1 Introduction

Most of the content of the first lecture is contained in the slides that were used in class, which aimed to give a broad overview of the theory and applications of elliptic curves. The purpose of these notes is to summarize the formal definitions we will use in future lectures and to provide additional details on using the Newton polygon to compute the genus of a plane curve. They imply, in particular, that all nonsingular cubics, including the Weierstrass equation $y^2 = x^3 + Ax + B$ with $-16(4A^3 + 27B^2) \neq 0$, are curves of genus 1, as are Edwards curves $x^2 + y^2 = 1 + cx^2y^2$ with $c \neq 0, 1$, which are the main cases of interest to us.

1.1 Formal definition of an elliptic curve

Definition 1.1. Let k be a field. An *elliptic curve* E/k is a smooth projective curve of genus 1 defined over k with a distinguished k-rational point O.

Note that the field k and the k-rational point O are part of the definition. To make this precise we need to define the terms "smooth", "projective curve", "genus 1", and "k-rational point" that appear in the definition. For any field k we use $\bar{k}$ to denote an *algebraic closure* of k, which can be formed by adjoining the roots of all polynomials in $k[x]$ to k.

Definition 1.2. Let k be a field. A *projective point in* $\mathbb{P}^n$ is an equivalence class of tuples $(x_0, \dots, x_n) \in \bar{k}^{n+1}$ with at least one $x_i \neq 0$ given by the equivalence relation

$$(x_0, \dots, x_n) \sim (\lambda x_0, \dots, \lambda x_n)$$

for all $\lambda \in \bar{k}^\times$. A projective point in $\mathbb{P}^n$ is *k-rational* if it contains a representative with $(x_0, \dots, x_n) \in k^{n+1}$, equivalently, it is a tuple $(x_0, \dots, x_n)$ with $x_i/x_j \in k$ for all $x_j \neq 0$. We use the notation $(x_0 : \cdots : x_n)$ to denote the equivalence class of the tuple $(x_0, \dots, x_n)$. The set of k-rational projective points in $\mathbb{P}^n$ is denoted $\mathbb{P}^n(k)$.

For $n = 2$ we typically use the coordinates x, y, z rather than x_0, x_1, x_2 and call $\mathbb{P}^2$ the *projective plane*. It will be convenient to distinguish the subset $(x : y : 1)$ of projective points in $\mathbb{P}^2$ with nonzero z-coordinate as the *affine plane* $\mathbb{A}^2$. The projective points in $\mathbb{P}^2$ that do not lie in the affine plane (those with z-coordinate zero) make up the *line at infinity*, which is isomorphic to the projective line $\mathbb{P}^1$. Of course the choice of the coordinate z is arbitrary, we could have chosen x or y, but z is most commonly used.

Definition 1.3. Let R be a commutative ring. A polynomial $f \in R[x_0, \dots, x_n]$ is *homogeneous* if every nonzero term of f has the same degree. For any nonzero polynomial $f \in R[x_0, \dots, x_n]$ we use $\deg f$ to denote the maximum of the degrees of its nonzero terms. For each $f \in R[x_0, \dots, x_{n-1}]$ there is a unique homogeneous polynomial $f^* \in R[x_0, \dots, x_n]$ with $\deg f^* = \deg f$ that satisfies $f^*(x_0, \dots x_{n-1}, 1) = f$. It can be computed by replacing each term t of f with the term $tx_n^{\deg f - \deg t}$. The polynomial f^* is the *homogenization* of f, and f is a *dehomogenization* of f^*.

Definition 1.4. Let k be a field. A *plane projective curve* $X\colon f(x, y, z) = 0$ is defined by a nonzero homogeneous polynomial $f \in k[x, y, z]$ that is irreducible as an element of $\bar{k}[x, y, z]$. For any extension K/k the set of *K-rational points* of X is the zero locus of f in $\mathbb{P}^2(K)$:

$$X(K) \coloneqq \{(x : y : z) \in \mathbb{P}^2(K) | f(x, y, z) = 0\}.$$

Because f is homogeneous, we have $f(\lambda x, \lambda y, \lambda z) = \lambda^{\deg f} f(x, y, z)$ for any nonzero λ. It follows that either f vanishes at every element of the equivalence class $(x : y : z)$, or it vanishes at none of them; this ensures that $X(K)$ is well defined.

For any nonzero $\lambda \in k$ the polynomial λf has the same zero locus in $\mathbb{P}^2(K)$ for every extension K/k. The polynomials f and λf thus define the same curve X because they have the same *functor of points* $K \mapsto X(K)$, which sends each field extension K/k to the set $X(K)$. Conversely, the functor of points $K \mapsto X(K)$ determines f up to multiplication by $\lambda \in \bar{k}^\times$ (in fact $X(\bar{k})$ is enough). A slightly more general perspective is to view the curve X as being defined by the ideal (f) that f generates; note that $(f) = (g)$ if and only if $g = \lambda f$ for some nonzero λ.

Our requirement that f is irreducible ensures that f generates a prime ideal in $k[x, y, z]$, equivalently, that the quotient ring $k[z, y, z]/(f)$ is an integral domain (has no zero divisors). The quotient ring $k[x, y, z]/(f)$ is the *coordinate ring* of the curve X, denoted $k[X]$, which will play a role in future lectures. We want it to be an integral domain so that we can consider its field of fractions, which we will use to define the *function field* $k(X)$.

We impose the stronger condition that f is irreducible in $\bar{k}[x, y, z]$ to ensure that f generates a prime ideal in $K[x, y, z]$ for any field extension K/k, so that $K[x, y, z]/(f)$ is always an integral domain (hence also has a field of fractions). Note that irreducibility in $\bar{k}[x, y, z]$ is sufficient even if K is not contained in $\bar{k}$; a polynomial in $\mathbb{Q}[x, y, z]$ that is irreducible in $\overline{\mathbb{Q}}[x, y, z]$ will also be irreducible in $\mathbb{C}[x, y, z]$, for example. But irreducibility in $k[x, y, z]$ is not sufficient: the polynomial $x^2 + y^2$ is irreducible in $\mathbb{Q}[x, y]$ but not in $\overline{\mathbb{Q}}[x, y]$, where it factors as $(x + iy)(x - iy)$, for example. Requiring irreducibility over $\bar{k}$ ensures that our curves are always *geometrically irreducible*.

We will often define plane curves using an affine equation of the form $g(x, y) = h(x, y)$ with $g, h \in k[x, y]$ distinct. Such an equation should be interpreted as defining the curve associated to the homogeneous polynomial $f(x, y, z) \coloneqq g^*(x, y, z) - h^*(x, y, z)$. In this course all curves will be plane projective curves, even when they are defined by an affine equation.

Definition 1.5. A plane projective curve X/k defined by $f \in k[x, y, z]$ is *smooth* at a point $P \in X(\bar{k})$ if at least one of the partial derivatives $\partial f/\partial x$, $\partial f/\partial y$, $\partial f/\partial z$ is nonzero at P (note that the partial derivatives are all homogeneous polynomials); otherwise P is a *singular* point of X. The curve X is *smooth* if it is smooth at every point in $X(\bar{k})$, equivalently, there are no points in the common zero locus of f and its partial derivatives.

Remark 1.6. One can define the notion of a projective curve in $\mathbb{P}^n$, for any $n \geq 2$, as an algebraic variety of dimension one. For $n > 2$ one uses the zero locus of a set of homogeneous polynomials (or more precisely, the ideal I that they generate, which we require to be a prime ideal in $\bar{k}[x_0, \ldots, x_n]$) to define the functor of points, and uses the *Krull dimension* (the maximal length of a chain of prime ideals) of the ring $\bar{k}[x_0, \ldots, x_n]/I$ to compute its dimension, which we require to be one. One then defines the notion of a smooth point using a matrix of partial derivatives with a row for each polynomial. Plane projective curves are the only curves we will consider in this course, in which case we can assume that I is generated by a nonzero homogeneous polynomial $f \in k[x, y, z]$ that is irreducible in $\bar{k}[x, y, z]$.

