SELECTIONS FROM THE ARITHMETIC GEOMETRY OF SHIMURA CURVES I: MODULAR CURVES

PETE L. CLARK

These are notes accompanying a sequence of three lectures given in the UGA number theory seminar in August and September of 2006. I originally planned to give a single lecture on some Diophantine problems involving Atkin-Lehner twists of modular curves, but have been persuaded by Dino Lorenzini to expand it into three lectures and include Shimura curves as well.

This first lecture is on (classical) modular curves. It is a rather "straight up" expository account of the subject, suitable for people who have heard about modular curves before but seen little about them. After reviewing some truly classical[1] results about modular curves over the complex numbers, we will discuss rational and integral canonical models, focusing in on the case of $X_0(N)$ for squarefree N. We hope to end this first lecture by revealing a (presumably) surprising connection between the geometry of modular curves and the arithmetic of quaternion algebras.

1. Uniformization of Riemann surfaces

I will for the most part be making a bunch of interrelated statements of fact. Most of them are nontrivial and the reader should not be alarmed by their inability to supply a proof. Sometimes when some statement follows especially easily from another statement, I rapidly indicate a proof.

Convention: All algebraic varieties we meet will be *connected* and *nonsingular*. In this section, all varieties will be defined over $\mathbb{C}$. The first fact that we will be taking for granted is that every (nonsingular!) algebaraic variety $V_{/\mathbb{C}}$ gives rise to a complex manifold, whose set of points is $V(\mathbb{C})$. Moreover, the connectedness of V in the algebraic sense – or, if you like, the connectedness of $V(\mathbb{C})$ endowed with the *Zariski* topology - is equivalent to the connectedness of the associated complex manifold – or, if you like, the connectedness of $V(\mathbb{C})$ endowed with the *analytic* topology (which is strictly finer than the Zariski topology unless $V(\mathbb{C})$ is finite).[2]

Now suppose C is a complex algebraic curve, so that $C(\mathbb{C})$ can be endowed with the structure of a Riemann surface (i.e., a one-dimensional complex manifold). On the other hand, not every Riemann surface arises this way. The precise condition for a Riemann surface to arise in this way – to be algebraic – is that it be the complement of a finite set in a compact Riemann surface. In this case the compactification is well-determined up to isomorphism – in particular, its genus $g(M)$

[1]Classical does not imply trivial, of course, nor even devoid of contemporary research interest.

[2]Note that this is a special property of the complex numbers $\mathbb{C}$. It does not even hold, e.g., for varieties over $\mathbb{R}$. In fact the (finite) number of real-analytic components of a (Zariski-)connected variety $V_{/\mathbb{R}}$ is an interesting invariant.

is well-determined – as is the number $p(M)$ of points required to compactify. (The corresponding algebraic curve is *projective* if $p(M) = 0$ and affine if $p(M) > 0$.)

To say a bit more: given any connected Riemann surface M, the *uniformization theorem* asserts the following: (i) the universal covering $\tilde{M}$ can be endowed with the structure of a Riemann surface in a unique way so that the covering map $\pi : \tilde{M} \to M$ is a local $\mathbb{C}$-analytic isomorphism; and (ii) there are, up to equivalence, precisely three simply connected Riemann surfaces: the Riemann sphere $\mathbb{P}^1$; the Euclidean plane $\mathbb{A}^1$; and the upper half-plane $\mathcal{H}$.[3]

$\mathbb{P}^1$ covers only itself.[4] Given an unramified covering $\mathbb{A}^1 \to M$, we can view the deck transformation group Λ as a lattice in $\mathbb{A}^1$ and hence $M = \mathbb{A}^1/\Lambda$. Aside from the trivial case of $\Lambda = 0$, we have the case in which $\Lambda \cong \mathbb{Z}$ and the quotient is isomorphic to $\mathbb{C} \setminus 0$, the multiplicative group; otherwise $\Lambda \cong \mathbb{Z}^2$ and M is an elliptic curve.

In contrast to $\mathbb{P}^1$ and $\mathbb{A}^1$, the upper halfplane $\mathcal{H}$ is not itself algebraic, which makes it a much richer source of Riemann surfaces: indeed, by the process of elimination, every Riemann surface which is not $\mathbb{P}^1, \mathbb{A}^1$, $\mathbb{G}_m$ or an elliptic curve is uniformized by $\mathcal{H}$: in particular every compact Riemann surface of genus at least 2.

2. Fuchsian groups

In real life, most group actions have fixed points; in particular, it is important to expand our notation of "uniformized Riemann surface" to include (finite) ramification. This goes as follows: first, the group of biholomorphic automorphisms of $\mathcal{H}$ is $PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/\pm 1$ acting by usual linear fractional transformations: $z \mapsto \frac{az+b}{cz+d}$.

(Remarks: More generally we have $GL_2(\mathbb{C})$ acting on $\mathbb{P}^1(\mathbb{C}) = \mathbb{C} \cup \infty$, $GL_2(\mathbb{R})$ acting on $\mathcal{H}^{\pm} = \mathbb{P}^1(\mathbb{C}) \setminus \mathbb{P}^1(\mathbb{R})$, and $GL_2(\mathbb{R})^+$, the matrices of positive determinant, acting on $\mathcal{H}$. However, scalar matrices give rise to the trivial linear transformation, and $GL_2(\mathbb{R})^+/\text{scalars} = PSL_2(\mathbb{R})$.)

By definition, a **Fuchsian group** Γ is a discrete subgroup of $PSL_2(\mathbb{R})$, so every Fuchsian group acts on $\mathcal{H}$, and we may consider the map

$$\pi_\Gamma : \mathcal{H} \mapsto \Gamma\backslash\mathcal{H} = Y(\Gamma).$$

It turns out that $Y(\Gamma)$ can be given the structure of a Riemann surface, uniquely specified by the condition that π_Γ be a quotient map – that is, for any Riemann surface M, a function $f : Y(\Gamma) \to M$ is holomorphic iff its pullback $f \circ \pi_\Gamma$ to $\mathcal{H}$ is holomorphic.[5] Note well that this map is in general ramified (over a discrete but

[3]The Riemann mapping theorem asserts that any simply connected proper domain in $\mathbb{C}$ is biholomorphic to $\mathcal{H}$, so can be used in place of $\mathcal{H}$. Occasionally it is also useful to work with the open unit disk.

[4]Exercise: Give as many different proofs as you can, using (for instance) Euler characteristic, the Riemann-Hurwitz theorem, the Lefschetz Fixed point theorem, Luroth's Theorem...

[5]This "categorical" perspective on quotients of manifolds with extra structure seems to be better known among European mathematicians, since it is used systematically in the Bourbaki books.

at the moment possibly infinite set of points).

Remark: In some sense this is a "lucky fact," turning on the especially simple local structure of ramified finite-to-one maps of Riemann surfaces (up to local biholomorphism, there are only the maps $z \mapsto z^n$). If I am not mistaken, given a locally compact group G acting properly discontinuously on a locally compact space X, the only general assertion that can be made about the quotient $G\backslash X$ is that it is a Hausdorff space. In particular, the quotient of even a finite group G of biholomorphic automorphisms of a higher-dimensional complex manifold can have singularities ("orbifold points").

What is the condition on Γ such that $Y(\Gamma)$ be an algebraic Riemann surface? It is most easily phrased as follows: there exists on $\mathcal{H}$ a very nice hyperbolic[6] metric

$$ds^2 = \frac{dxdy}{y^2};$$

in particular it is $PSL_2(\mathbb{R})$-invariant. (After remarking that conformal automorphisms of planar regions preserve angles and ds^2, being a rescaling of the Euclidean metric, measures angles in the same way, what remains to be checked is that linear fractional transformations preserve hyperbolic *length*, which, if you like, is the explanation for why the factor of y^2 has been inserted in the denominator.) We have the notion of a *fundamental domain* D for Γ – this is the closure D of a connected open subset of $\mathcal{H}$ such that the Γ-translates of D tile $\mathcal{H}$ (i.e., $\mathcal{H} = \bigcup_{\gamma\in\Gamma} \gamma D$, and distinct translates intersect only along ∂D) and of its volume v with respect to the hyperbolic metric.[7]

Theorem 1. *For a Fuchsian group Γ, the quotient surface $Y(\Gamma) = \Gamma\backslash\mathcal{H}$ is algebraic iff Γ has finite hyperbolic volume.*

The special case in which $Y(\Gamma)$ is compact – i.e., Γ has a compact fundamental domain – will be especially important in what follows. In older terminology, these are said to be Fuchsian groups **of the first kind**.

