\phi, \epsilon)$ is just the subset of $G(\mathbb{Q}_\ell)/K_\ell$ fixing γ_ℓ under the right action, and so $X^p(\phi) \simeq Y^p(\gamma) = \{g \in G(\mathbb{A}_f^p)/K^p | \gamma g \equiv g \pmod{K^p}\}$.

To understand $X_p(\phi, \epsilon)$, we need to do a little more. Let $\mathcal{O} = W(\mathbb{F}_q)$, and $B = \operatorname{Frac} \mathcal{O} = L_{[\mathbb{F}_q:\mathbb{F}_p]}$. By definition, $X_p(\phi, \epsilon)$ is the subset of $\{g \in G(L)/G(\mathcal{O}_L) | g^{-1} b(\phi) g \in G(\mathcal{O}_L) \cdot \mu(p) \cdot G(\mathcal{O}_L)\}$ on which ϵ acts by Φ^r, where $b(\phi)$ is the image (up to σ-conjugacy) in $G(\mathbb{Q}_p^{\mathrm{unr}})$ of the lift of the Frobenius $\sigma \in \Gamma_p^{\mathrm{unr}}$ to D under $D \to G(\mathbb{Q}_p^{\mathrm{unr}}) \rtimes \Gamma_p^{\mathrm{unr}}$. The condition on the action of ϵ reduces to restricting g to $G(B)/G(\mathcal{O})$, in which case each element of the conjugacy class $g^{-1}bg$ can be taken over $G(B)$, i.e. $b(\phi)$ is represented by some $\delta \in G(B)$, i.e. $g^{-1}bg = g^{-1}\delta\sigma g$, and so we can write $X_p(\phi, \epsilon)$ as the set $Y_p(\delta) = \{g \in G(B)/G(\mathcal{O}) | g^{-1}\delta\sigma g \in G(\mathcal{O}) \cdot \mu(p) \cdot G(\mathcal{O})\}$. Write $I = I(\phi, \epsilon)$; on $\mathbb{A}_f^p$-points, this is just the centralizer $G(\mathbb{A}_f^p)_\gamma$ of γ, and on $\mathbb{Q}_p$-points it is the σ-centralizer $G_{\delta\sigma}(\mathbb{Q}_p)$ of δ, i.e. g such that $g^{-1}\delta\sigma g = \delta$. Thus we can rewrite

$$S(\phi, \epsilon) = I(\phi, \epsilon)(\mathbb{Q}) \backslash Y^p(\gamma) \times Y_p(\delta).$$

The order of this is now something we can actually compute: let f^p be the indicator function of K^p on $G(\mathbb{A}_f^p)$, and ϕ_r be the indicator function of $C_p(\mathcal{O}) = G(\mathcal{O}) \cdot \mu(p) \cdot G(\mathcal{O})$. Then

$$\begin{aligned} |I(\mathbb{Q}) \backslash Y^p(\gamma) \times Y_p(\delta)| &= \int_{I(\mathbb{Q}) \backslash G(\mathbb{A}_f^p) \times G(B)} f^p(g_1^{-1}\gamma g_1)\phi_r(g_2^{-1}\delta\sigma g_2) \\ &= \int_{(I(\mathbb{Q}) \backslash G(\mathbb{A}_f^p)_\gamma \times G_{\delta\sigma}(B)) \times (G(\mathbb{A}_f^p)_\gamma \backslash G(\mathbb{A}_f^p)) \times (G_{\delta\sigma}(B) \backslash G(B))} f^p(g_2^{-1}\gamma g_2)\phi_r(g_3^{-1}\delta\sigma g_3) \\ &= \mathrm{vol}(I(\mathbb{Q}) \backslash G(\mathbb{A}_f^p)_\gamma \times G_{\delta\sigma}(B)) \cdot \int_{I(\mathbb{A}_f^p) \backslash G(\mathbb{A}_f^p)} f^p(g^{-1}\gamma g)\, dg \\ &\qquad \cdot \int_{I(\mathbb{Q}_p) \backslash G(B)} \phi_r(g^{-1}\delta\sigma g)\, dg, \end{aligned}$$