1.2 The genus of a plane curve

To formally define the genus of a curve over an arbitrary field requires material that is beyond the scope of this course (one needs the Riemann-Roch theorem). In this section we give a simple criterion for determining the genus of a plane projective curve defined by an

affine equation $f(x,y)=0$ for suitable polynomials $f\in k[x,y]$ that involves counting integer lattice points in the interior of its Newton polygon. This method can be used to compute the genus of all the curves we will consider. For those not familiar with the Riemann-Roch theorem, Proposition 1.11 below can be taken as the definition of the genus of a plane projective curve defined by a suitable polynomial $f\in k[x,y]$.

Let k be a field with algebraic closure $\bar{k}$. As above, for a polynomial $f\in k[x,y]$ we use $f^*\in k[x,y,z]$ to denote its homogenization.

Definition 1.7. For a polynomial $f(x,y)=\sum a_{ij}x^iy^j\in k[x,y]$, the *Newton polygon* $\Delta(f)$ of f is the convex hull of the set $\{(i,j):a_{ij}\neq 0\}\subseteq\mathbb{Z}^2$ in $\mathbb{R}^2$. The interior and boundary of $\Delta(f)$ are denoted $\Delta^\circ(f)$ and $\partial\Delta(f)$, respectively, and for each edge $\gamma\subseteq\Delta(f)$ we define the polynomial $f_\gamma(x,y):=\sum_{(i,j)\in\gamma}a_{ij}x^iy^j$.

Theorem 1.8 (Baker's Theorem)**.** *Let $f(x,y)\in k[x,y]$ be irreducible in $\bar{k}[x,y]$, and let $F:=\mathrm{Frac}(k[x,y]/(f))$ denote the corresponding function field, with genus $g(F)$. Then*

$$g(F)\leq \#\{\Delta^\circ(F)\cap\mathbb{Z}^2\}.$$

Proof. See [1, Theorem 2.4] for a short proof based on the Riemann–Roch theorem. □

Definition 1.9. A polynomial $f\in k[x,y]$ is *nondegenerate* with respect to an edge γ of $\partial\Delta(f)$ if the polynomials $f_\gamma, x\frac{\partial f_\gamma}{\partial x}, y\frac{\partial f_\gamma}{\partial y}$ have no common zero in $(\bar{k}^\times)^2$. The polynomial f is *nondegenerate* with respect to $\Delta(f)$ if it is nondegenerate with respect to every edge of $\partial\Delta(f)$ and not divisible by x or y.

Remark 1.10. For any edge γ of $\Delta(f)$, if either of the partial derivatives of $f_\gamma(x,y)$ is a monomial, then f is nondegenerate with respect to γ, since monomials have no zeros in $(\bar{k}^\times)^2$.

Proposition 1.11. *Let $f(x,y)\in k[x,y]$ be an irreducible polynomial in $\bar{k}[x,y]$ that is nondegenerate with respect to $\Delta(f)$, and suppose $f^*(x,y,z)$ has no singularities outside $\{(0:0:1),(0:1:0),(1:0:0)\}$. Then*

$$g(F)=\#\{\Delta^\circ(f)\cap\mathbb{Z}^2\}.$$

Proof. See [2, Theorem 4.2]. □

Example 1.12. Let $f(x,y)=y^2-x^3-Ax-B$, with $A,B\in k$, and $-16(4A^3+27B^2)\neq 0$. Then $f(x,y)$ is irreducible in $\bar{k}[x,y]$, and $\partial\Delta(f)$ has the three edges $\gamma_1=[(0,0),(3,0)]$, $\gamma_2=[(0,0),(0,2)]$, and $\gamma_3=[(0,2),(3,0)]$. We have

$$\begin{aligned}f_{\gamma_1}(x,y)&=-x^3-Ax-B,\\ f_{\gamma_2}(x,y)&=y^2-B,\\ f_{\gamma_3}(x,y)&=y^2-x^3.\end{aligned}$$

The polynomial $f(x,y)$ is not divisible by x or y, and the fact that the discriminant of x^3+Ax+B is nonzero implies that f is nondegenerate with respect to γ_1. By Remark 1.10, f is also nondegenerate with respect to the edges γ_2 and γ_3. Thus $f(x,y)$ is nondegenerate, and $f^*(x,y,z)$ has no singularities at all, so Proposition 1.11 implies that

$$g(f)=\#\{\Delta^\circ(f)\cap\mathbb{Z}^2\}=\#\{(1,1)\}=1.$$

Example 1.13. Let $f(x,y) = x^2 + y^2 - 1 - cx^2y^2$ with $c \neq 0, 1$. Then $f(x,y)$ is irreducible in $\bar{k}[x,y]$, and $\partial\Delta(f)$ has the four edges $\gamma_1 = [(0,0),(2,0)]$, $\gamma_2 = [(0,0),(0,2)]$, $\gamma_3 = [(0,2),(2,2)]$, and $\gamma_4 = [(2,0),(2,2)]$. We have

$$\begin{aligned} f_{\gamma_1}(x,y) &= x^2 - 1, \\ f_{\gamma_2}(x,y) &= y^2 - 1, \\ f_{\gamma_3}(x,y) &= y^2 - cx^2y^2, \\ f_{\gamma_4}(x,y) &= x^2 - cx^2y^2. \end{aligned}$$

The polynomial $f(x,y)$ is not divisible by x or y and Remark 1.10 applies to all four f_{γ_i}, thus f is nondegenerate. The homogenized polynomial $f^*(x,y,z)$ is singular only at $(0:1:0)$ and $(1:0:0)$, so f satisfies the hypothesis of Proposition 1.11 and

$$g(F) = \#\{\Delta^\circ(F) \cap \mathbb{Z}^2\} = \#\{(1,1)\} = 1.$$

References

[1] Peter Beelen, *A generalization of Baker's theorem*, Finite Fields and Their Applications **15** (2009), 558–568.

[2] Peter Beelen and Ruud Pellikaan, *The Newton polygon of plane curves with many rational points*, Designs, Codes and Cryptography **21** (2000), 41–67.

2 Elliptic curves as abelian groups

In Lecture 1 we defined an elliptic curve as a smooth projective curve of genus 1 with a distinguished rational point. An equivalent definition is that an elliptic curve is an abelian variety of dimension one. An *abelian variety* is a smooth projective variety equipped with a group structure defined by rational maps (we will make this definition more precise below). Remarkably, the fact that we are working with projective varieties rather than affine varieties forces the group operation to be commutative, which is why they are called abelian varieties.

In this lecture we will prove that elliptic curves are abelian varieties by explicitly deriving the rational maps that define the group law. In the course of doing so we will verify that they do in fact satisfy the axioms required of a group operation.

2.1 The group law for Weierstrass curves

Recall from Lecture 1 that the group law for an elliptic curve defined by a Weierstrass equation is given by the following rule:

Three points on a line sum to zero, which is the point at infinity.

For convenience let us assume we are working over a field k whose characteristic is not 2 or 3. In this case we may assume that we are working with an elliptic curve E/k defined by a short Weierstrass equation

$$E : y^2 = x^3 + Ax + B.$$

The case of a general Weierstrass equation $y^2+a_1xy+a_3y = x^3+a_2x^2+a_4x+a_6$ is essentially the same, but the formulas are slightly more complicated; see [8, III.2.3] for details and a proof that every elliptic curve can be defined by a Weierstrass equation.

Recall that although we typically specify our curves using an affine equation in the variables x and y, we are really working with the corresponding projective curve, which in this case is given by the homogeneous equation

$$E : y^2z = x^3 + Axz^2 + Bz^3.$$

In order to specify an elliptic curve we need not only an equation defining the curve, but also a distinguished rational point, which acts as the identity of the group. For curves in Weierstrass form we always take the point $O := (0 : 1 : 0)$ at infinity as our distinguished point; this is the unique point on the curve E that lies on the line $z = 0$ at infinity: if $z = 0$ then $x = 0$ and we may assume $y = 1$ after scaling the projective point $(0 : y : 0)$ by $1/y$; note that $x = z = 0$ forces $y \neq 0$, since $(0 : 0 : 0)$ is (by definition) not a projective point.