Example: Take $\Gamma(1) = PSL_2(\mathbb{Z})$. Then, as is well-known, a fundamental domain for Γ can be obtained by taking $|\Re(z)| \leq \frac{1}{2}$, $|z| \geq 1$, and the identifications along the boundary are such that $\Gamma(1)\backslash\mathcal{H}$ is, as a topological space, homeomorphic to $\mathbb{A}^1$. Moreover, the volume is $\frac{\pi}{3} < \infty$, so it is isomorphic to $\mathbb{A}^1$ as a Riemann surface (and not to the other, nonalgebraic Riemann surface to which it is homeomorphic, namely $\mathcal{H}$). Note that the uniformization map $\pi : \mathcal{H} \to Y(1)$ is ramified over two points of $Y(1)$ (it is also, in a sense in which we are trying to avoid discussing in detail, infinitely ramified over the missing point at ∞), namely the distinguished points $e^{2\pi i/4}$ and $e^{2\pi i/6}$ of the fundamental region. These are, respectively, the unique fixed points in $\mathcal{H}$ of the elements $S = \ldots$ and $T = \ldots$. S and T are, up to conjugacy, the only nontrivial elements of $\Gamma(1)$ of finite order. It follows that if $\Gamma \subset \Gamma(1)$ is any subgroup containing no conjugacy class of S or T – and in

[6]We will ignore the temptation – quite strong in August 2006 – to digress on the general topic of geometric structures on manifolds.

[7]Since there are many different fundamental domains for a given Γ, it formally speaking must be checked that the volume is independent of the chosen domain, but this is in fact rather easy.

particular if Γ is normal and contains neither S nor T – then the map

$$\pi_\Gamma : \mathcal{H} \to Y(\Gamma)$$

is the universal covering map.

For any $N \geq 1$, we define the subgroup $\Gamma(N) \subset PSL_2(\mathbb{Z})$ as the projectivization of the kernel of the (surjective) homomorphism $SL_2(\mathbb{Z}) \to SL_2(\mathbb{Z}/N\mathbb{Z})$, $Y(N) = Y(\Gamma(N))$, and $X(N)$ the compactification of $Y(N)$. For $N \geq 2$, $\Gamma(N)$ has no elements of finite order.

Example: $X(2)$ is the projective line and $Y(2)$ is the projective line with three points removed. It follows that $\Gamma(2)$ is the fundamental group of the complex projective line with three points removed, i.e., a free group on 2 generators.

Now a remarkable theorem of Belyi states that every (compact, connected, nonsingular) algebraic curve C defined over some number field – i.e., over $\overline{\mathbb{Q}}$ – admits a map to $\mathbb{P}^1$ branched over 3 points. Removing the branch points, we get an unramified cover from C minus its branch points to $Y(2)$. This means that there is a finite index subgroup Γ of $\Gamma(2)$ such that the compactification of $Y(\Gamma)$ is isomorphic to C! In other words, if we allow a modular curve to be the compactification of some finite-index subgroup of the modular group $\Gamma(1)$, then every *algebraic* algebraic curve is a modular curve (and conversely, because the ramification is over only three points, one can see that every modular curve can be defined over some number field).

Although on the face of it this result says that all algebraic curves over $\overline{\mathbb{Q}}$ have a "hidden modular structure," in practice it is not clear how to exploit this "extra structure" and it seems that we should read this result in the other direction: i.e., to get interesting results we should restrict the class of finite index subgroups of $\Gamma(1)$. This brings us to the definition of a **congruence subgroup** $\Gamma \subset \Gamma(1)$, which is a subgroup containing $\Gamma(N)$ for some N.

Exercise 1: Use the above reasoning to explain why there must be many noncongruence finite-index subgroups of $\Gamma(1)$. (One can in fact show that the probability that an index d subgroup of $\Gamma(1)$ is a congruence subgroup goes to 0 as $d \to \infty$.)

Exercise 2: Show that there are infinitely many finite-index subgroups $\Gamma \subset \Gamma(1)$ such that $Y(\Gamma)$ has genus 0.

Remark: Henceforth we will only discuss noncongruence subgroups as a source of counterexamples, and by a *modular curve* we shall mean a curve of the form $Y(\Gamma)$ for a congruence subgroup $\Gamma \subset_2 (\mathbb{Z})$.

Particular congruence subgroups of interest are $\Gamma_0(N)$ and $\Gamma_1(N)$; the corresponding curves are denoted $Y_0(N)$ and $Y_1(N)$ and their compactified versions as $X_0(N)$ and $X_1(N)$. We thus have, for any N, finite covering maps

$$X(N) \to X_1(N) \to X_0(N) \to X_1(1) \cong \mathbb{P}^1$$

as well as, for any $M \mid N$, natural maps $X_\bullet(N) \to X_\bullet(M)$.

Cusps: Since $Y_0(1)$ has genus zero and needs precisely one point in order to be compact, there exists a coordinate function J on $X_0(1)$ – i.e., an isomorphism $X_0(1) \xrightarrow{\sim} \mathbb{P}^1$ which carries the unique point of $X_0(1) \setminus Y_0(1)$ to the point at ∞, which carries $e^{2\pi i/4}$ to 1728 and $e^{2\pi i/6}$ to 0 (since the automorphism group of $\mathbb{P}^1$ is triply transitive, we can send any three points anywhere we like). Indeed, the classical theory tells us that the function $\tau \mapsto j(E_\tau)$ descends to give such a coordinate function which has especially nice arithmetic properties: in particular, the field of moduli of $_\tau$ is $\mathbb{Q}(j(E_\tau)$. We are avoiding mentioning this (important!) classical construction in the interest of saving time.

Having fixed such a coordinate, we speak of the point at ∞ of $X_0(1)$ and call it the **cusp** of $X_0(1)$. Moreover, for any $\Gamma \subset \Gamma(1)$, the **cusps** on $X(\Gamma)$ are precisely the preimages of the single cusp ∞ on $X_0(1)$. In other words, the cusps are precisely the points we must add in order to compactify $Y(\Gamma)$. The theory of Fuchsian groups describes the cusps explicitly in terms fixed points of certain "parabolic" elements of Γ on the boundary $\mathbb{RP}^1 \cong S^1$ of $\mathcal{H}$, another important and useful construction we will omit for lack of time.

We will be most interested in the family of curves $X_0(N)$ for squarefree N. In this case it can be shown that the number of cusps is precisely equal to the number of positive divisors of N, i.e., 2^r if $N = \prod_{i=1}^r p_i$. (This would follow quite easily if only we described cusps in the usual way. Later we will deduce it as a consequence of the structure of the modular automorphism group of $X_0(N)$.)

It is not especially hard to compute the genus of $X_0(N)$. The degree of the covering $X_0(N) \to X_1(N)$ is $\psi(N) = \prod_i (p_i + 1)$, so by Riemann-Hurwitz it suffices to figure out the ramification over $J = 0,\ 1728$ and ∞. This can be done in an elementary way, although it is probably more enlightening to do the calculation after one has the moduli interpretation in mind. Anyway, here is the well-known answer: put

$$e_2(N) = \prod_{p \mid N} (1 + \left(\frac{-4}{p}\right)),$$

$$e_3(N) = \prod_{p \mid N} (1 + \left(\frac{-3}{p}\right)).$$

Then

$$g(X_0(N)) = 1 + \frac{\psi(N)}{12} - \frac{e_2(N)}{4} - \frac{e_3(N)}{3} - 2^{r-1}.$$

In other words the genus is roughly $\frac{N}{12}$, and is indeed asymptotic to it when we bound the number of prime divisors of N. In particular it tends to ∞ with N. The (squarefree) values of N for which $X_0(N)$ has genus 0 (i.e., is itself isomorphic to $\mathbb{P}^1$) are:

$$N = 1, 2, 3, 5, 6, 7, 10, 13$$

and the values for which it has genus 1 are

$$N = 11, 14, 15, 17, 19, 21.$$

Remark: For any N, there are known formulas for the genera of $X_0(N)$, $X_1(N)$, $X(N)$ and other similarly related curves. In the case of $X_0(N)$ for non-squarefree N, the formulas become a bit more complicated (and especially so when N is divisible by 4 or 9). As we shall see later, this is indicative of the fact that the *integral geometry* of $X_0(N)$ is completely understood when N is squarefree, whereas the case of general N is a topic of contemporary research interest.

Exercise 3:
a) Suppose p is prime. Show that the $PSL_2(\mathbb{F}_p)$-Galois cover $X(p) \to X(1)$ is ramified over $J = 1728$ with ramification index 2, over $J = 0$ with r. index 3, and over $J = \infty$ with index p.
b) Using the Riemann-Hurwitz formula, show that

$$g(X(p)) = 1 + \frac{(p^2-1)(p-6)}{24}.$$

In particular, $X(p)$ has genus zero $\iff$ $p \leq 5$.
c) (Not so easy) Show[8] that the genus 3 curve $X(7)$ is nothing else than Klein's quartic curve, defined e.g. by the equation

$$X^3Y + Y^3Z + Z^3X = 0.$$

3. Studying the modular curves over $\mathbb{C}$

Complex geometers tend to study properties of "the generic curve" of genus g rather than specific curves of genus g. Although as arithmetic geometers we care most about the additional arithmetic structure on modular curves to be discussed in the next section, it is worth pointing out that modular curves are *already* of deep interest as compact Riemann surfaces. Here are some sample results and conjectures:

Definition: The **gonality** of a complex (projective, nonsingular, connected) algebraic curve is the least degree of a morphism to $\mathbb{P}^1$. It is easy to see, using the canonical divisor, that (if $g(C) > 1$), the gonality is at most $2g(C) - 2$.[9]

Theorem 2. *(Abramovich) Let $\Gamma \subset PSL_2(\mathbb{Z})$ be a congruence subgroup. Then the gonality of $X(\Gamma)$ is at least $\frac{7}{800}[PSL_2(\mathbb{Z}) : \Gamma]$.*

Remark: Note that by Exercise 2 (and the fact that there are only finitely many subgroups of $\Gamma(1)$ of any given finite index) this result is false without the "congruence" hypothesis.