where the integrals are with respect to Haar measures on $G(\mathbb{A}_f^p)$ giving measure 1 to K^p, on $I(\mathbb{A}_f^p)$ and $I(\mathbb{Q}_p)$ giving rational measure to compact open subgroups, and on $G(B)$ giving measure 1 to $G(\mathcal{O})$. We call these integrals $\mathrm{O}_\gamma(f^p)$ and $\mathrm{TO}_\delta(\phi_r)$ respectively, and writing $I(\mathbb{A}_f) = I(\mathbb{A}_f^p) \times I(\mathbb{Q}_p)$ conclude that

$$|I(\mathbb{Q}) \backslash Y^p(\gamma) \times Y_p(\delta)| = \mathrm{vol}(I(\mathbb{Q}) \backslash G(\mathbb{A}_f)) \cdot \mathrm{O}_\gamma(f^p) \cdot \mathrm{TO}_\delta(\phi_r).$$

We define this quantity to be $\mathrm{I}(\gamma_0; \gamma, \delta)$.

Note that $Y^p(\gamma)$ and $Y_p(\delta)$ depend only on γ and δ respectively, as written, and not on the original choice of (ϕ, ϵ). We'd like to be able to work only in terms of γ and δ; they are not arbitrary elements, but have the property that in $G(\overline{\mathbb{Q}}_\ell)$ each γ_ℓ is conjugate to the same element coming from $G(\mathbb{Q})$, which we call γ_0, and δ is such that

$$\delta\sigma\delta\sigma^2\delta \cdots \sigma^{[\mathbb{F}_q:\mathbb{F}_p]-1}\delta$$

is conjugate to the same γ_0 in $G(\overline{\mathbb{Q}}_p)$. We also impose an admissibility condition, which implies that γ_0 is elliptic in $G(\mathbb{R})$, i.e. contained in an elliptic torus of $G_\mathbb{R}$ (i.e. anisotropic in $G_\mathbb{R}^{\mathrm{ad}}$).

We can also define I purely in terms of this triple $(\gamma_0; \gamma, \delta)$. Set $I_0 = G_{\gamma_0}$, the centralizer of γ_0 in G. Since γ_0 is semisimple and G^{der} is simply connected, this is a connected and reductive group. We set I_∞ to be the inner form of $(I_0)_\mathbb{R}$ whose image in $G_\mathbb{R}^{\mathrm{ad}}$ is anisotropic; more precisely, if T is an elliptic maximal torus of $G_\mathbb{R}$ containing γ_0 and x is such that h_x factors through T, then $\mathrm{ad}h_x(i)$ preserves $(I_0)_\mathbb{R}$ and induces a Cartan involution on its image in $G_\mathbb{R}^{\mathrm{ad}}$, which gives a suitable twist of $(I_0)_\mathbb{R}$. For $\ell \neq p$, we let I_ℓ be the centralizer of γ_ℓ in $G_{\mathbb{Q}_\ell}$, and I_p is the inner form of G consisting of $g \in G(B)$ such that $g^{-1}\delta\sigma g = \delta$. The remainder of the admissibility condition is that I_0 admits an inner form I such that $I_{\mathbb{Q}_v} \simeq I_v$ for every place v, including p and ∞.

We've seen that we can attach such a triple to an admissible pair (ϕ, ϵ), with isomorphic pairs giving isomorphic triples; the inner form I is given by $I(\phi, \epsilon)$ in this case. We get a map from the set of isomorphism classes of admissible pairs (ϕ, ϵ) to the set of isomorphism classes of triples; we call a triple effective if its isomorphism class is in the image of this map.

In general, this map is not a bijection, but it is finite-to-one. For any reductive group Υ over $\mathbb{Q}$ we have a map

$$H^i(\mathbb{Q}, \Upsilon) \to \prod_v H^i(\mathbb{Q}_v, \Upsilon),$$

and we write $\mathrm{Ker}^i(\mathbb{Q}, \Upsilon)$ for its kernel.

Proposition 4.1. *Fix an effective triple $(\gamma_0; \gamma, \delta)$. Then the number of isomorphism classes of admissible pairs mapping to the class of $(\gamma_0; \gamma, \delta)$ is finite and given by the cardinality of the kernel*

$$\ker(\mathrm{Ker}^1(\mathbb{Q}, I_0) \to H^1(\mathbb{Q}, G))$$

of the map induced by the inclusion $I_0 \to G$.