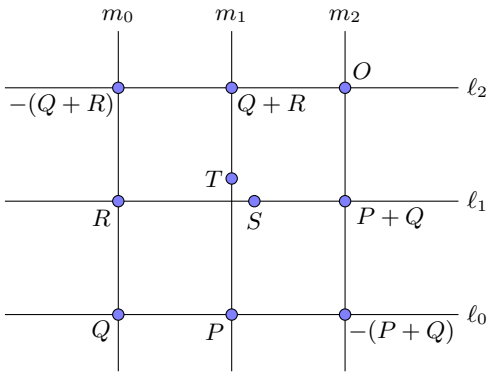
Every point $P \neq O$ on the curve E thus has a nonzero z-coordinate which we can scale to be 1, and we use the notation $P = (x_0, y_0) := (x_0 : y_0 : 1)$ to denote these affine points. Notice that the point $Q = (x_0, -y_0)$ also lies on the curve E, and the projective line through P and Q is defined by $x = x_0z$, which also passes through $O = (0 : 1 : 0)$. The three points P, Q, O lie on a line, so $P + Q + O = P + Q = O$, and therefore $Q = -P$.

Commutativity of the group law follows immediately from our definition, but associativity is not obvious. We will give two proofs. The first is geometric and depends on a genericity assumption that does not apply in special cases, but the proof is short and it provides some intuition as to *why* the group law is associative. The second is algebraic and handles all cases over any field whose characteristic is not 2 or 3, but we will rely on the computer algebra system Sage to do the heavy lifting. A formally verified proof of associativity that works in all characteristics is now available in the Lean `mathlib` library [1, 6].

2.1.1 A geometric proof of associativity in the generic case

This is an adaptation of the proof in [4, p. 28]. Let P, Q, R be points on an elliptic curve E over a field k that we may assume is algebraically closed (if the group law is associative over $\bar{k}$ then it is certainly also associative when we restrict to k). We shall also assume that P, Q, R, and the zero point O are all in *general position*. This means that in the diagram below there are no relationships among the points other than those that necessarily exist by construction; in particular the eight points $P, Q, R, O, \pm(P+Q), \pm(Q+R)$ are distinct and no three are collinear.

The line ℓ_0 through P and Q meets the curve E at a third point, $-(P+Q)$, and the line m_2 through O and $-(P+Q)$ meets E at $P+Q$. Similarly, the line m_0 through Q and R meets E at $-(Q+R)$, and the line ℓ_2 through O and $-(Q+R)$ meets E at $Q+R$. Let S be the third point where the line ℓ_1 through $P+Q$ and R meets E, and let T be the third point where the line m_1 through P and $Q+R$ meets E. See the diagram below:



We have $S = -((P+Q)+R)$ and $T = -(P+(Q+R))$. It suffices to show $S = T$. Suppose not. Let $g(x,y,z)$ be the cubic polynomial formed by the product of the lines ℓ_0, ℓ_1, ℓ_2 in homogeneous coordinates, and similarly let $h(x,y,z) = m_0 m_1 m_2$. We claim $g(T) \neq 0$. Indeed, if $g(T) = 0$ then T must lie on ℓ_1, since if it lies on ℓ_0 then so does $Q+R$ which is then collinear with P and Q, contradicting our general position assumption, and if it lies on ℓ_2 then so does P which is then collinear with O and $Q+R$, another contradiction; if T lies on ℓ_1 it must be equal to S, contrary to our supposition, because it cannot be equal to R or $P+Q$ (since neither is collinear with P and $Q+R$), and there are only three points in the intersection of ℓ_1 with E (by Bézout's theorem). Similarly, $h(S) \neq 0$.

It follows that g and h are linearly independent elements of the k-vector space V of homogeneous cubic polynomials in $k[x,y,z]$. The space V has dimension $\binom{3+2}{2} = 10$, thus the subspace of homogeneous cubic polynomials that vanish at the eight distinct points O, P, Q, R, $\pm(P+Q)$, and $\pm(Q+R)$ has dimension 2 and is spanned by g and h. The polynomial $f(x,y,z) = x^3 + Axz^2 + Bz^3 - zy^2$ that defines E is a nonzero element of this subspace, so we may write $f = ag + bh$ as a linear combination of g and h. Now $f(S) = f(T) = 0$, since S and T are both points on E, but $g(S), h(T) = 0$ and $g(T), h(S) \neq 0$, which implies that a and b are both zero, but this is a contradiction because f is not the zero polynomial.

This completes our geometric proof of associativity (in the generic case). In order to give a more general algebraic proof, and to be able to actually perform group operations explicitly, we need explicit formulas for computing the sum of two points.

2.2 The group law in algebraic terms

Let P and Q be two points on our elliptic curve $E\colon y^2 = x^3 + Ax + B$. We want to compute the point $R = P + Q$ by expressing the coordinates of R as rational functions of the coordinates of P and Q. If either P or Q is the point O at infinity, then R is simply the other point, so we assume that P and Q are affine points $P = (x_1, y_1)$ and $Q = (x_2, y_2)$. There are two cases.

Case 1. $x_1 \neq x_2$. The line $\overline{PQ}$ has slope $m = (y_2 - y_1)/(x_2 - x_1)$, which yields the linear equation $y - y_1 = m(x - x_1)$ for $\overline{PQ}$. This line is not vertical, so it intersects the curve E in a third affine point $-R = (x_3, -y_3)$. Plugging the equation for the line $\overline{PQ}$ into the equation for the curve E yields

$$(m(x - x_1) + y_1)^2 = x^3 + Ax + B.$$

Expanding the LHS and moving every term to the RHS yields a cubic equation

$$g(x) := x^3 - m^2x^2 + \cdots = 0,$$

where the ellipsis hides lower order terms in x. The monic cubic polynomial $g(x)$ has two roots $x_1, x_2 \in k$ and therefore factors in $k[x]$ as

$$g(x) = (x - x_1)(x - x_2)(x - x_3),$$

where $x_3 \in k$ is the x-coordinate of the third point $-R$ on the intersection of $\overline{PQ}$ and E. Comparing the coefficient of x^2 in the two expressions for $g(x)$ shows that $x_1 + x_2 + x_3 = m^2$, and therefore $x_3 = m^2 - x_1 - x_2$. We can then compute the y-coordinate $-y_3$ of $-R$ by plugging this expression for x_3 into the equation for $\overline{PQ}$, and we have

$$\begin{aligned} m &= (y_2 - y_1)/(x_2 - x_1), \\ x_3 &= m^2 - x_1 - x_2, \\ y_3 &= m(x_1 - x_3) - y_1, \end{aligned}$$

which expresses the coordinates of $R = P + Q$ as rational functions of the coordinates of P and Q as desired. To compute $P + Q = R$, we need to perform three multiplications (one of which is squaring m) and one inversion in the field k. We'll denote this cost $3\mathbf{M}+\mathbf{I}$; we are ignoring the cost of additions and subtractions because these are typically negligible compared to the cost of multiplications and (especially) inversions.

Case 2. $x_1 = x_2$. We must have $y_1 = \pm y_2$. If $y_1 = -y_2$ then $Q = -P$ and $P + Q = R = 0$. Otherwise $P = Q$ and $R = 2P$, and the line $\overline{PQ}$ is the tangent at P on the equation for E, whose slope we can compute by implicit differentiation. This yields

$$2y\,dy = 3x^2dx + A\,dx,$$

so at the point $P = (x_1, y_1)$ the slope of the tangent line is

$$m = \frac{dy}{dx} = \frac{3x_1^2 + A}{2y_1},$$

and once we know m we can compute x_3 and y_3 as above. Note that we require an extra multiplication (a squaring) to compute m, so computing $R = 2P$ has a cost of $4\mathbf{M}+\mathbf{I}$.