It is remarkable that the proof, although rather short, uses deep results from differential geometry (due to Li and Yau) and spectral theory (Selberg's celebrated bound $\lambda_1 \geq \frac{3}{16}$ on the smallest positive eigenvalue of the hyperbolic Laplacian).

In the case of $X_0(N)$, this gives an linear lower bound on the gonality in terms

[8]For a solution, see Noam Elkies's beautiful survey article on the arithmetic geometry of the Klein quartic.

[9]Actually this holds for the k-gonality for curves over any field k; in the special case $k = \mathbb{C}$, one can (I am told) use Brill-Noether theory to save a factor of 4.

of the genus. If one is content with a $g^{\frac{1}{2}}$-type lower bound, then there is an argument due to Ogg which stays within the realm of arithmetic geometry.[10]

Since the generic curve of genus g will have gonality approximately equal to g, this shows that modular curves *are* "typical" in this geometric sense.

However, there is another sense in which they are not typical. Given a (discrete) family C_i of compact Riemann surfaces, one can ask which elliptic curves E are covered by this family, i.e., for which there exists for some i a finite map $C_i \to E$. Note that this is equivalent to E occurring up to isogeny as a direct factor of the Jacobian $J(C_i)$. In particular, for any fixed C_i, only countably many isomorphism classes of elliptic curves are covered by C_i, so assuming our family is itself countable, in all only countably many elliptic curves could be covered by the family (whereas there are uncountably many elliptic curves in all, as we shall shortly see). We might as well assume that each C_i has genus at least 2; then, it seems reasonable to expect that *generically* they do not cover any elliptic curves at all. (As far as I know this is purely geometric problem – concerning how certain loci of principally polarized abelian varieties with extra endomorphisms intersect the Torelli, or Jacobian, locus – is open.)

On the other hand the elliptic modularity theorem (Wiles-Breuil-Conrad-Diamond-Taylor) asserts that at least every elliptic curve which can be defined over $\mathbb{Q}$ is covered by some $X_0(N)$. Are there other elliptic curves covered by $X_0(N)$ (for suitable N)? Yes indeed: there is a conjectural characterization due to Ribet, which would follow from Serre's conjecture (i.e., it looks to be pretty close to a theorem).

Finally, here is another interesting question with a "classical" geometric flavor:

Question 1. *Where are the Weierstrass points on $X_0(N)$?*

In the 1960's, Oliver Atkin asked whether, for any squarefree N, a cusp can be a Weierstrass point of $X_0(N)$.[11] Atkin showed that this does not happen when $X_0(N)$ has genus 0 and Ogg extended this to the case of $X_0(pN)$ where $(p, N) = 1$ and $X_0(N)$ has genus zero (i.e., infinitely many cases), as well as the case of $X_0(11p)$. (To make things even more intriguing, Atkin showed that when N is sufficiently far from being squarefree, there are cuspidal Weierstrass points on $X_0(N)$.) There are closely related recent results of Ahlgren-Ono and El-Guindy about mod p reductions of Weierstrass points. But if they are not cusps, then (as we shall see presently) the Weierstrass points correspond to *specific* elliptic curves. Which ones? What are their fields of moduli? There are, as far as I know, only two papers addressing these questions, by Silverman and Burnol. Notwithstanding the interesting mathematics that these two papers contain, the subject remains essentially wide open.

[10]Note mostly to myself: I have a paper in which I use this result for $X_1(N)$ as well as $X_0(N)$ – or, in fact for the Shimura curve analogues thereof – and cited Abramovich's result rather than Ogg's for this reason. A while ago, Matt Baker indicated to me that Ogg's arguments can be made to work for $X_1(N)$ (and possibly all congruence subgroups?) as well; at some point I should try to write this up.

[11]Computations of W.A. Stein confirm that for squarefree N at most 2000 or so cusps are never Weierstrass points, so it seems reasonable to upgrade "Atkin's Question" to "Atkin's Conjecture." A proof seems to be totally out of current reach.

4. Connections with elliptic curves

In fact $Y(1) = PSL_2(\mathbb{Z})\backslash\mathcal{H}$ naturally parameterizes isomorphism classes of elliptic curves and congruence coverings parameterize isomorphism classes of elliptic curves together with some extra torsion structure.

Let us at least explain the first statement. Fix $\tau \in \mathcal{H}$. Then we can define an elliptic curve E_τ which is the quotient of $\mathbb{C}$ by the lattice $\Lambda_\tau = \mathbb{Z}1 \oplus \mathbb{Z}\tau$. We don't get all lattices in this way, but that's okay, because two lattices $\Lambda_1,\ \Lambda_2$ will uniformize isomorphic elliptic curves iff they are **homothetic**, i.e., if there exists $\alpha \in \mathbb{C}$ such that $\alpha\Lambda_1 = \Lambda_2$. Given any basis $\tau_1,\ tau_2$ for a lattice Λ in $\mathbb{C}$ – i.e., τ_1 and τ_2 are two elements of $\mathbb{C}$ which are $\mathbb{R}$-linearly independent – we can uniquely rescale to make $\tau_1 = 1$ (i.e., divide by τ_1!) and then $\tau := \tau_2/\tau_1 \in \mathbb{C} \setminus \mathbb{R}$. τ is then not necessarily in $\mathcal{H}$; the extra (necessary and sufficient) condition here is that the original basis $\tau_1,\ \tau_2$ be *positively oriented*, a notion which homothety does not disturb. Of course any lattice has a positively oriented basis: if the given basis is not positively oriented, interchange the basis elements.

This shows that the map $\tau \mapsto E_\tau$ is surjective onto all elliptic curves. More precisely, it gives an isomorphism between $\mathcal{H}$ and the set of all homothety classes of positively oriented lattice bases. The right hand side has "too much structure" – we would rather just have each homothety class appearing once. But we have natural $SL_2(\mathbb{Z})$ actions on both sides: on the left by linear fractional transformations and on the right since $SL_2(\mathbb{Z})$ acts simply transitively on the set of positively oriented bases for a given lattice. Working this out a bit, one deduces:

Proposition 3. *The assignment $\tau \mapsto \Lambda_\tau$ descends to an isomorphism from $Y(1) = PSL_2(\mathbb{Z})\backslash\mathcal{H}$ to the set of all isomorphism classes of complex elliptic curves.*

Thus $Y(1)$ is, at least in some naive sense, a moduli space of elliptic curves.[12] If we mod out by some subgroup of $\Gamma(1)$, we are then obtaining some structure which is intermediate between the choice of a positively oriented basis for Λ_τ and just the homothety class of Λ_τ (i.e., the elliptic curve itself). When Γ is one of the congruence subgroups, this extra structure can be nicely interpreted. We will just give the answers:

I. $Y(N) = \Gamma(N)\backslash\mathcal{H}$ parameterizes elliptic curves E together with an isomorphism

$$E[N] \cong \Lambda_\tau[N] = \frac{1/N\mathbb{Z}1 \oplus 1/N\mathbb{Z}\tau}{\mathbb{Z}1 \oplus \mathbb{Z}\tau}.$$

Such an isomorphism is called a **full level N structure** on E.

II. $Y_1(N) = \Gamma_1(N)\backslash\mathcal{H}$ parameterizes isomorphisms $(E, P) \mapsto (\Lambda_\tau, \frac{1}{N} \cdot 1)$, or, informally, elliptic curves together with a distinguished point of exact order N.

III. $Y_0(N) = \Gamma_0(N)\backslash\mathcal{H}$ parameterizes isomorphisms $(E, C_N) \mapsto (\Lambda_\tau, \langle\frac{1}{N} \cdot 1\rangle)$, or, informally, elliptic curves together with a distinguished order N cyclic subgroup.

[12] It is not within the scope of these lectures to carefully define a (coarse) moduli space. Indeed, in a department rife with algebraic geometers, there are plenty of people better qualified than I to explain these ideas.

One can consider intermediate and composite level structures.

In fact, there is nothing formally to stop us from making the following generalization: given any subgroup $\Gamma \subset SL_2(\mathbb{Z})$, $Y(\Gamma)$ parameterizes Γ-orbits of isomorphisms from a positively oriented basis of a uniformizing lattice for E to Λ_τ. Thus, as described earlier, from Belyi's theorem it follows that every algebraic curve over $\overline{\mathbb{Q}}$ is technically a moduli space of elliptic curves.[13]

5. Canonical models

We would like to study the modular curves arithmetically; in particular, at least for $X_0(N)$ and $X_1(N)$ we would like to know them as algebraic curves over $\overline{\mathbb{Q}}$ (rather than over $\mathbb{C}$, or, by "the obvious direction of Belyi's theorem," over $\overline{\mathbb{Q}}$).

