# LECTURES ON SHIMURA CURVES 0: MODULAR CURVES

PETE L. CLARK

General remarks about these documents: These are *lecture notes*. Let us reflect separately on the meaning of both words. The first word is meant to indicate some correspondence between what is said in the lectures and what appears in these notes. This correspondence is not precise: on the one hand, in response to a question or a comment, I may say things in class which do not make it into these documents. (If I or someone else says something *important* that does not appear here, let me know, and I will try to incorporate it into a later draft.) On the other hand in certain places the notes will go into significantly more depth than what is said in class. It is also hoped that gaps in arguments which get pointed out in class will be filled in the notes. The meaning of the second word is that these documents are provisional and works in progress – we hope that the mathematics presented in here is clear, and is complete at least where it claims to be complete. It would be nice if every unproved fact were accompanied by a precise reference, but this may be too much to hope for. An important part of your responsibility in taking the class is to look through the notes and to suggest improvements. Perhaps some day these notes will be turned into a formal written document, but that day is not especially near at hand.

Interspersed throughout the notes are **Exercises** and **Problems**. If you want to keep up with the material, then working on the exercises is very strongly recommended. However, solving the exercises is not required in any formal sense, or at least not by me. Not all of the exercises have the same status: some exercises ask for the proof of a stated result (the implication here is usually that it follows rather easily from what has already been said), whereas others are more ancillary. I have put stars in front of certain exercises or parts of exercises. A star means the exercise is likely to require some extra knowledge, or in other words a possibly "unfair" question, so be warned. A problem is more serious than an exercise, either in difficulty or scale. A problem may not be phrased so as to have a clearly defined solution, or the solution may be unknown (or unknown to me).

## 1. Introducing modular curves

In one sentence, the goal of this course is to become acquainted with certain **Shimura varieties** which are generalizations of the classical modular curves (like $X_0(N)$) in that they parameterize certain higher-dimensional abelian varieties, and to concentrate especially on the case of (quaternionic) Shimura curves.

### 1.1. **Complex elliptic curves.**

What is meant by an elliptic curve over the complex numbers? There are at least three quite different looking definitions:

(EC1) It is given by a Weierstrass equation

$$y^2z = x^3 + Ax^2z + Bz^3$$

whose discriminant $\Delta = 4A^3 + 27B^2$ is nonzero.

Advantages: (i) elementary nature; (ii) amenability to computation; (iii) generality (works over every field of characteristic not 2 or 3 and over every field with minor modifications).
Disadvantages: (i) The equation is not intrinsic, necessitating a theory of changing between all possible equations. (ii) does not suggest generalizations (other than hyperelliptic curves!).

(EC2) It is both a compact Riemann surface (= complex manifold of dimension one) and a complex Lie group.

It is not too hard to show that all such objects (EC2) are of the form $\mathbb{C}/\Lambda$ (i.e., topologically a torus) where $\Lambda \cong \mathbb{Z}^2$ is a *lattice* in $\mathbb{C}$ viewed as a two-dimensional $\mathbb{R}$-vector space, and conversely such quotients clearly give objects (EC2).

Advantage: The nice geometric picture makes many properties obvious (e.g., the structure of the torsion subgroup) obvious for which purely algebraic proofs would be much harder fought (or impossible).
Disadvantage: Again suggests "the wrong" higher-dimensional generalization, to complex tori rather than abelian varieties.

(EC3) The instance $k = \mathbb{C}$ of the following definition: An elliptic curve over an arbitrary field $k$ is a nice (= nonsingular, connected) projective curve of genus one endowed with a $k$-rational point $O \in C(k)$.

(EC4) The instance $k = \mathbb{C}$ of: An elliptic curve is a one-dimensional abelian variety, i.e., a nice projective group variety $A_{/k}$.

The last two definitions I would regard as "correct" but more technically demanding and not as immediately useful as (EC1) and (EC2).

These four definitions are in fact equivalent, although the equivalence is by no means obvious. (The proofs can be found, e.g., in Silverman's two books.) It might be educational to at least think about what is involved in proving the equivalence of these four criteria.