Remark 2.1. You might object that we have not formally defined implicit differentiation over an arbitrary field, nor have we shown that this gives us the slope of the tangent line. One can rigorously justify this (using Kähler differentials, for example), but it is easy to verify that it works in our case: if you plug $y = m(x - x_1) + y_1$ into the curve equation $E\colon y^2 = x^3 + Ax + B$ using the slope $m = (3x_1^2 + A)/2y_1$ we computed using implicit differentiation, you will find that x_1 is a double root, and since the point $(x_1, -y_1)$ does not lie on the line $L\colon y = m(x - x_1) + y_1$ unless $y_1 = 0$, the point (x_1, y_1) has multiplicity 2 in the intersection $E \cap L$, which implies that L is tangent to E at (x_1, y_1) as claimed.

With these equations in hand, we can now prove associativity as a formal identity, treating $x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3, A, B$ as indeterminates subject to the three relations implied by the fact that P, Q, R lie on the curve E. See the Sage worksheet

Lecture 2 Proof of associativity

for details, which includes checking all the special cases.

The equations above can be converted to projective coordinates by replacing x_1, y_1, x_2, and y_2 with x_1/z_1, y_1/z_1, x_2/z_2, and y_2/z_2 respectively, and then writing the resulting expressions for x_3/z_3 and y_3/z_3 with a common denominator. When $P \neq Q$ we obtain

$$\begin{aligned}
x_3 &= (x_2z_1 - x_1z_2)((y_2z_1 - y_1z_2)^2z_1z_2 - (x_2z_1 - x_1z_2)^2(x_2z_1 + x_1z_2))\\
y_3 &= (y_2z_1 - y_1z_2)((x_2z_1 - x_1z_2)^2(x_2z_1 + 2x_1z_2) - (y_2z_1 - y_1z_2)^2z_1z_2) - (x_2z_1 - x_1z_2)^3y_1z_2\\
z_3 &= (x_2z_1 - x_1z_2)^3z_1z_2
\end{aligned}$$

and for $P = Q$ we obtain

$$\begin{aligned}
x_3 &= 2y_1z_1(A^2(z_1^2 + 3x_1^2)^2 - 8x_1y_1^2z_1)\\
y_3 &= A(z_1^2 + 3x_1^2)(12x_1y_1^2z_1 - A^2(z_1^2 + 3x_1^2)^2) - 8y_1^4z_1^2\\
z_3 &= (2y_1z_1)^3
\end{aligned}$$

These formulas are more complicated, but they have the advantage of avoiding inversions, which are more costly than multiplications (in a finite field of cryptographic size inversions may be 50 or even 100 times more expensive than multiplications). With careful reuse of common subexpressions these formulas lead to a cost of 12**M** for addition (of distinct points) and 14**M** for doubling.

2.3 Elliptic curves as abelian varieties

An abelian variety is a smooth projective variety G/k equipped with morphisms $\mu\colon G \times G \to G$ and $i\colon G \to G$ and a k-rational point O such that for every field extension K/k the set $G(K)$ of K-rational points has the structure of a group with composition law given by μ, inverses given by i, and O as the identity element.

We have not formally defined what it means to be a smooth projective variety, but we have defined smooth projective plane curves C: these are defined by a polynomial in $f \in k[x, y, z]$ that is irreducible in $\bar{k}[x, y, z]$ such that there is no point $P \in C(\bar{k})$ at which the three (formal) partial derivatives of f simultaneously vanish. This example of a smooth projective variety suffices for our present purpose, as it includes the case of an elliptic curve. For the morphism μ we can take the rational maps defined by the polynomial expressions

we derived above for x_3, y_3, z_3 in terms of the projective coordinates x_1, y_1, z_1 and x_2, y_2, z_2, and for the inverse morphism i we simply take the map $(x : y : z) \mapsto (x : -y : z)$.

In the case of elliptic curves, the group law is commutative by construction. In fact commutativity holds for all abelian varieties [7, §4.3], which justifies their nomenclature, even though it is not obviously implied by the definition; indeed, with affine algebraic groups, which are defined exactly as abelian varieties but with the underlying algebraic variety affine rather than projective, the group operation is typically not commutative.

Remark 2.2. We have shown that elliptic curves are abelian varieties of dimension one (curves are algebraic varieties of dimension one by definition, regardless of their genus). We have not shown that every abelian variety of dimension one is an elliptic curve, which is beyond the scope of this course, but this is indeed the case (abelian varieties of dimension one are smooth projective curves, but one needs to show that they have genus 1, and that the group operation on the abelian variety necessarily coincides with that induced by the elliptic curve group law when we take the identity element as our distinguished point).

2.4 Edwards curves

Various alternative models of elliptic curves other than Weierstrass equations have been proposed over the years; each leads to different formulas for the group law that are ultimately equivalent to the formulas for curves in Weierstrass form, after applying a suitable isomorphism, but which may be more efficient to compute or have other advantages.

We give just one example here, a particular form of an *Edwards curve* [2, 3, 5]. Let a be a non-square element of a field k whose characteristic is not 2. Then the equation

$$x^2 + y^2 = 1 + ax^2y^2 \tag{1}$$

defines an elliptic curve with distinguished point $(0, 1)$.

Remark 2.3. The plane projective curve defined by equation (1) has two singular points at infinity, violating our requirement that an elliptic curve be smooth. However, this plane curve can be *desingularized* by embedding it in $\mathbb{P}^3(k)$. The points at infinity are then no longer rational, and do not play a role in the group operation on $E(k)$, whose elements can all be uniquely represented as solutions (x, y) to equation (1) above.

If we define

$$w := (ax^2 - 1)y, \qquad X := \frac{-2(w-1)}{x^2}, \qquad Y := \frac{4(w-1) + 2(a+1)x^2}{x^3},$$

then for any solution (x_0, y_0) to (1) with $x_0 \neq 0$ we obtain an affine point (X_0, Y_0) on the elliptic curve E/k defined by the Weierstrass equation

$$Y^2 = (X - a - 1)(X^2 - 4a).$$

(this is not a short Weierstrass equation, since the coefficient of X^2 is not zero, but for $\mathrm{char}(k) \neq 3$ the substitution $X = X' + a + 1$ yields a short Weierstrass equation).

If we map the solution $(0, 1)$ to the point at infinity on E and the solution $(0, -1)$ to the point $(a+1, 0)$ on E we obtain a bijection between the set of k-rational solutions to (1) and $E(k)$ (and similarly for all field extensions K of k, even though a may be a square in K). It is straightforward to check that this is in fact a bijection: if two points (x_0, y_0) map to

the same value of $X_0 := X(x_0, y_0)$ they must be of the form $(\pm x_0, y_0)$, but then the values of $Y_0 := Y(\pm x_0, y_0)$ will differ in sign unless $x_0 = 0$, but $(0, 1)$ and $(0, -1)$ are distinguished by the fact that one is mapped to the point at infinity and the other is not.

It follows that we can use the group law on E (three points on a line to sum to zero) to give the k-rational solutions to (1) the structure of a group isomorphic to $E(k)$ (and similarly if we replace k with an extension field K). One can then work out explicit formulas for this group law in terms of coordinates on the Edwards curve (1). We shall omit the details of these derivations (which are best done using a computer algebra system) and simply present the final result, which is quite pleasing.

The formula for adding points (x_1, y_1) and (x_2, y_2) in $E(k)$ is

$$(x_1, y_1) + (x_2, y_2) = \left(\frac{x_1 y_2 + x_2 y_1}{1 + a x_1 x_2 y_1 y_2}, \frac{y_1 y_2 - x_1 x_2}{1 - a x_1 x_2 y_1 y_2} \right), \tag{2}$$

which implies that the inverse of (x_1, y_1) is $(-x_1, y_1)$. In contrast to the formulas for curves in Weierstrass form, the formula in (2) is well defined for every pair of points (x_1, y_1) and (x_2, y_2) in $E(k)$.