There are many ways of defining canonical $\mathbb{Q}$-models on $X_0(N)$ and $X_1(N)$ (all compatible with each other). The most classical methods exploit one or both of the following properties of the modular curves $Y_\bullet(N)$: (i) they are non-compact, i.e., there are always cusps; (ii) they all cover the affine line, namely $Y(1)$. Let us briefly recall some of these methods:

1) (q-expansion principle): For a Fuchsian group Γ, define a Γ-modular function $f : \mathcal{H} \to \mathbb{C}$ to be a function which is invariant under Γ – so they descend to $Y(\Gamma)$ – and are meromorphic at the cusps – so they extend to functions $X(\Gamma) \to \mathbb{P}^1$. Thus we get a complex-analytic description of the field of functions $\mathbb{C}(Y(\Gamma))$ (which, recall, is enough to determine the compactified curve up to isomorphism). If Γ is a congruence subgroup of $PSL_2(\mathbb{Z})$ then it contains $z \mapsto z + N$ for some N, and this periodicicity leads to a Fourier series expansion relative to each of the cusps. Consider the subfield $\mathbb{Q}_\Gamma$ of modular functions whose Fourier series expansions at each cusp have $\mathbb{Q}$-coefficients, i.e., are elements of $\mathbb{Q}((q))$. Then, if $\Gamma = \Gamma_0(N)$ or $\Gamma_1(N)$, one can show that $\mathbb{Q}_\Gamma$, which is evidently a subfield of $\mathbb{C}(X(\Gamma))$, is a regular extension of $\mathbb{Q}$ of transcendence degree 1, so defines a $\mathbb{Q}$-rational model.

Advantages: This generalizes nicely to Hilbert and Siegel modular forms, and is an important part of the theory of p-adic modular forms.

Disadvantages: If Γ has no cusps, the rationality condition is vacuous and thus $Q_\Gamma = C(X(\Gamma))$: no good.

2) (explicit functions approach): Recall that we chose a special function J on $X(1) = \mathbb{C}(t)$ by normalizing its value at the three ramification points of the uniformization map $\mathcal{H} \to Y(1)$. Thus we can certainly put a rational model on $Y(1)$ corresponding to the function field $\mathbb{Q}(J)$.

Begin digression of David Foster Wallacian proportions We could have taken any nonconstant function on $Y(1)$. Why was the J-function well-chosen? Because for $\tau \in \mathcal{H}$, $J(\tau) = J(E_\tau) = j(E_\tau)$ generates the field of moduli of the

[13] Again, when Γ is a non-congruence subgroup it seems unlikely that this is of any use, but I don't *guarantee* the uselessness of this approach (whatever that might mean!).

elliptic curve E_τ. Note that this is true for J iff it is true for any function J which we normalize to take values in $\mathbb{P}^1(\mathbb{Q})$ at the three ramification points. Why did we make the specific choices $J(e^{pii/4}) = 1728$, $J(e^{2\pi i/6}) = 0$? (Or, if you like, why not $3J$ or $\frac{17J}{J+5}$?) Because then the J function not only has good rationality properties, it has good *integrality* properties: (e.g. if $j(E) \in \mathbb{Q}$, then E has potentially good reduction at p iff $j(E) \in \mathbb{Z}_p$), so is thus the correct choice for an *integral* canonical model of $X(1)$.

The importance of finding the "integrally correct" normalization of the coordinate function on a genus zero modular curve is a topic that the late 19th century number theorists (notably Weber, Fricke, Klein) were well aware of (probably better aware of than most present-day number theorists), a favorable consequence of the more down-to-earth approach to mathematical objects that was the spirit of the day. As far as I know, the classical theorists did indeed find such a *distinguished* generator of all genus zero modular function fields: such a thing is called a *Hauptmodul*. In some sense, the modern "explanation" for what a Hauptmodul is – namely, a choice of *integral* canonical model well before the notions of mod p reduction of curves were considered.[14] – is anachronistic and limited. Rather, if you actually work *explicitly* with modular curves, then certain nice functions naturally present themselves to you, and working with the "naturally nice" functions will turn out to have advantages beyond any particular axiomatic framework. For instance, the Moonshine phenomenon began when John Thompson noticed a simple arithmetic relationship between the Fourier coefficients of the J-function and the dimensions of the irreducible representations of the Monster (i.e., the largest sporadic simple group). One of the truly puzzling things about Moonshine is that it is a remarkable, but finite, coincidence: there are, of course, only finitely many irreducible representations of the monster which we are comparing to the first finitely many coefficients of J: how do you tell if it is all just some ridiculous concidence? What you can do is look for similar structure among Fourier coefficients of other Hauptmoduls, as did Conway, Norton and McKay, and they found them![15] In some sense, then, although as far as rational canonical models are concerned, $\frac{17J}{J+5}$ is as good as J, the Monster prefers J, and similarly for finitely many other genus zero modular function fields. It is all a bit spooky, and (as I have heard Conway himself remark, in paraphrase) although Borcherds' work is spectacular mathematics, it leaves most of the spookiness of the situation intact.

End D.F.W.-style digression.

Anyway, if $J(\tau)$ is the j-invariant of E_τ, then $J(N\tau)$ is the j-invariant of $E_{n\tau}$, which is canonically cyclically N-isogenous to E. (Check for yourself.) Thinking about this a bit, it is not too hard to see that the function $J_N : \tau \mapsto J(N\tau)$ is a $\Gamma_0(N)$-modular function, and that $\mathbb{C}(J, J_N) = \mathbb{C}(Y_0(N))$, hence it makes sense to define $\mathbb{Q}_{\Gamma_0(N)}$ as $\mathbb{Q}(J, J_N)$ and not so hard to show that this gives a $\mathbb{Q}$-model of

[14] I am no historian, but I would be shocked if Weber, Fricke and Klein were, in any reasonable sense, interested in characteristic p.

[15] It was left to Richard Borcherds to actually find the mathematics linking these objects, using the intermediary of vertex operator algebras (which I will not pretend to understand).

$\mathbb{C}(Y_0(N))$. This is a sort of maximally explicit approach: there is an actual polynomial $\Phi_N[X,Y] \in \mathbb{Z}[X,Y]$ out there which gives the algebraic relation between J and J_N, and

$$\Phi_N[X,Y] = 0$$

is an explicit equation for $X_0(N)$.

Advantages: This is a maximally elementary approach, understandable if you know only the definitions of elliptic curve, j-invariant and modular function. We have not really used (i) (the existence of cusps): replacing J by any element of $\mathbb{Q}(J)$ would not change the function field generated by the modular polynomial.

Disadvantages: We did, of course use (ii), the existence of a mapping down to a genus zero curve. Moreover actually computing the modular polynomial is incredibly time-consuming (I don't have precise references here, but unless you are a professional computational number theorist showing off your clever new approach, even something like $N = 101$ is hopeless), and the model given by $\Phi_N[X,Y] = 0$ will have horrible singularities.[16]

(3) Galois Theory approach: take E to be an elliptic curve over $\mathbb{Q}(J)$ with j-invariant J, i.e., a so-called *generic* elliptic curve. Let $K_N/\mathbb{Q}(J)$ be the field extension cut out by trivializing the N-torsion on E. One can show that $K_N/\mathbb{Q}(J)$ is Galois with group $GL_2(\mathbb{Z}/N\mathbb{Z})$. Now let us work with congruence subgroups

$$K_N \subset \tilde{G} \subset GL_2(\mathbb{Z})$$

where K_N is the kernel of the reduction map $GL_2(\mathbb{Z}) \to GL_2(\mathbb{Z}/N\mathbb{Z})$. Such subgroup $\tilde{G}$ clearly correspond to subgroups $G \subset GL_2(\mathbb{Z}/N\mathbb{Z})$. We can define $\mathbb{Q}(\tilde{G}) = K_N^{\pm G}$.

Advantages: this approach shows us why we have been staying away from the $\mathbb{Q}$-canonical model of $X(N)$: namely, the full extension $K_N/\mathbb{Q}(J)$ contains the constant extension $\mathbb{Q}(\zeta_N)$. (In general, the theory of the Weil pairing implies that to trivialize the N-torsion on an elliptic curve in characteristic 0, one necessarily also trivializes the N-torsion in the multiplicative group, i.e., all N-torsion fields contain a primitive Nth root of unity, say ζ_N.) Thus we get a canonical model for $X(N)$ over $\mathbb{Q}(\zeta_N)$. (In general, we say that a modular curve Γ has *level* N if N is the least positive integer M for which $\Gamma \subset \Gamma(N)$. Then the argument shows that any modular curve of level N is defined over $\mathbb{Q}(\zeta_N)$.) Looking more carefully, one can see that the constant extension of $\mathbb{Q}$ contained in $K_N^G(\Gamma)$ is cut out by the image of the determinant map $\det : GL_2(\mathbb{Z}/N\mathbb{Z}) \to (\mathbb{Z}/N\mathbb{Z})^\times$ restricted to G, and this map is surjective for $\tilde{G} = \tilde{G}_0(N)$ (defined in the same way as $\Gamma_0(N)$, i.e., matrices which are upper triangular modulo N) and $\Gamma_1(N)$ (defined as matrices in $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with $a \cong 1 \pmod N, c \cong 0 \pmod N$).