Proof. Since $(\gamma_0; \gamma, \delta)$ is effective, there exists at least one admissible pair (ϕ_0, ϵ_0) whose associated triple is in the same class as $(\gamma_0; \gamma, \delta)$. Let (ϕ, ϵ) be another such pair. Then $\mathrm{Hom}((\phi_0, \epsilon_0), (\phi, \epsilon))$ is an $\mathrm{Aut}(\phi_0, \epsilon_0) = I(\phi_0, \epsilon_0)$-torsor and so gives a class in $H^1(\mathbb{Q}, I(\phi_0, \epsilon_0))$. Since the two pairs define the same triple, after localizing at each place the torsor is trivial, and so gives an element of $\mathrm{Ker}^1(\mathbb{Q}, I(\phi_0, \epsilon_0))$. There is a natural map $I(\phi_0, \epsilon_0) \hookrightarrow G \twoheadrightarrow G^{\mathrm{ab}}$; for any torus there is only one admissible ϕ by Proposition 1.5, so our torsor in fact gives an element of $\ker(\mathrm{Ker}^1(\mathbb{Q}, I(\phi_0, \epsilon_0)) \to H^1(\mathbb{Q}, G^{\mathrm{ab}}))$. Any torsor is given by such a homomorphism, and the condition that the corresponding class vanish after localizing at each place implies that the corresponding (ϕ, ϵ) must map to the same triple, so this is a surjection; if (ϕ', ϵ') gives rise to the same class, it defines the same torsor and so is isomorphic to (ϕ, ϵ), so this is also a surjection.

The short exact sequence

$$1 \to G^{\mathrm{der}} \to G \to G^{\mathrm{ab}} \to 1$$

induces

$$G^{\mathrm{ab}}(\mathbb{Q}) \to H^1(\mathbb{Q}, G^{\mathrm{der}}) \to H^1(\mathbb{Q}, G) \to H^1(\mathbb{Q}, G^{\mathrm{ab}}) \to H^2(\mathbb{Q}, G^{\mathrm{der}}),$$

so the vanishing of our class in $H^1(\mathbb{Q}, G^{\mathrm{ab}})$ means that it lifts to an element of the image of $H^1(\mathbb{Q}, G^{\mathrm{der}})$ in $H^1(\mathbb{Q}, G)$. Since G^{der} is simply connected, by the Hasse principle

$$H^1(\mathbb{Q}, G^{\mathrm{der}}) \to \prod_v H^1(\mathbb{Q}_v, G^{\mathrm{der}})$$

is injective, so the vanishing of our class at each v implies that it is trivial in $H^1(\mathbb{Q}, G)$, i.e. we can replace G^{ab} with G in the above. An elementary lemma of Langlands and Rapoport shows that we can replace $I(\phi_0, \epsilon_0)$ with I_0 (of which it is an inner form) without changing the cardinality, so we get a bijection between the fiber over $(\gamma_0; \gamma, \delta)$ and the kernel claimed. □

Call this cardinality $c(\gamma_0; \gamma, \delta)$.

Corollary 4.2. *Let (G,X) be a Shimura datum satisfying the hypotheses of Conjecture 3.2, and suppose that Conjecture 3.2 is true for $\overline{\mathrm{Sh}}_p(G,X)$. Then*

$$|\overline{\mathrm{Sh}}_p(G,X)(\mathbb{F}_q)| = \sum_{[\gamma_0;\gamma,\delta]} c(\gamma_0;\gamma,\delta)\cdot \mathrm{I}(\gamma_0;\gamma,\delta),$$

where the sum is over isomorphism classes of effective triples.

Again, we can also take both sides with level structure, as in the description above.