To prove this, let us suppose for the sake of obtaining a contradiction that one of the denominators in (2) is zero for some pair of inputs $(x_1, y_1), (x_2, y_2)$. Then we must have

$$(1 + a x_1 x_2 y_1 y_2)(1 - a x_1 x_2 y_1 y_2) = 1 - a^2 x_1^2 x_2^2 y_1^2 y_2^2 = 0,$$

so $a^2 x_1^2 x_2^2 y_1^2 y_2^2 = 1$, and therefore x_1, x_2, y_1, y_2 are all nonzero. Applying this and the curve equation (twice) yields

$$x_1^2 + y_1^2 = 1 + a x_1^2 y_1^2 = 1 + \frac{1}{a x_2^2 y_2^2} = \frac{x_2^2 + y_2^2}{a x_2^2 y_2^2}.$$

By adding or subtracting $2 x_1 y_1 = \pm 2/(a x_2 y_2)$ to both sides we can obtain

$$(x_1 \pm y_1)^2 = \frac{(x_2 \pm y_2)^2}{a x_2^2 y_2^2},$$

with either choice of sign on the LHS (the sign on the RHS may vary, but in any case the numerator of the RHS is a square). Since x_1 and y_1 are nonzero, one of $x_1 + y_1$ and $x_1 - y_1$ is nonzero, and this implies that a is a square in k, but this is a contradiction, since we assumed from the beginning that a is not a square in k.

Remark 2.4. The formula in (2) works over extension fields at all points where it is well defined, but it is only for extensions K/k where a is not a square that it is guaranteed to be well defined at every K-rational point (and if a is a square the desingularization of the projective curve defined by (1) will have two rational points at infinity not handled by (1)).

As written, the group law involves five multiplications and two inversions (ignoring the multiplication by a, which we can choose to be small), which is greater than the cost of the group operation in Weierstrass form. However, in projective coordinates we have

$$\frac{x_3}{z_3} = \frac{z_1 z_2 (x_1 y_2 + x_2 y_1)}{z_1^2 z_2^2 + a x_1 x_2 y_1 y_2}, \qquad \frac{y_3}{z_3} = \frac{z_1 z_2 (y_1 y_2 - x_1 x_2)}{z_1^2 z_2^2 - a x_1 x_2 y_1 y_2}.$$

There are a bunch of common subexpressions here, and in order to compute z_3, we need a common denominator. Let $r = z_1z_2$, let $s = x_1y_2 + x_2y_1$, let $t = ax_1y_2x_2y_1$, and let $u = y_1y_2 - x_1x_2$. We then have

$$x_3 = rs(r^2 - t), \qquad y_3 = ru(r^2 + t), \qquad z_3 = (r^2 + t)(r^2 - t).$$

This yields a cost of 12**M**. If we compute s as $s = (x_1 + y_1)(x_2 + y_2) - x_1x_2 - y_1y_2$, the cost is reduced to 11**M**.

A simple Sage implementation of these formulas can be found here:

Lecture 2 Group law on Edwards curves

Because the expression in (2) is well defined at every point in $E(k)$, we do not need separate formulas for addition and doubling.[1] Moreover, we don't even need to check the cases where one or both points is the identity element, or one is the negation of the other; the same formula works in every case. Such formulas are said to be *complete*, and they have two distinct advantages. First, they can be implemented very efficiently as a straight-line program with no branching. Second, they protect against what is known as a *side-channel* attack. If you are using different formulas for addition and doubling, it is possible that an adversary may be able to externally distinguish these cases, e.g. by monitoring the CPU (electronically, thermally, or even acoustically) and noticing the difference in the time required or energy used by each operation. They can then use this information to break a cryptosystem that performs scalar multiplication by an integer n that is meant to be secret (as in Diffie-Hellman key exchange, for example), because the sequence of doubling and1 adding used in scalar multiplication effectively encodes the binary representation of n. Using complete formulas prevents a side-channel attack because exactly the same sequence of instruction is executed for every group operation.

Having said that, if you know you want to double a point and are not concerned about a side-channel attack, there are several optimizations that can be made to the formulas above (these include replacing $1 + cx^2y^2$ with $x^2 + y^2$). This reduces the cost of doubling on an Edwards curve to 7**M**, half the 14**M** cost of doubling a point in Weierstrass coordinates [2].

The explicit formulas database contains optimized formulas for Edwards curves and various generalizations, as well as many other forms of elliptic curves. Operation counts and verification scripts are provided with each set of formulas.

We should note that, unlike Weierstrass equations, not every elliptic curve can be defined by an equation in Edwards form. In particular, an Edwards curve always has a rational point of order 4, the point $(1, 0)$, but most elliptic curves do not have a rational point of order 4.

References

[1] David Kurniadi Angdinata and Junyan Xu, *An elementary formal proof of the group law on Weierstrass elliptic curves in any characteristic*, arXiv:2302.10640, 2023.

[2] Daniel J. Bernstein and Tanja Lange, *Faster addition and doubling on elliptic curves*, Advances in Cryptology - ASIACRYPT 2007, Lecture Notes in Computer Science **4833**, Springer-Verlag, New York (2007), 29–50.

[1] See [3] for formulas that achieve this for every point in $E(\bar{k})$ without assuming a is a non-square, and also for the more general case of *twisted Edwards curves* $dx^2 + y^2 = 1 + ax^2y^2$.

[3] Daniel J. Bernstein and Tanja Lange, *A complete set of addition laws for incomplete Edwards curves*, Journal of Number Theory **131** (2011), 858–872.

[4] J. W. S. Cassels, *Lectures on elliptic curves*, London Mathematical Society Student Texts **24**, Cambridge University Press, 1991.

[5] Harold M. Edwards, *A normal form for elliptic curves*, Bulletin of the American Mathematical Society **44** (2007), 393–422.

[6] The Mathlib Community, *The Lean mathematical library*, available at `https://github.com/leanprover-community/mathlib4`.

[7] Igor R. Shafarevich, *Basic algebraic geometry*, Springer, 1994.

[8] Joseph H. Silverman, *The arithmetic of elliptic curves*, 2nd ed., Springer, 2009.

3 Finite field arithmetic

In order to perform explicit computations with elliptic curves over finite fields, we first need to understand how to compute in finite fields. In many of the applications we will consider, the finite fields involved will be quite large, so it is important to understand the computational complexity of finite field operations. This is a huge topic, one to which an entire course could be devoted, but we will spend just one or two lectures on this topic, with the goal of understanding the most commonly used algorithms and analyzing their asymptotic complexity. This will force us to omit many details, but references to the relevant literature will be provided for those who want to learn more.

Our first step is to fix an explicit representation of finite field elements. This might seem like a technical detail, but it is actually quite crucial; questions of computational complexity are meaningless otherwise.

Example 3.1. By Theorem 3.12 below, the multiplicative group of a finite field $\mathbb{F}_q$ is cyclic. One way to represent the nonzero elements of a finite field is as explicit powers of a fixed generator, in which case it is enough to know the exponent, an integer in $[0, q-2]$. With this representation multiplication and division are easy, solving the discrete logarithm problem is trivial, but addition is costly (not known to be polynomial-time). We will instead choose a representation that makes addition (and subtraction) very easy, multiplication slightly harder but still easy, division slightly harder than multiplication but still easy (all these operations take quasi-linear time). But solving the discrete logarithm problem will be hard (no polynomial-time algorithm is known).

For the sake of brevity, we will focus primarily on finite fields of large characteristic, and prime fields in particular, although the algorithms we describe will work in any finite field of odd characteristic (most will also work in characteristic 2). Fields of characteristic 2 are quite important in many applications (coding theory in particular), and there are specialized algorithms that are optimized for such fields, but we will not address them here.[1]

3.1 Finite fields

We begin with a quick review of some basic facts about finite fields, all of which are straightforward but necessary for us to establish a choice of representation; we will also need them when we discuss algorithms for factoring polynomials over finite fields. Those already familiar with this material should feel free to skim this section.

Definition 3.2. For each prime p we define $\mathbb{F}_p$ to be the quotient ring $\mathbb{Z}/p\mathbb{Z}$.