Let us recall some of the relations between the first two definitions. For an elliptic curve given by a Weierstrass equation as above, one can define its $j$-invariant

$$j(E) = \frac{2^8 3^3 A^3}{4A^3 + 27B^2},$$

and as the terminology suggests, $j$ is independent of the choice of the Weierstrass equation so gives a complex number which is an invariant of $E$. Moreover, the following facts are known:

(J1) If $E$ and $E'$ are $\mathbb{C}$-elliptic curves, then $E \cong E' \iff j(E) = j(E')$.

(J2) For every $j \in \mathbb{C}$, there exists an $E$ with $j(E) = j$.

In other words, over the complex numbers, isomorphism classes of elliptic curves are naturally paramterized by a copy of the complex numbers – or, to rephrase things in a more pretentious way – by $\mathbb{A}^1$, the affine line over the complex numbers. The slogan here is: "The moduli space of elliptic curves is the $j$-line."

What we are looking for is moduli spaces of higher-dimensional abelian varieties, but this purely algebraic discussion doesn't suggest anything. Much more enlightening is the corresponding discussion of the $j$-invariant in the language of (EC2), which we will now review.

Namely, let us try to construct (roughly speaking) "the moduli space of lattices" $\Lambda$ in $\mathbb{C}$. Viewing $\mathbb{C} = \mathbb{R}^2$, we can specify a lattice by giving its basis, which will be two vectors $v_1,\ v_2 \in \mathbb{R}^2$, and the only condition is that they be $(\mathbb{R})$-linearly independent, or in other words that the matrix $[v_1|v_2]$ should have nonzero determinant: this gives us $GL_2(\mathbb{R})$.

We are certainly not done, since on the one hand we are giving ourselves not just a lattice but a "framed lattice" (i.e., a choice of basis), and on the other hand we need to consider the relationship between two lattices $\Lambda,\ \Lambda'$ such that $\mathbb{C}/\Lambda$ and $\mathbb{C}/\Lambda'$ are isomorphic elliptic curves. It is not hard to see that such an isomorphism determines and is determined by a $\mathbb{C}$-linear isomorphism $\alpha : \mathbb{C} \to \mathbb{C}$ (here we view $\mathbb{C}$ as the universal cover of the two elliptic curves) with the property that $\alpha(\Lambda) \subset \Lambda'$. The $\mathbb{C}$-linear maps from $\mathbb{C}$ to $\mathbb{C}$ are just the constants, so the right equivalence relation on lattices is **homothety**: for any $\alpha \in \mathbb{C}^\times$, $\Lambda$ and $\alpha\Lambda$ give the same elliptic curve. In other words, we need to consider the quotient $GL_2(\mathbb{R})/\mathbb{C}^\times$.

Exercise 0.1: Show that $GL_2(\mathbb{R})/\mathbb{C}^\times$ is naturally isomorphic to the "double half plane" $\mathcal{H}^\pm = \{x + yi \mid y \neq 0\}$.

More precisely, $GL_2(\mathbb{R})$ acts on $\mathcal{H}^\pm$ by linear fractional transformations:

$$\left[\begin{array}{cc} a & b \\ c & d \end{array}\right](z) = \frac{az+b}{cz+d}$$

so that $GL_2(\mathbb{R})^+$ (the matices with positive determinant) act on $\mathcal{H}$ by fractional transformations, and the above exercise is equivalent to the claim that this action is transitive and the stabilizer of $i$ is $SO(2)$, i.e., the complex numbers of norm 1.

The passage from $GL_2(\mathbb{R})$ to $GL_2(\mathbb{R})^+$ can be seen as requiring our basis vector $[z_1 \mid z_2]$ to be positively oriented. Thus, by dividing by $z_1$, we get a reprentative of any elliptic curve by the lattice $\Lambda_\tau = [1 \mid \tau]$ for $\tau \in \mathcal{H}$. Thus we get a $\mathbb{C}$-valued function on the upper halfplane,

$$J(\tau) := j(E_\tau).$$

This function is not an injection, because $\mathbb{Z} + \mathbb{Z}\tau$ and $\mathbb{Z} + \mathbb{Z}\tau'$ can be the same lattice without $\tau = \tau'$. The ambiguity here is in the choice of an oriented basis for the lattice, and this is easy to account for:

Exercise 0.2: Show that $\Lambda_\tau = \Lambda_{\tau'}$ if and only if there exists a matrix $m \in SL_2(\mathbb{Z})$ such that $m\tau = \tau'$.