Let's apply this description in a concrete case: counting elliptic curves with level structure. Let $(G,X)=(\mathrm{GL}_2,\mathcal{H})$, so that $\mathrm{Sh}_K(G,X)$ are the modular curves. The fiber over a triple corresponds to abelian varieties (in this case elliptic curves) isogenous to a fixed representative E_0; for example, if we take E_0 supersingular, since all supersingular elliptic curves form an isogeny class we can count supersingular curves with level structures as $I(\gamma_0;\gamma,\delta)$. In this case we can take γ_0 trivial; in this case the centralizer is all of GL_2, so $\mathrm{O}_\gamma(f^p)$ is just the measure evaluated on $\{1\}$, i.e. $\frac{1}{\operatorname{vol} K^p}$. For simplicity, let's work over $\overline{\mathbb{F}_p}$; then the Frobenius is trivial and so $\mathrm{TO}_\delta(\phi_r)$ is also trivial, so $\mathrm{I}(\gamma_0;\gamma,\delta)=\operatorname{vol}(I(\mathbb{Q})\backslash G(\mathbb{A}_f))\cdot\frac{1}{\operatorname{vol}(K^p)}$. Since we want to vary K^p, make a different choice of Haar measure on $G(\mathbb{A}_f^p)$: choose the measure on each factor such that $\mathrm{GL}_2(\mathbb{Z}_\ell)$ has measure 1. This makes evaluating the first factor easier and the second factor more complicated (rather than trivial).

In this case $I=\mathrm{Aut}(\phi,\epsilon)=\mathrm{Aut}(E_0)$ is a quaternion algebra over $\mathbb{Q}$, split away from p and ∞. We have $\operatorname{vol}(I(\mathbb{Q})\backslash G(\mathbb{A}_f))=\operatorname{vol}((I(\mathbb{Q})\backslash I(\mathbb{A}_f))\operatorname{vol}(I(\mathbb{A}_f)\backslash G(\mathbb{A}_f))$; since I is split away from p and ∞, the second factor is just $\operatorname{vol}(I(\mathbb{Q}_p)\backslash G(\mathbb{Q}_p^{\mathrm{unr}}))$, and I splits over $\mathbb{Q}_p^{\mathrm{unr}}$ so this is also 1; and the first factor is the order of $I(\mathbb{Q})\backslash I(\mathbb{A}_f)/\,\mathrm{O}(\mathbb{A}_f)$ where O is a maximal order of I (given by honest endomorphisms of E_0). By p-adic uniformization this is in bijection with the isogeny class of E_0, i.e. the set of supersingular curves over $\overline{\mathbb{F}_p}$ with no level structure, so adding K^p-level structure changes the number of curves by a factor of $\frac{1}{\operatorname{vol} K^p}$. We can compute this explicitly for $K^p=\Gamma(N)$ (with $p\nmid N$): for $\ell\nmid N$, the local factor is just $\mathrm{GL}_2(\mathbb{Z}_\ell)$ since N is invertible in $\mathbb{Z}_\ell$, so $\operatorname{vol}(K_\ell)=\operatorname{vol}(\mathrm{GL}_2(\mathbb{Z}_\ell))=1$. If ℓ divides N a times, then modulo ℓ^a we have $\operatorname{vol}(\mathrm{GL}_2(\mathbb{Z}_\ell)/K_\ell)=\operatorname{vol}(\mathrm{GL}_2(\mathbb{Z}/\ell^a\mathbb{Z})/B)$ for B the Borel subgroup over $\mathbb{Z}/\ell^a\mathbb{Z}$. This quotient classifies full flags in $(\mathbb{Z}/\ell^a\mathbb{Z})^2$, i.e. one-dimensional subspaces of $(\mathbb{Z}/\ell^a\mathbb{Z})^2$, i.e. the projective line over $\mathbb{Z}/\ell^a\mathbb{Z}$, which has $\ell^{a-1}(\ell+1)$ rational points, so $\operatorname{vol} K_\ell=\frac{\operatorname{vol}(\mathrm{GL}_2(\mathbb{Z}_\ell))}{\ell^{a-1}(\ell+1)}=\frac{1}{\ell^{a-1}(\ell+1)}$. Therefore this factor is

$$\prod_{\ell^a\|N}\ell^{a-1}(\ell+1)=N\prod_{\ell|N}\left(1+\frac{1}{\ell}\right).$$

References

[1] James S Milne. Introduction to Shimura varieties. *Harmonic analysis, the trace formula, and Shimura varieties*, 4:265–378, 2005.