Theorem 3.3. *The ring $\mathbb{F}_p$ is a field, and every field of characteristic p contains a canonical subfield isomorphic to $\mathbb{F}_p$. In particular, all fields of cardinality p are isomorphic.*

Proof. To show that the ring $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ is a field we just need to show that every nonzero element is invertible. If $[a] := a + p\mathbb{Z}$ is a nontrivial coset in $\mathbb{Z}/p\mathbb{Z}$ then a and p are coprime and $(a, p) = (1)$ is the unit ideal. Therefore $ua+vp = 1$ for some $u, v \in \mathbb{Z}$ with $ua \equiv 1 \bmod p$, so $[u][a] = [1]$ in $\mathbb{Z}/p\mathbb{Z}$ and $[a]$ is invertible. To justify the second claim, note that in any field of characteristic p the subring generated by 1 is isomorphic to $\mathbb{Z}/p\mathbb{Z} = \mathbb{F}_p$, and this subring is clearly unique (any other must also contain 1), hence canonical. □

[1] The recent breakthrough in computing discrete logarithms in finite fields of small characteristic in quasi-polynomial time [1] has greatly diminished the enthusiasm for using such fields in cryptographic applications.

The most common way to represent $\mathbb{F}_p$ for computational purposes is to pick a set of unique coset representatives for $\mathbb{Z}/p\mathbb{Z}$, such as the integers in the interval $[0, p-1]$.

Definition 3.4. For each prime power $q = p^n$ we define $\mathbb{F}_q = \mathbb{F}_{p^n}$ to be the field extension of $\mathbb{F}_p$ generated by adjoining all roots of $x^q - x$ to $\mathbb{F}_p$ (the splitting field of $x^q - x$ over $\mathbb{F}_p$). Equivalently, $\mathbb{F}_q := \overline{\mathbb{F}_p}^{\sigma_q}$ is the subfield of the algebraic closure of $\mathbb{F}_p$ fixed by the q-power Frobenius automorphism $\sigma_q \colon x \mapsto x^q$.

Remark 3.5. We note that this definition makes sense for $n = 1$, with $q = p$: the polynomial $x^p - x$ splits completely over $\mathbb{F}_p$, and $\mathbb{F}_p$ is the subfield of $\overline{\mathbb{F}_p}$ fixed by σ_p.

Theorem 3.6. *Let $q = p^n$ be a prime power. The field $\mathbb{F}_q$ has cardinality q and every field of cardinality q is (non-canonically) isomorphic to $\mathbb{F}_q$.*

Proof. The map $x \mapsto x^q = x^{p^n}$ is an automorphism σ_q of $\mathbb{F}_q$, since in characteristic p we have

$$(a+b)^{p^n} = a^{p^n} + b^{p^n} \qquad \text{and} \qquad (ab)^{p^n} = a^{p^n} b^{p^n},$$

where the first identity follows from the binomial theorem: $\binom{p^n}{r} \equiv 0 \bmod p$ for $0 < r < p^n$. Let $k := \mathbb{F}_q^{\sigma_q}$ be the subfield of $\mathbb{F}_q$ fixed by σ_q. We have $\mathbb{F}_p \subseteq k$, since

$$(1 + \cdots + 1)^q = 1^q + \cdots + 1^q = 1 + \cdots + 1,$$

and it follows that $\mathbb{F}_q \subseteq k$, since σ_q fixes $\mathbb{F}_p$ and every root of $x^q - x$, and therefore $k = \mathbb{F}_q$. The polynomial $x^q - x$ has no roots in common with its derivative $(x^q - x)' = qx^{q-1} - 1 = -1$, so it has q distinct roots, which are precisely the elements of $\mathbb{F}_q$ (they lie in $\mathbb{F}_q$ by definition, and every element of $\mathbb{F}_q = \mathbb{F}_q^{\sigma_q}$ is fixed by σ_q and therefore a root of $x^q - x$).

Now let k be a field of cardinality $q = p^n$. Then k must have characteristic p, since the set $\{1, 1+1, \ldots\}$ is a subgroup of the additive group of k, so the characteristic divides $\#k = p^n$, and in a finite ring with no zero divisors the characteristic must be prime. By Theorem 3.3, the field k contains $\mathbb{F}_p$. The order of each $\alpha \in k^\times$ divides $\#k^\times = q - 1$; thus $\alpha^{q-1} = 1$ for all $\alpha \in k^\times$, so every $\alpha \in k$, including $\alpha = 0$, is a root of $x^q - x$. It follows that k is isomorphic to a subfield of $\mathbb{F}_q$, and $\#k = \#\mathbb{F}_q$, so $k \simeq \mathbb{F}_q$ (this isomorphism is not canonical because when q is not prime there are many ways to embed k in $\mathbb{F}_q$). □

Remark 3.7. Now that we know all finite fields of cardinality q are isomorphic, we will feel free to refer to any and all of them as *the* finite field $\mathbb{F}_q$, with the understanding that there are many ways to represent $\mathbb{F}_q$ and we will need to choose one of them.

Theorem 3.8. *The finite field $\mathbb{F}_{p^m}$ is a subfield of $\mathbb{F}_{p^n}$ if and only if m divides n.*

Proof. If $\mathbb{F}_{p^m} \subseteq \mathbb{F}_{p^n}$ then $\mathbb{F}_{p^n}$ is an $\mathbb{F}_{p^m}$-vector space of (integral) dimension n/m, so $m|n$. If $m|n$ then $p^n - p^m = (p^m - 1)(p^{n-m} + p^{n-2m} + \cdots + p^{2m} + p^m)$ is divisible by $p^m - 1$ and

$$x^{p^n} - x = (x^{p^m} - x)(1 + x^{p^m - 1} + x^{2(p^m-1)} + \cdots + x^{p^n - p^m})$$

is divisible by $x^{p^m} - x$. Thus every root of $x^{p^m} - x$ is also a root of $x^{p^n} - x$, so $\mathbb{F}_{p^m} \subseteq \mathbb{F}_{p^n}$. □

Theorem 3.9. *For any irreducible $f \in \mathbb{F}_p[x]$ of degree $n > 0$ we have $\mathbb{F}_p[x]/(f) \simeq \mathbb{F}_{p^n}$.*

Proof. The ring $k := \mathbb{F}_p[x]/(f)$ is an $\mathbb{F}_p$-vector space with basis $1, x, \ldots, x^{n-1}$ and therefore has dimension n and cardinality p^n. The ring $\mathbb{F}_p[x]$ is a principal ideal domain and f is irreducible and not a unit, so (f) is a maximal ideal and $\mathbb{F}_p[x]/(f)$ is a field with p^n elements, hence isomorphic to $\mathbb{F}_{p^n}$ by Theorem 3.6. □

Theorem 3.9 allows us to explicitly represent $\mathbb{F}_{p^n}$ as $\mathbb{F}_p[x]/(f)$ using any irreducible polynomial $f \in \mathbb{F}_p[x]$ of degree n, and it does not matter which f we pick; by Theorem 3.6 we always get the same field (up to isomorphism). We also note the following corollary.

Corollary 3.10. *Every irreducible $f \in \mathbb{F}_p[x]$ of degree n splits completely in $\mathbb{F}_{p^n}$.*

Proof. We have $\mathbb{F}_p[x]/(f) \simeq \mathbb{F}_{p^n}$, so every root of f is a root of $x^{p^n} - x$ and lies in $\mathbb{F}_{p^n}$. □

Remark 3.11. This corollary implies that $x^{p^n} - x$ is the product over the divisors $d|n$ of all monic irreducible polynomials of degree d in $\mathbb{F}_p[x]$. This can be used to derive explicit formulas for the number of irreducible polynomials of degree d in $\mathbb{F}_p[x]$ using Möbius inversion. It also implies that, even though we defined $\mathbb{F}_{p^n}$ as the splitting field of $x^{p^n} - x$, it is also the splitting field of every irreducible polynomial of degree n.