Overall this is a quite enlightening approach, and I recommend reading Rohrlich's article, which takes this tack to develop much of the basic theory of modular curves.

[16]On the other hand, there is enough literature on computing modular polynomials that there must be *some* other advantage that I am not aware of.

Disadvantages: We are still assuming (ii), that we have a $\mathbb{P}^1$ at the bottom of our tower.

(4) Moduli space approach: We have been working complex-analytically with elliptic curves over $\mathbb{C}$ but there is a perfectly well-defined moduli problem of elliptic curves over $\mathbb{Q}$, whose coarse moduli space gives rise to the canonical model $Y(1)_{\mathbb{Q}}$, which is indeed just the affine line over $\mathbb{Q}$. Moreover, the moduli problem of isomorphism classes of elliptic curves over $\mathbb{Q}$-schemes together with a distinguished point of order N (i.e., an isomorphism $(E, P) \to (E', P')$ is an isomorphism of elliptic curves (over some base $\mathbb{Q}$-scheme) which carries the point $P \mapsto P'$) gives rise to the curves $Y_1(N)$ and $X_1(N)_{/\mathbb{Q}}$. (Having exact order N is a condition that, by definition if you like, we check on geometric fibers.) Finally, the moduli problem of an elliptic curve together with a cyclic, order N subgroup scheme (i.e., a finite flat subgroup scheme each of whose geometric fibers yields a cyclic order N subgroup) gives rise to $Y_0(N)$ and $X_0(N)$ over $\mathbb{Q}$.

Advantages: This will work for Shimura curves over $\mathbb{Q}$, once we define the analogous moduli problem. Moreover it works nicely for Hilbert moduli space, Siegel moduli space (and in general for PEL-type Shimura varieties), and is in general such an important technique in arithmetic and algebraic geometry that one needs to know about it no matter what. Perhaps most importantly, it points the way to a similar approach to *integral canonical models*.

Advantage/Disadvantage?: The uniqueness of the model is solved immediately; it is then tempting to slough over the existence problem as being relatively abstract nonsense. The latter cannot really be the case: consider, for instance that the moduli problem makes perfect sense even for non-congruence subgroups, so at some point in the proof of the existence of a coarse moduli scheme over $\mathbb{Q}$, we must be making nontrivial use of the congruence subgroup property. (It is not especially mysterious how this goes – the strategy will be (i) to show that $X(N)$ for $N \geq 3$ is a fine moduli space and then (ii) use descent theory to solve the corresponding *relative* moduli problem – but I don't want to enter into the details here.) But in fact this may be an advantage: if, as a student of modular curves, you are looking for corners to cut in order to learn the material, then wondering about the details of the existence of an object guaranteed to be unique by general principles and which has been studied by many people over the years, then perhaps the proof that such canonically defined objects do indeed exist is a relatively safe corner to cut!

Disadvantage: The method still does not work for the most general quaternionic Shimura curves, which are in fact not moduli spaces, at least not over their ground field. The truly satisfactory theory of rational canonical models of Shimura varieties – as laid down by Shimura himself and clarified and extended by Deligne – uses a yet more complicated approach, essentially using the explicit reciprocity law of Shimura-Taniyama for abelian varieties with complex multiplication to predict what the field of definition of each of a certain Zariski-dense countable subset of "special points" should be and then showing that there exists a unique model over an appropriate number field giving the correct field of moduli at each of the special

points.

How important is this special point construction, both technically in the definition of Shimura varieties (including, to be sure, certain Shimura curves that are rather closely related to the ones we shall discuss) and philosophically in understanding what's going on in things like the André-Oort Conjecture, Heegner point constructions on elliptic curves, work of Mazur, Wiles, Rubin, Bertolini, Darmon, Vatsal, Cornut, Dasgupta.....? It is of the highest level of importance. But it is technically demanding enough to be ever so far beyond the scope of these notes.

6. Some enormous theorems

We are now well entitled to consider $X_0(N)$ and $X_1(N)$ as curves defined over $\mathbb{Q}$. It would be silly to spend so many pages describing the rational canonical models of these curves without reviewing at least some of the spectacular results concerning these models.

First, a (relatively elementary) key observation:

Fact 1. *For any N, there is at least one $\mathbb{Q}$-rational cusp on $X_1(N)$, and hence also on $X_0(N)$.*

We have not said enough about the cusps in order to get into the proof. But morally, a theme in algebraic geometry is that most naturally occurring moduli spaces are not compact, and in order to compactify them we have to enlarge our moduli problem by considering degenerate, or (hopefully mildly) singular objects. This is possible in our case: we can consider the cuspidal points of $X_1(N)$ as parameterizing "Néron N-gons" and from this moduli interpretation it is easy to find that there is a distinguished cusp (still called ∞) which is $\mathbb{Q}$-rational.[17]

Now for the pyrotechnics:

Theorem 4. *(Mazur) For any positive integer N, the following are equivalent:*
a) $X_1(N) \cong \mathbb{P}^1$.
a′) $X_1(N)$ *has genus zero.*
b) $N \leq 10$ *or* $N = 12$.
c) $Y_1(N)(\mathbb{Q})$ *is nonempty.*

Remark: The equivalence of a) and a′) follows from the existence of $\mathbb{Q}$-rational cusps; the equivalence of these conditions with b) follows from known genus formulas for $X_1(N)$.[18] That these conditions imply c) is obvious, so the meat (and there is about a hundred pages worth of meat) is in c) $\implies$ b).

Remark: Mazur proved this theorem in 1974. It had long been part of the folklore (for a while it was called Ogg's conjecture, but it was later pointed out that it went back at least as far as Beppo-Levi).

[17] More elementary proofs are certainly possible.

[18] Although we did not these write down explicitly, they are not too hard to derive; easier still is to see that $g(X_1(N))$ is approximately quadratic in N and to give an effective lower bound.

Theorem 4 leads rather easily to a characterization of all possible torsion subgroups of rational elliptic curves: namely, cyclic of order N for the above permitted values of N or $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\nu\mathbb{Z}$ for $1 \leq \nu \leq 4$.

Remark: One can seek to generalize this result by studying points on $X_1(N)$ over number fields of higher degree. Work of Kamienny, Abramovich, Merel and Parent culminated in a proof of the *uniform boundedness conjecture*: there exists a function $F(d)$ such that if $K/\mathbb{Q}$ is any number field of degree d, then the order of the torsion subgroup of any elliptic curve $E_{/K}$ is at most $F(d)$.

Exercise XX: Show that the uniform boundedness of the size of the entire torsion subgroup in $[K : \mathbb{Q}]$ is equivalent to the finiteness of the set of all N such that $Y_1(N)(K) \neq \emptyset$ for all number fields $K$of any given degree.

The bounds of Merel and Parent are exponential in d, which is presumably far from the truth, so the problem of finding the *optimal* bound $F(d)$ remains wide open.[19]

Yet more impressive is the following:

Theorem 5. *(Mazur) Suppose that $N > 163$ is prime. Then $Y_0(N)(\mathbb{Q}) = \emptyset$. More precisely, $Y_0(N)(\mathbb{Q}) \neq \emptyset$ iff any one of the following holds:*
a) $Y_0(N)$ has genus 0.
b) $\mathbb{Q}(\sqrt{-N})$ has class number 1.
c) $N = 17$ (genus one) or $N = 37$ (genus two).

Remark: To discuss the "easy" half of Mazur's theorem: if a) holds, then $X_0(N)$ has genus 0 and a rational cusp so is $\mathbb{P}^1$ and hence has infinitely many rational points, only finitely many of which are cusps. Part b) follows easily from the theory of complex multiplication. The case of $N = 17$ is not too surprising, since $X_0(17)$ is an elliptic curve, and current beliefs are that the "chance" that an elliptic curve has infinitely many rational points is $\frac{1}{2}$, whereas rational points on curves of higher genus are thought to be significantly more rare. (On the other hand, the Mordell-Weil group of the elliptic curve $X_0(17)_{/\mathbb{Q}}$ is not infinite; it is $\mathbb{Z}/4\mathbb{Z}$, which means there are precisely two noncuspidal rational points.) The case of $N = 37$ is a famous anomaly which had earlier been studied in a paper of Mazur and Swinnerton-Dyer.

Theorem 6. *(Elliptic Modularity Theorem) For every elliptic curve $E_{/\mathbb{Q}}$, there exists, for some N, a finite morphism $X_0(N) \to E$.*

Remark: This is a somewhat more precise form of the result alluded to in §X. One can be even more precise: there exists a finite morhpism $X_0(N) \to E$ iff N is a multiple of the conductor $N(E)$ of E, an integer which is (almost) characterized by the following: for any prime $p \geq 5$, $N(E)$ is exactly divisible by p^0 if E has good reduction at p, by p if E has semistable (aka multiplicative) bad reduction at p, and by p^2 if E has nonsemistable (aka additive) bad reduction at p. (Things are more complicated at 2 and 3.)