In other words, $J$ factors through to give a function

$$J : \mathcal{H}/SL_2(\mathbb{Z}) \to \mathbb{C}.$$

Let us write $Y(1)$ for the quotient space $\mathcal{H}/SL_2(\mathbb{Z})$: it is an example of a Riemann surface uniformized by a Fuchsian group. The argument shows that the descended function $J$ is an injection. In fact, it can be shown that $J$ is an isomorphism.[1]

In summary, we obtain the $j$-line, the moduli space for elliptic curves, by considering the quotient of the upper halfplane by the discrete subgroup $SL_2(\mathbb{Z})$, or equivalently, as the double coset space $GL_2(\mathbb{Z})\backslash GL_2(\mathbb{R})/\mathbb{C}^\times$.

1.2. **Modular curves of higher level.** Put $\Gamma(1) := SL_2(\mathbb{Z})$ and let $N$ be a positive integer. We define a subgroup $\Gamma(N) \subset SL_2(\mathbb{Z})$ as the kernel of the homomorphism of groups $SL_2(\mathbb{Z}) \to SL_2(\mathbb{Z}/N\mathbb{Z})$ obtained by reducing entries modulo $N$. In other words, $\Gamma(N)$ is the set of matrices $\left[ \begin{array}{cc} a & b \\ c & d \end{array} \right] \in SL_2(\mathbb{Z})$ such that $a-1,\ b,\ c,\ d-1$ are all divisible by $N$. A group $\Gamma$ such that $\Gamma(N) \subset \Gamma \subset \Gamma(1)$ is called a **congruence subgroup** of $SL_2(\mathbb{Z})$.[2] To such a $\Gamma$, we can associate the Riemann surface

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

Notice that $Y(\Gamma)$ comes equipped with a natural map to $Y(1)$, via the quotient map $\Gamma\backslash\mathcal{H} \to SL_2(\mathbb{Z})\backslash\mathcal{H}$, which is finite of degree either $[\Gamma(1) : \Gamma]$ or $\frac{1}{2}[\Gamma(1) : \Gamma]$. The curve $Y(\Gamma)$ is a moduli space for elliptic curves together with some extra structure on the $N$-torsion subgroup. This can be explained (probably too briefly for those who have not seen the material before) as follows: on the "framed" elliptic curve $E_\tau$ we can, for any $N \geq 2$ write down a canonical basis of the $N$-torsion, namely $(\frac{1}{N}, \frac{\tau}{N})$. If, however, we change the lattice basis by some $m \in SL_2(\mathbb{Z})$, we get a different canonical basis, and the new and old bases can be compared via a matrix $T \in SL_2(\mathbb{Z}/N\mathbb{Z})$. It is thus at least plausible that if we restricted to $m$ which are congruent to the identity modulo $N$, the transition matrix $T$ will also be the identity, so in other words $Y(\Gamma(N))$ parameterizes triples $(E, P, Q)$, where $E$ is an elliptic curve and $P,\ Q$ give a basis for $E[N]$.

Exercise 0.3:
a) Work out the details of this moduli interpretation of $Y(\Gamma(N))$.
b) Let $\Gamma_0(N) = \{ \left[ \begin{array}{cc} a & b \\ c & d \end{array} \right] \in SL_2(\mathbb{Z}) \mid c \equiv 0 \pmod{N} \}$.
Show that $Y_0(N) := Y(\Gamma_0(N))$ parameterizes elliptic curves together with a distinguished cyclic order $N$ subgroup.
c) Let $\Gamma_1(N) = \{ \left[ \begin{array}{cc} a & b \\ c & d \end{array} \right] \in \Gamma_0(N) \mid a \equiv 1 \pmod{N} \}$.

[1] The surjectivity is the hardest part; see Corollary I.4.3 of Silverman's second book on elliptic curves.

[2] It is *not* the case that every finite index subgroup is a congruence subgroup, as we shall discuss later on.

Show that $Y_1(N) := Y(\Gamma_1(N))$ parameterizes elliptic curves together with a distinguished point of order $N$.