Theorem 3.12. *Every finite subgroup of the multiplicative group of a field is cyclic.*

Proof. Let k be a field, let G be a subgroup of $k^\times$ of order n, and let m be the exponent of G (the least common multiple of the orders of its elements), which necessarily divides n. Every element of G is a root of $x^m - 1$, which has at most m roots, so $m = n$. Every finite abelian group contains an element of order equal to its exponent, so G contains an element of order $m = n = \#G$ and is therefore cyclic. □

Corollary 3.13. *The multiplicative group of a finite field is cyclic.*

If α is a generator for the multiplicative group $\mathbb{F}_q^\times$, then it generates $\mathbb{F}_q$ as an extension of $\mathbb{F}_p$, that is, $\mathbb{F}_q = \mathbb{F}_p(\alpha)$, and we have $\mathbb{F}_q \simeq \mathbb{F}_p[x]/(f)$, where $f \in \mathbb{F}_p[x]$ is the minimal polynomial of α, but the converse need not hold. This motivates the following definition.

Definition 3.14. A monic irreducible polynomial $f \in \mathbb{F}_p[x]$ whose roots generate the multiplicative group of the finite field $\mathbb{F}_p[x]/(f)$ is called a *primitive polynomial*.

Theorem 3.15. *For every prime p and positive integer n there exist primitive polynomials of degree n in $\mathbb{F}_p[x]$. Indeed, the number of such polynomials is $\phi(p^n - 1)/n$.*

Here $\phi(m)$ is the Euler function that counts the generators of a cyclic group of order m, equivalently, the number of integers in $[1, m-1]$ that are relatively prime to m.

Proof. Let α be a generator for $\mathbb{F}_{p^n}^\times$ with minimal polynomial $f_\alpha \in \mathbb{F}_p[x]$; then f_α is primitive. There are $\phi(p^n - 1)$ possible choices for α. Conversely, if $f \in \mathbb{F}_p[x]$ is a primitive polynomial of degree n then each of its n roots is a generator for $\mathbb{F}_q^\times$. We thus have a surjective n-to-1 map $\alpha \to f_\alpha$ from the set of generators of $\mathbb{F}_{p^n}^\times$ to the set of primitive polynomials over $\mathbb{F}_p$ of degree n; the theorem follows. □

The preceding theorem implies that there are plenty of irreducible (and even primitive) polynomials $f \in \mathbb{F}_p[x]$ that we can use to represent $\mathbb{F}_q = \mathbb{F}_p[x]/(f)$ when q is not prime. The choice of the polynomial f has some impact on the cost of reducing polynomials in $\mathbb{F}_p[x]$ modulo f; ideally we would like f to have as few nonzero coefficients as possible. We can choose f to be a binomial whenever its degree divides $p - 1$, and we can usually (although not always) choose f to be a trinomial; see [8]. Finite fields in cryptographic standards are often specified using an $f \in \mathbb{F}_p[x]$ that makes reduction modulo f particularly efficient.

For mathematical purposes it is more useful to fix a universal choice of primitive polynomials once and for all; this simplifies the task of migrating data from one computer algebra

system to another, as well as the restoration of archived data. One way to do this is to take the lexicographically minimal primitive polynomial $f_{p,n} \in \mathbb{F}_p[x]$ of each degree n, where we represent monic $f_{p,n}(x) = \sum a_i x^{n-i}$ as a sequence of integers $(1, a_1, \ldots, a_n)$ with $0 \le a_i < p$.

There are two downsides to this simple-minded approach. First (and most significantly), we would like to be able to easily embed $\mathbb{F}_{p^m}$ in $\mathbb{F}_{p^n}$ when $m|n$, which means that if α is a root of $f_{p,n}(x)$ then we really want $\alpha^{(p^n-1)/(p^m-1)}$ to be a root of $f_{p,m}(x)$, including when $m = 1$. Secondly (and less significantly), we would like the root r of $f_{p,1} = x - r$ to be the least primitive root modulo p, which will not be the case if we use the lexicographic ordering defined above, but will be the case if we tweak our sign convention and take $(1, a_1, \ldots, a_n)$ to represent the polynomial $x^n - a_1 x^{n-1} + \cdots + (-1)^n a_n$ with terms $(-1)^i a_i x^{n-i}$. This leads to the following recursive definition due to Richard Parker (named in honor of John Conway).

Definition 3.16. Order polynomials $f(x) = x^n - a_1 x^{n-1} + \cdots + (-1)^n a_n \in (\mathbb{Z}/p\mathbb{Z})[x]$ with $0 \le a_i < p$ according to the lexicographic order on integer sequences $(1, a_1, \ldots, a_n)$. For each prime p and $n > 0$ the *Conway polynomial* $f_{p,n}(x)$ is defined by:

- For $n = 1$, let $f_{p,1}(x) := x - r$, where r is the least positive integer generating $(\mathbb{Z}/p\mathbb{Z})^\times$;
- For $n > 1$, let $f_{p,n}(x)$ be the least primitive polynomial of degree n such that for every $0 < m < n$ dividing n and every root α of $f_{p,m}(x)$ we have $f_{p,n}(\alpha^{(p^n-1)/(p^m-1)}) = 0$.

That $f_{p,n}(x)$ exists is a straightforward proof by induction that we leave as an exercise.

Conway polynomials are now used by most computer algebra systems, including GAP, Magma, Macaulay2, and SageMath. One downside to their recursive definition is that it is quite time consuming to compute any particular Conway polynomial on demand; instead, each of these computer algebra systems includes a list of precomputed Conway polynomials. The key point is that, even in a post-apocalyptic scenario where all these tables are lost, they can all be readily reconstructed from the succinct definition above.

Having fixed a representation for $\mathbb{F}_q$, every finite field operation can ultimately be reduced to integer arithmetic: elements of $\mathbb{F}_p$ are represented as integers in $[0, p-1]$, and elements of $\mathbb{F}_q = \mathbb{F}_p[x]/(f)$ are represented as polynomials of degree less than $\deg f$ whose coefficients are integers in $[0, p-1]$.

Before leaving our review of finite fields, we want to recall one other key fact about finite fields, which is that every finite field $\mathbb{F}_q$ is a Galois extension of its prime field $\mathbb{F}_p$, and the Galois group $\mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_p)$ is cyclic of order $[\mathbb{F}_q : \mathbb{F}_p]$, generated by the p-power Frobenius automorphism $\sigma_p \colon x \mapsto x^p$. This follows immediately from our definition of $\mathbb{F}_q$ as the splitting field of $x^q - x$ over $\mathbb{F}_p$, provided we know that $\mathbb{F}_q/\mathbb{F}_p$ is Galois. This follows from the fact that $x^q - x$ is a *separable polynomial*.

Definition 3.17. Let k be a field and let $f = \sum f_i x^i \in k[x]$ be a polynomial. We say that f is *separable* if any of the following equivalent conditions hold:

- f has $\deg f$ distinct roots in any algebraic closure $\bar{k}$ of k;
- f is squarefree over every extension of k;
- $\gcd(f, f')$ is a unit in $k[x]$, where $f'_i := \sum i f_i x^{i-1}$ denotes the formal derivative of f.

A polynomial that is not separable is said to be *inseparable*.

Remark 3.18. We will typically write $\gcd(f, f') = 1$ to indicate that $\gcd(f, f')$ is a unit. The gcd of two elements in a ring is defined only up to units (if a divides b and c then so

does ua for any unit u), and in a principal ideal domain it is standard to take $\gcd(a,b)$ to be a unique representative of the ideal (a,b). For the ring $\mathbb{Z}$ there is a unique positive representative (the only units are ± 1), and in the ring $k[x]$ there is a unique monic representative (units are elements of $k^\times$).

Remark 3.19. Some older textbooks (notably including Bourbaki) define a polynomial to be separable if its irreducible factors are separable, which would make polynomials like $(x-1)^2$ separable, but for us this is not a separable polynomial. On the other hand, it is clear that if a polynomial f is separable under our definition, then all its irreducible factors are separable, since if f has distinct roots in $\bar{k}$ then so does every divisor of f.