Following a complaint/suggestion of Dino Lorenzini, let us discuss a bit of the

[19]Some remarks on this are made in [**?**] and [**?**].

history of this all-important theorem. To every elliptic curve $E_{/\mathbb{Q}}$ one can associate its *L-series* $L(E,s)$, a certain meromorphic function. In the first half of the twentieth century Hecke was able to identify the L-series of an elliptic curve with *complex multiplication* as the L-series obtained from a character of the maximal abelian extension of the CM field K (a "Grossencharacter"). This showed in particular that $L(E,s)$, a priori defined for $s > 1$ by an Euler product, had much nicer analytic properties: in particular, it extends to an entire function on the complex plane and satisfies a certain functional equation.

In the 1950's Eichler and Shimura made the following construction: given a weight two modular (new)form f on $\Gamma_0(N)$ with $\mathbb{Q}$-rational Fourier coefficients, they defined an elliptic curve $E(f)_{/\mathbb{Q}}$ in such a way that $L(E(f),s) = L(f,s)$, where the latter is the natural Dirichlet series $\sum_{n=1}^{\infty} \frac{a_n}{n^s}$ associated to the modular form $\sum_{n=1}^{\infty} a_n q^n$. This elliptic curve is constructed *a priori* as a quotient of the Jacobian of $X_0(N)$, so it admits a finite morphism from $X_0(N)$. (This construction generalizes Hecke's construction in a way that we will not attempt to discuss.) Conversely, it is not so hard to see that any elliptic curve E which is the target of a morphism $X_0(N) \to E$ is obtained via the Eichler-Shimura construction.

It is therefore natural to ask for the *image* of the Eichler-Shimura construction, i.e., which rational elliptic curves arise in this way? Apparently the bold suggestion that *all* rational elliptic curves should arise in this way is due to Yutaka Taniyama. Taniyama's original thoughts in this direction were vaguer (and the first precision of them turned out to be incorrect), and the precise form of the conjecture was, I believe, developed in collaboration with Shimura. In its early days the conjecture met with great skepticism, which seems a little odd from our modern eyes. (Given $E_{/\mathbb{Q}}$, to disprove the strong form of the conjecture, namely that for any $E_{/\mathbb{Q}}$ there exists a weight 2 newform f on $\Gamma_0(N(E))$ such that $L(f,s) = L(E,s)$, would be a matter of finite computation. Similarly, given any $E_{/\mathbb{Q}}$, one can compute $N(E)$ and show that there is at most one newform f of level $N(E)$ whose coefficients match the L-series coefficients of E. One cannot prove the modularity this way, but one can match the a_p's for as many primes p as one likes, so where are the grounds for disbelief? Of course, such computations were far from routine back then.)

One of the controversies in this subject is whether and/or to what extent Weil's name should be associated with this conjecture. I was not around at the time so it's not for me to try to sort out the history of this, but let me instead point out a very important *theorem* of Weil's. Remember that one of the reasons that the modularity conjecture is so exciting is that it implies that the L-function $L(E,s)$ has desirable analytic properties. What Weil showed is that if you instead postulate the $L(E,s)$, together with all (or sufficiently many) of its twists by Dirichlet characters, has these properties – namely analytic continuation and a precise functional equation – then one can *deduce* the modularity of E. This theorem ("Weil's converse theorem") is important enough in the subject (and in particular, gives strong philosophical evidence for the conjecture, whether Weil thought so or not!) that I find it appropriate to speak of the *Taniyama-Shimura-Weil Conjecture*.

Of course the T-S-W conjecture was proved in the semistable case (squarefree conductor) by Andrew Wiles with the assistance of Richard Taylor in 1995, and the full theorem was proven by Christophe Breuil, Brian Conrad, Fred Diamond and Taylor in 1999. Wiles' work was undoubtedly *un des plus grand tours de force* in mathematical history, but like any modern mathematician his work relies on that of many others, in particular that of Mazur and Langlands-Tunnell.

7. Transition to Shimura curves

Why is the modularity theorem so important for the study of the *arithmetic* (rather than merely than analytic number theory) of rational elliptic curves? Given any elliptic curve $E_{/\mathbb{Q}}$, there is a map

$$X_0(N) \to E$$

defined over $\mathbb{Q}$, so that we can do something remarkable: study a fixed elliptic curve using a family of all elliptic curves. At first it might seem that Theorem XX limits the usefulness of this: if $N > 163$, there are no-noncuspidal rational points on $X_0(N)$, and moreover one can show that cusps always map to torsion points on E.

However, using the fact that E is an algebraic group, we can employ the following trick: let $P \in X_0(N)(L)$ be a point defined over any number field L. Then $\varphi(P) \in E(L)$. However, if $\sigma_i : L \hookrightarrow \overline{\mathbb{Q}}$ are all the embeddings, then $\sum_i \sigma_i(P) \in E(\mathbb{Q})$. In other words, even irrational points on $X_0(N)$ produce rational points on E!

This is however a bewildering array of riches: if we take a breath and remember that $E(\mathbb{Q})$ is finitely generated, then it must be that for most L and points $P \in E(L)$, we are getting the same points on E, or multiples thereof. Where to look for "special points"? I wish I could remember the reference, but there is an old paper of Birch in which this question occurs non-rhetorically. He mentions the family of points on curves which algebraic geometers regard as special, namely Weierstrass points. Indeed, one can certainly ask the following

Question 2. *Let $\varphi : X_0(N(E)) \to E$ be an (optimal) modular parameterization of a rational elliptic curve. Let $W \subset E(\mathbb{Q})$ be the subgroup generated by the images of the traces of Weierstrass points on $X_0(N(E))$. What can be said about W?*

This is a perfectly interesting question. It is however one that we seem not to be ready to answer, because as mentioned earlier we don't really know "where the Weierstrass points are" on $X_0(N)$, nor anything about their arithmetic properties.

Better would be a supply of points whose arithmetic we understand. Now comes what is (to my mind) the single most beautiful idea in modern elliptic curve theory: in so many ways elliptic curves with complex multiplication are much easier to understand. Let us then, via modular parameterizations, study *all* rational elliptic curves using these easiest elliptic curves.[20]

[20]To make a Langlands-style slogan out of it, let us study GL_2-type arithmetic objects using $GL_1 = \mathbb{G}_m$-type arithmetic objects.

It turns out that the particular class of CM points on $X_0(N)$ we want to study is the following: we want pairs $E \to E'$ of cyclically N-isogenous elliptic curves so that $\mathrm{End}(E)$ and $\mathrm{End}(E')$ have CM by the *same* order in an imaginary quadratic field K. Without getting too bogged down into the particulars of CM theory, suppose we want both to have CM by the maximal order: then $E' = E/\mathcal{I}$, where $\mathcal{I}$ is an ideal of $\mathfrak{o}_K$ of norm N. The condition then is that every prime dividing N should *split* in K, the so-called *Heegner condition*. For any such K, we can consider the images of Heegner points on $E(K)$. The details are of course beyond the scope of this survey, but suffice it to say that, thanks to beautiful work of Gross-Zagier and Kolyvagin, we get a highly interesting and useful supply of rational points on $E(K)$ in this way.

But what about $E(\mathbb{Q})$? Well, given any quadratic field, we get modulo 2-torsion, a decomposition of $E(K)$ into the direct sum of $E(\mathbb{Q})$ and $E^{\chi_K}(\mathbb{Q})$, so the arithmetic over K and over $\mathbb{Q}$ is closely related. The original perspective on this was that K and $E(K)$ were auxiliary data to study $E(\mathbb{Q})$. The main theorem that comes out of all this is that if the analytic rank of $E_{/\mathbb{Q}}$ is at most 1, then the rank is equal to the analytic rank, and to prove this one can always choose a suitable K satisfying the Heegner hypothesis.

On the other hand we would also like to understand the arithmetic of E not just over $\mathbb{Q}$ but over all number fields and in particular over all imaginary quadratic fields. What if we are interested in an imaginary quadratic field K such that some primes of bad reduction do not split in K? The modern answer is that, at least when $N(E)$ is squarefree, we consider not just the parameterization $X_0(N(E)) \to E$, but for all factorizations $N(E) = N \cdot D$, the Shimura curve parameterization

$$X_0^D(N) \to E.$$

This is the "mainstream" motivation for studying Shimura curves.

However, just as one can be interested in modular curves in their own right and not just as a supply of rational points on elliptic curves, one can be interested in intrinsic the Diophantine geometry of Shimura curves. In fact, as we shall see, there is one respect in which the Shimura curves are *more* interesting than the classical modular curves: their lack of cusps leads to to the possibility that they may fail to have points rational over some given number field K and/or over all completions K_v of K.