1.3. **Rational models.** In particular, each $Y(\Gamma)$ is an (affine) algebraic curve over the complex numbers. It turns out that elliptic curves over the complex numbers are no longer so interesting to algebraic geometers – their structure is well understood. However, elliptic curves over other fields (e.g. $\mathbb{Q}$) remain of paramount interest to number theorists. For an arbitrary field $k$, our definition (EC2) does not make sense, although the other three do. However, if $k$ has characteristic zero then to an elliptic curve over $k$ one can associate an elliptic curve over $\mathbb{C}$, so that such an elliptic curve has all the structure of a complex elliptic curve and more (in particular, complex analytic methods can still be used!).

If $E$ is an elliptic curve over $k$, then its $j$-invariant evidently lies in $k$, and the proof of $(J2)$ actually shows that for any $j \in k$, there exists an elliptic curve $E$ defined by a Weierstrass equation all of whose coefficients are in $k$, such that $j(E) = k$. Thus in some sense $\mathbb{A}^1_{/k}$ is still a moduli space for elliptic curves over $k$. (However, $(J1)$ does not hold: if $j(E) = j(E')$, then one may need to extend the base field in order for $E$ and $E'$ to be isomorphic. This is the phenomenon of (usually quadratic) twists.)

The task therefore is, for a congruence subgroup $\Gamma$, to realize $X(\Gamma)$ not only as a complex curve but also as a curve defined over $\mathbb{Q}$. The details of this are certainly to be left for later: for now, suffice it to say that this goal succeeds without reservations when $\Gamma = \Gamma_0(N)$ or $\Gamma_1(N)$ (for any $N$): one does indeed get curves $Y_0(N)$, $Y_1(N)$ defined over $\mathbb{Q}$. In general, if $\Gamma$ contains $\Gamma(N)$, then one gets a nice model of $X(\Gamma)$ over $\mathbb{Q}(\zeta_N)$ (where $\zeta_N$ is a primitive $N$th root of unity) and a model over $\mathbb{Q}$ which may be disconnected.

A major problem in arithmetic geometry is to classify the points on these curves – especially $Y_0(N)$ and $Y_1(N)$ – rational over $\mathbb{Q}$ and over various fields of higher degree. Here are some of the spectacular results:

**Theorem 1.** *(Mazur) The possible structures of the torsion subgroup of an elliptic curve over $\mathbb{Q}$ are as follows: $\mathbb{Z}/N\mathbb{Z}$ for $1 \leq N \leq 10$ or $N = 12$, or $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2N\mathbb{Z}$ for $1 \leq N \leq 4$.*

The complete answer is also known over quadratic fields; I am not sure if the work is complete over cubic fields.

**Theorem 2.** *(Merel) There exists a function $C(d)$ such that if $K$ is a number field with $[K : \mathbb{Q}] \leq d$, then $N \geq C(d) \implies Y_1(N)(K) = \emptyset$.*

**Theorem 3.** *(Mazur) If $N > 163$ is prime, then $Y_0(N)(\mathbb{Q}) = \emptyset$.*

The curve $Y(\Gamma)$ is always affine, but like any affine curve it can be "completed" in a canonical way. By this, I mean that if $C$ is any nice complex algebraic curve, then there exists a unique integer $n \geq 0$ such that $C(\mathbb{C})$ is topologically an orientable surface of genus $g$ with $n$ points removed. In our context, the compactification can also be performed as follows: namely, $Y(1) \cong \mathbb{A}^1$, whose completion is the familiar object $\mathbb{P}^1$. Call this $X(1)$. Then $X(\Gamma)$ can be constructed as the fiber product $Y(\Gamma) \times_{Y(1)} X(1)$, or (equivalently) as the normalization of $X(1)$ in $Y(\Gamma)$.

(Of course, there is a nicer way to describe the added points, namely as *cusps*, but it is worth making the point that this theory is not necessary for constructing the curves $X(\Gamma)$.)

We can now state what is arguably the most important single result of twentieth century number theory:

**Theorem 4.** *(Taylor, Wiles, Breuil, Conrad, Diamond...) Let $E_{/\mathbb{Q}}$ be a rational elliptic curve. Then there exists a positive integer $N$ and a finite map $X_0(N) \to E$ defined over $\mathbb{Q}$.*

In other words, every rational elliptic curve is modular. Work of Darmon (and Bertolini, Dasgupta, ...) has explored what might be the right analogue for elliptic curves defined over more general number fields. On the other hand, one can ask for the generalization of this enormous theorem to abelian varieties, and the statement of this conjecture is one of the goals of the first part of the course.