8. Introduction to Shimura curves

Let us go back to Fuchsian groups. In some naive sense, the subgroups of the modular group $PSL_2(\mathbb{Z})$ are the easiest Fuchsian groups (of the first kind) to construct. There are evidently many more, because every compact Riemann surface of genus at least 2 is uniformized by a Fuchsian group *of hyperbolic type* (i.e., with a compact fundamental domain and without elements of finite order). Nevertheless it is not so easy to get our hands on them, e.g. to write down generators. For instance, if you give me a hyperelliptic curve $y^2 = P(x)$ for $P \in \mathbb{Q}[x]$ a polynomial of degree at least 5, then the corresponding algebraic curve is uniformized by some unique (up to

$PSL_2(\mathbb{R})$-conjugacy) Fuchsian group of hyperbolic type. Can we write down generators? I don't know how to do it, and especially there is no easy correspondence between the arithmetic properties of the curve and of the uniformizing Fuchsian group.

8.1. **An interesting family of Fuchsian groups.** There is one class of examples of Fuchsian groups which one can write down explicitly, which generalizes $PSL_2(\mathbb{Z})$: namely, choose any elements $a,\ b,\ c$ in $\mathbb{Z}^+ \cup \infty$ such that $\frac{1}{a} + \frac{1}{b} + \frac{1}{c} < 1$ (here our convention is that $\frac{1}{\infty} = 0$, of course). Then, through the magic of hyperbolic geometry, there is up to isometry a unique hyperbolic triangle with angles $\frac{\pi}{a}$, $\frac{\pi}{b}$, $\frac{\pi}{c}$. Here if any of a, b, c are zero then we are getting an "ideal" triangle, namely with angle 0 and with vertex on the boundary of the hyperbolic plane (this is an instance where the unit disc model of the hyperbolic plane is easier to visualize). For instance, if we take $(a, b, c) = (2, 3, \infty)$ then the corresponding hyperbolic triangle bounds the region in $\mathcal{H}$ which is the right half of the standard fundamental region for $PSL_2(\mathbb{Z})$.

Consider the group $\tilde{\Delta}$ of isometries of $\mathcal{H}$ generated by reflections $\tau_1,\ \tau_2,\ \tau_3$ through these three sides. Reflections are orientation-reversing, so these are not elements of $PSL_2(\mathbb{R})$. However, there is an index 2 subgroup of orientation-preserving isometries which is a Fuchsian group: it is generated by $x = \tau_3 \circ \tau_2,\ y = \tau_1 \circ \tau_3,\ z = \tau_2 \circ \tau_1$ and these are respectively, rotations through the three vertices of the triangle of angles $2\pi/a,\ 2\pi/b, 2\pi/c$. We get a very explicit and interesting Fuchsian group $\Delta(a, b, c)$ with presentation

$$\Delta(a, b, c) = \langle x, y, z \mid x^a = y^b = z^c = xyz = 1 \rangle.$$

In particular $\Delta(2, 3, \infty) = PSL_2(\mathbb{Z})$ (up to conjugacy) and $\Delta(a, b, c)$ is cocompact iff a, b, c are all finite.

It is not hard to write down explicit generators for $\Delta(a, b, c)$ whose entries lie in $R = R_{a,b,c} = \mathbb{Z}[2\cos(\pi/a), 2\cos(\pi/b), 2\cos(\pi/c)]$. One can then define congruence subgroups with respect to ideals of R: in particular for every ideal $\mathcal{N}$ one can define $\Gamma_0(\mathcal{N})$, $\Gamma_1(\mathcal{N})$, $\Gamma(\mathcal{N})$.

The corresponding family of algebraic curves is, I believe, a very interesting one, but it is not the family that we are meant to be talking about: this is a generalization of modular curves which is in general different from the Shimura curves: there are precisely 85 triples (a, b, c) such that the curves uniformized by $\Delta(a, b, c)$ are Shimura curves. In particular there is, as far as I know, no good theory of Heegner points on a general such curve, so for elliptic curve applications this is the wrong generalization. On the other hand, they are much easier to compute with than arbitrary Shimura curves; in fact, in the burgeoning theory of Shimura curve computations, the vast majority of work concerns those 85 "Shimura" triples. It is an interesting question how much of the arithmetic theory of modular curves and Shimura curves can be generalized to this family of curves. (In fact I have been working on this on and off for several years.)

8.2. **Shimura curves (over $\mathbb{Q}$).** To get the real thing, then, we need to take a somewhat more abstract perspective. Here is one way to motivate the construction: consider the class of Fuchsian groups whose matrix entries are as close as possible to being integral, without actually being in $\mathbb{Z}$. What could this mean? Well, the

characteristic polynomial of any element of $SL_2(\mathbb{Z})$ clearly has integral coefficients, so let us (just for the sake of exposition) define a $\mathbb{Z}$-*group* to be a Fuchsian group $\Gamma \subset SL_2(\mathbb{Z})$ of the first kind such that for all $\gamma in \Gamma$, the characteristic polynomial of Γ has $\mathbb{Z}$-entries. Since the determinant is always 1, this means precisely that we want integral trace. Now the rational canonical form tells us that any particular element γ of a $\mathbb{Z}$-group can up to to conjugacy be expressed as a matrix with $\mathbb{Z}$-entries. However, we may not be able to *simultaneously* conjugate all elements of Γ into $\mathbb{Z}$. Indeed, here is what we can say about a $\mathbb{Z}$-group:

Proposition 7. *Let Γ be a $\mathbb{Z}$-group. Put $\mathbb{Q}[\Gamma]$ to be the $\mathbb{Q}$-subalgebra of $M_2(\mathbb{R})$ generated by the entries of Γ and similarly put $\mathbb{R}[\Gamma]$ to be the $\mathbb{R}$-subalgebra. Then $\mathbb{R}[\Gamma] = M_2(\mathbb{R})$.*

Since $\mathbb{R}[\Gamma] = \mathbb{Q}[\Gamma] \otimes_{\mathbb{Q}} \mathbb{R}$, this tells us precisely that $\mathbb{Q}[\Gamma]$ is an indefinite rational quaternion algebra. That is, it is a four-dimensional $\mathbb{Q}$-algebra, which is simple (no two-sided ideals), whose center is precisely $\mathbb{Q}$.

Example: $\Gamma = SL_2(\mathbb{Z})$ and $\mathbb{Q}[\Gamma] = M_2(\mathbb{Q})$. This is called the *split* quaternion algebra. Any other rational quaternion algebra is a division algebra.

It is of course, not yet clear that any $\mathbb{Z}$-groups other than $SL_2(\mathbb{Z})$ and its subgroups of finite index exist. But notice that we can express the construction of a Fuchsian group $SL_2(\mathbb{Z})$ from $M_2(\mathbb{Q})$ in a way which generalizes to arbitrary indefinite rational quaternion algebras: namely, to say that a rational quaternion algebra $B_{/\mathbb{Q}}$ is indefinite is to say that there exists an isomorphism $B \otimes_{\mathbb{Q}} \mathbb{R} = M_2(\mathbb{R})$ (equality here means we pick one such isomorphism, and since all such are $SL_2(\mathbb{R})$-conjugate, it doesn't matter which one we pick), and in particular an injection

$$B \hookrightarrow M_2(\mathbb{R}).$$

Now select a maximal $\mathbb{Z}$-order $\mathcal{O}$ of B – that is a sub-$\mathbb{Z}$-algebra of B such that the canonical map $\mathcal{O} \times \mathbb{Q} \to B$ is an isomorphism, and which is not properly contained in any other such order – and take $\Gamma(1)$ to be the units of $\mathcal{O}$ which, under the embedding into $GL_2(\mathbb{R})$, have determinant 1 (as in the case of $SL_2(\mathbb{Z})$, they would have to have determinant, or "reduced norm" ± 1), modulo ± 1. Now the basic fact is:

Proposition 8. *a) $\Gamma(1) \subset PSL_2(\mathbb{R})$ is a Fuchsian group, i.e., is discrete.*
b) Up to $PSL_2(\mathbb{R})$-conjugacy, $\Gamma(1)$ is independent of the choices made.
c) Except in the case where $B = M_2(\mathbb{Q})$, $\Gamma(1)$ is cocompact.

For any quaternion algebra $B_{/\mathbb{Q}}$, we define its discriminant to be the product of the primes p such that $B \otimes \mathbb{Q}_p$ is a division algebra. Then it is a basic (but nontrivial; it is closely related to classfield theory) fact that there is a unique quaternion algebra of any given squarefree discriminant which is *indefinite* if the number of finite ramified primes is even and *definite* (i.e., $B \otimes \mathbb{R}$ is the unique (Hamilton) division quaternion algebra over $\mathbb{R}$). Let us rename $\Gamma(1)$ $\Gamma^D(1)$ where D is the discriminant of the quaternion algebra B. So, for any squarefree D which is divisible by an even number of primes, we get a well-determined complex algebraic curve $Y(\Gamma^D(1))$; if $D = 1$ this is just $Y(1)$ as before, but for $D > 1$ it is cocompact.

We can again define the notion of a congruence subgroup of $\Gamma^D(1)$. Namely, for any

positive integer N, $N\mathcal{O}$ is a 2-sided $\mathcal{O}$-ideal, so we may define $\Gamma^D(N)$ to be those elements of $\Gamma^D(1)$ which are congruent to 1 modulo $N\mathcal{O}$. Although this definition makes sense for arbitrary N, it is both more transparent and more useful in the case when N is prime to the discriminant D, which we will assume from now on. We then find that $\Gamma^D(1)/\Gamma^D(N) \cong PSL_2(\mathbb{Z}/N\mathbb{Z})$ as in the classical case, which means that we can also define $\Gamma_0^D(N)$ – i.e., matrices which modulo N become upper triangular – and $\Gamma_1^D(N)$ – i.e., matrices which modulo N become unipotent. Therefore we may define

$$Y_\bullet^D(N) = \Gamma_\bullet^D(N)\backslash\mathcal{H}$$

and $X_\bullet^D(N)$ to be the compactification of $Y_\bullet^D(N)$ (which, in every case but the classical $D = 1$ case, is the same as $Y_\bullet^D(N)$). These are the *Shimura curves*.

Here is a genus formula for $X_0^D(N)$, for squarefree N, generalizing the case of $D = 1$:

$$g(X_0^D(N)) = 1 + \frac{1}{12}\varphi(D)\psi(N) - \frac{e_1(D,N)}{4} - \frac{e_3(D,N)}{4} - \frac{e_\infty(D,N)}{2},$$

where φ is the Euler ϕ function, as before $\psi(N) = \prod_{p \mid N}(p+1)$, e_∞ is the number of cusps on $X_0^D(N)$ – namely, the number of positive divisors of N if $D = 1$ and 0 if $D > 1$ – and finally, for $m = 1$ or $m = 3$,

$$e_m(D,N) = \prod_{p \mid D} 1 - \left(\frac{-d(m)}{p}\right) \prod_{q \mid N} 1 + \left(\frac{-d(m)}{q}\right).$$

It is again easy to see that the genus tends to infinity with $\max(N, D)$; and in fact, Abramovich's result holds verbatim here and gives a linear lower bound on the gonality in terms of the genus. An important difference is that there is usually not a projective line on the bottom. Indeed,

$$g(X_0^D(N)) = 0 \iff (D,N) = (6,1),\ (6,7),\ (10,1),\ (22,1);$$

$$g(X_0^D(N)) = 1 \iff$$

$$(D,N) = (6,13),\ (10,7),\ (14,1),\ (15,1),\ (21,1),\ (33,1),\ (34,1),\ (46,1).$$

9. Moduli interpretation, canonical models

We would now like to provide a moduli interpretation for the Shimura curves $X_\bullet^D(N)$. Perhaps surprisingly, we will show that these curves are moduli of certain *abelian surfaces*; in fact our interpretation will hold even for $D = 1$, so that we will get a new moduli interpretation of the classical modular curves! (On the other hand, we will be able to see almost immediately how the new interpretation is related to the old interpetation.) Time constraints will cause us to restrict to the case of no level structure – X^D. To understand the moduli interpretation of the level structure for N prime to D, one needs to exploit the $\mathcal{O}$-module structure on $A[N]$ and use (to put a fancy label on a relatively simple piece of linear algebra) "Morita equivalence." These ideas figure prominently in my PhD thesis, which I can give you if you really are interested in the details.

Okay, here goes: recall we have chosen an isomorphism $B \otimes \mathbb{R}$ with $M_2(\mathbb{R})$ (unique up to conjugacy). Earlier we exploited the fact that this yields an embedding of a maximal order (also unique up to conjugacy) into $M_2(\mathbb{R})$. Carrying this one step

further, we also get an embedding of $\mathcal{O}$ into $M_2(\mathbb{C})$. In particular, for any $\tau \in \mathcal{H}$, one can consider the column vector $v_\tau = [\tau 1]^t$ in $\mathbb{C}^2$, and then also the set $\mathcal{O}v_\tau$. A (not completely trivial) computation reveals that this is a full lattice in $\mathbb{C}^2$, so we get a two-dimensional complex torus

$$A_\tau := \mathbb{C}^2/\mathcal{O}v_\tau.$$

Now recall that unlike elliptic curves, the generic complex torus of dimension at least 2 is *not* an abelian variety. The necessary and sufficient condition is the existence of a Riemann form. On the other hand, an important general principle is that one can build a Riemann form out of a sufficiently large ring of endomorphisms ("real multiplication" is enough; here we have more than that). To write down the Riemann form is not so difficult by requires an extra bit of data: indeed, once we define a Riemann form we will get a corresponding positive ("Rosati") involution on the endomorphism algebra, so it turns out that the right thing to do is to specify this involution in advance, which requires a little bit of quaternion arithmetic. (See §0.3.2 of my thesis for a discussion.)

In any case, things are chosen so that A_τ has $\mathcal{O}$ as an endomorphism ring, and moreover positive norm units of $\mathcal{O}$ are just going to change the basis of the lattice. Therefore we get that

$$Y^D(1) = \Gamma^D(1)\backslash\mathcal{H}$$

parameterizes certain abelian surfaces with *quaternionic multiplication*, i.e., an injection

$$\iota : \mathcal{O} \hookrightarrow \mathrm{End}(A).$$

We remark that the choice of ι – called a "QM structure" on A – determines a canonical principal polarization but gives (unless $D = 1$) a finite amount of additional information: there will in general be more than one QM structure on a pp abelian surface (assuming it has at least one QM structure). Thus a point on $Y^D(1)$ includes a *choice* of QM structure.

What happens when $D = 1$? Then $B = M_2(\mathbb{Q})$ so we are talking about abelian surfaces which are isogenous to the square of an elliptic curve then. This is in fact what we do: given any elliptic curve, E, E^2 is an abelian surface whose endomorphism ring contains $M_2(\mathbb{Z})$. The lattice in this case is just the square of the corresponding two-dimensional lattice.

In summary (and with some technical details omitted): $Y^D(1)$ parameterizes isomorphism classes of QM abelian surfaces.

As above, this leads to a canonical $\mathbb{Q}$-model, since it makes sense to consider QM-abelian schemes over S (where S is a $\mathbb{Q}$-scheme). This gives the uniqueness of the $\mathbb{Q}$-rational model, and the existence follows as for elliptic curves (for instance $X_1^D(N)$ is a fine moduli scheme iff $X_1(N)$ is).

Note that because we, in general, neither cusps to give Fourier expansions nor an obvious map to $\mathbb{P}^1$, none of the more elementary descriptions of the canonical rational model will work here.

What is the analogue of Mazur's theorem for modular curves? Things are quite different for Shimura curves:

Theorem 9. *(Shimura) For all $D > 1$, $X^D(\mathbb{R}) = \emptyset$.*

Morally speaking, this is true for the same reason that a CM elliptic curve, together with all its endomorphisms, cannot be defined over $\mathbb{R}$: it is in fact easy to show the same for a QM surface. Unfortunately a nasty subtlety intervenes to make this not literally correct: unlike modular curves, a point on a Shimura curve defined over some field K does not necessarily correspond to a QM surface defined over K. Otherwise put, there is a generally nontrivial obtruction to the field of moduli of a QM surface being a field of definition.[21] The most natural proof uses the idea that an $\mathbb{R}$-model on a complex algebraic curve is given by an antiholomorphic involution, and in our setting such a thing is provided by taking an element of $\mathcal{O}^\times$ of determinant -1.

What about $\mathbb{Q}_p$-rational points? For primes p dividing D, there is a very nice necessary and sufficient condition for $X_0^D(N)(\mathbb{Q}_p)$ to be nonempty, due in special cases to Jordan-Livné and Ogg. In particular there is always at least one p dividing D such that $X_0^D(N)(\mathbb{Q}_p) = \emptyset$. For other primes there are answers, but they are less satisfactory: for instance, for primes p prime to DN, the Eichler-Selberg trace formula says that there is a $\mathbb{Q}_p$-rational point if and only if a certain finite sum is positive. This sum is elementary enough to be easily computable in any particular case, but there is no known "closed" expression, and there is presumably nothing so simple as a congruence condition that determines the answer. Presumably what one should do is explore the behavior of this sum on average, which is the sort of thing that analytic number theorists are quite good at, although to the best of my knowledge such a computation has never been done in this particular context (so might serve as a thesis problem).

All of these results rely on facts concerning the minimal regular model of $X_0^D(N)$. It turns out that these curves have smooth reduction at primes p not dividing DN (essentially a result of Igusa), have semistable bad reduction at primes dividing D, and have bad reduction at primes dividing N which is semistable iff p exactly divides N. One can do better: if N is squarefree, then the work of Igusa, Deligne-Rapoport, Cerednik-Drinfeld, and Buzzard (and to a lesser extent, of David Helm and myself) gives an explicit description of an integral canonical model of $X_0^D(N)$. The explicit description of the model at primes dividing D requires p-adic uniformization so lies outside the scope of these lectures. The model at p when p exactly divides N will make an appearance in the third lecture.

[21]This makes the theory of QM surfaces significantly more technically challenging than the theory of elliptic curves. There is, for instance, a mistake along these lines in the PhD thesis of Bruce Jordan, and a different, but similar mistake in my thesis.