Lemma 3.20. *An irreducible polynomial $f \in k[x]$ is inseparable if and only if $f' = 0$.*

Proof. Let $f \in k[x]$ be irreducible. Then f is nonzero and not a unit. If $f' = 0$ then $\gcd(f, f') = f$ is not a unit and f is inseparable. If f is inseparable then $g = \gcd(f, f')$ is a nonconstant divisor of f and f', and if f' is nonzero then $\deg g \le \deg f' < \deg f$, which is impossible because f is irreducible. □

The polynomial $x^q - x$ is separable because

$$\gcd(x^q - x, (x^q - x)') = \gcd(x^q - x, -1) = 1,$$

and it follows that its splitting field over $\mathbb{F}_p$ is a Galois extension of $\mathbb{F}_p$ (this is the basic tenet of Galois theory: splitting fields of separable polynomials $f \in k[x]$ are finite Galois extensions of k, and every finite Galois extension of k is the splitting field of some separable polynomial $f \in k[x]$). An important consequence of this fact is that finite fields are perfect.

Definition 3.21. A field k is *perfect* if every irreducible polynomial in $k[x]$ is separable, equivalently, has a nonzero derivative.

Fields of characteristic zero are always perfect, since there is no way for the derivative of a nonconstant polynomial to be zero in such fields. Fields of positive characteristic p need not be perfect (we will see many examples of this in later lectures), but finite fields are.

Theorem 3.22. *Finite fields are perfect.*

Proof. Let $f = \sum_i f_i x^i$ be an irreducible polynomial in $\mathbb{F}_q[x]$, and let

$$g \coloneqq \prod_{\sigma \in \mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_p)} f^\sigma,$$

where $f^\sigma := \sum_i \sigma(f_i) x^i$. Let $\mathbb{F}_p$ be the prime field of $\mathbb{F}_q$. We have $g \in \mathbb{F}_p[x]$, since it is invariant under the action of $\mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_p)$, and it is irreducible in $\mathbb{F}_p[x]$ since any non-trivial factor of g in $\mathbb{F}_p[x]$ would also be a non-trivial factor in $\mathbb{F}_q[x]$, none of which are invariant under the action of $\mathrm{Gal}(\mathbb{F}_q/\mathbb{F}_p)$ (note that each f^σ is irreducible in $\mathbb{F}_q[x]$). Now $f | g$, so if g is separable then f is separable, which means that if $\mathbb{F}_p$ is perfect then so is $\mathbb{F}_q$.

Let $g = \sum_i g_i x^i$. If g is inseparable then $g' = \sum_i i g_i x^{i-1} = 0$, which implies that $g_i = 0$ for i not divisible by p, meaning that $g = h(x^p)$ for some $h \in \mathbb{F}_p[x]$. But this cannot be the case because $h(x^p) = h(x)^p$ is not irreducible. □

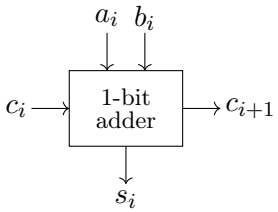
3.2 Integer addition

Every nonnegative integer a has a unique *binary representation* $a = \sum_{i=0}^{n-1} a_i 2^i$ with $a_i \in \{0,1\}$ and $a_{n-1} \neq 0$. The binary digits a_i are called *bits*, and we say that a is an *n-bit integer*; we can represent negative integers by including an additional sign bit.

To add two integers in their binary representations we apply the "schoolbook" method, adding bits and carrying as needed. For example, we can compute 43+37=80 in binary as

$$\begin{array}{r} {\scriptstyle 101111} \\ 101011 \\ +100101 \\ \hline 1010000 \end{array}$$

The carry bits are shown in red. To see how this might implemented in a computer, consider a 1-bit adder that takes two bits a_i and b_i to be added, along with a carry bit c_i.



$$c_{i+1} = (a_i \wedge b_i) \vee (c_i \wedge a_i) \vee (c_i \wedge b_i)$$

$$s_i = a_i \otimes b_i \otimes c_i$$

The symbols $\wedge$, $\vee$, and $\otimes$ denote the boolean functions AND, OR, and XOR (exclusive-or) respectively, which we may regard as primitive components of a boolean circuit. By chaining $n+1$ of these 1-bit adders together, we can add two n-bit numbers using $7n+7 = O(n)$ boolean operations on individual bits.

Remark 3.23. Chaining adders is known as *ripple* addition and is no longer commonly used, since it forces a sequential computation. In practice more sophisticated methods such as *carry-lookahead* are used to facilitate parallelism. This allows most modern microprocessors to add two 64 (or even 128) bit integers in a single clock cycle, and with the SIMD (Single Instruction Multiple Data) instruction sets available on newer AMD and Intel processors, one may be able to perform four (or even eight) 64 bit additions in a single clock cycle.

We could instead represent the same integer a as a sequence of words rather than bits. For example, write $a = \sum_{i=0}^{k-1} a_i 2^{64i}$, where $k = \left\lceil \frac{n}{64} \right\rceil$. We may then add two integers using a sequence of $O(k)$, equivalently, $O(n)$, operations on 64-bit words. Each word operation is ultimately implemented as a boolean circuit that involves operations on individual bits, but since the word-size is fixed, the number of bit operations required to implement any particular word operation is a constant. So the number of bit operations is again $O(n)$, and if we ignore constant factors it does not matter whether we count bit or word operations.

Subtraction is analogous to addition (now we need to borrow rather than carry), and has the same complexity, so we will not distinguish these operations when analyzing the complexity of algorithms. With addition and subtraction of integers, we have everything we need to perform addition and subtraction in a finite field. To add two elements of $\mathbb{F}_p \simeq \mathbb{Z}/p\mathbb{Z}$ that are uniquely represented as integers in the interval $[0, p-1]$ we simply add the integers and check whether the result is greater than or equal to p; if so we subtract p to obtain a value in $[0, p-1]$. Similarly, after subtracting two integers we add p if the result is negative.

The total work involved is still $O(n)$ bit operations, where $n = \lg p$ is the number of bits needed to represent a finite field element.

To add or subtract two elements of $\mathbb{F}_q \simeq (\mathbb{Z}/p\mathbb{Z})[x]/(f)$ we simply add or subtract the corresponding coefficients of the polynomials, for a total cost of $O(d \lg p)$ bit operations, where $d = \deg f$, which is again $O(n)$ bit operations, if we put $n = \lg q = d \lg p$.

Theorem 3.24. *The time to add or subtract two elements of $\mathbb{F}_q$ in our standard representation is $O(n)$, where $n = \lg q$ is the size of a finite field element.*

Remark 3.25. We will discuss the problem of reducing an integer modulo a prime p using fast Euclidean division in the next lecture. But this operation is not needed to reduce the sum or difference of two integers in $[0, p-1]$ to a representative in $[0, p-1]$; it is faster (both in theory and practice) to simply subtract or add p as required (at most once).

3.3 A quick refresher on asymptotic notation

Let f and g be two real-valued functions whose domains include the positive integers. The *big-O* notation "$f(n) = O(g(n))$" is shorthand for the statement:

There exist constants c and N such that for all $n \geq N$ we have $|f(n)| \leq c|g(n)|$.

This is equivalent to

$$\limsup_{n\to\infty} \frac{|f(n)|}{|g(n)|} < \infty.$$

Warning 3.26. "$f(n) = O(g(n))$" is an abuse of notation; in words we would say $f(n)$ is $O(g(n))$, where the word "is" does not imply equality (e.g., "Aristotle is a man"), and it is generally better to write this way. Symbolically, it would make more sense to write $f(n) \in O(g(n))$, regarding $O(g(n))$ as a set of functions. Some do, but the notation $f(n) = O(g(n))$ is far more common and we will occasionally use it in this course, with one caveat: we will never write a big-O expression to the left of the equal sign. It may be true that $f(n) = O(n \log n)$ implies $f(n) = O(n^2)$, but we avoid writing $O(n \log n) = O(n^2)$ because $O(n^2) \neq O(n \log n)$.

We also have *big-*Ω notation "$f(n) = \Omega(g(n))$", which means $g(n) = O(f(n))$,[2] as well as *little-o* notation "$f(n) = o(g(n))$," which is shorthand for

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = 0.$$

An alternative notation that is sometimes used is $f \ll g$, but depending on the author this may mean $f(n) = o(g(n))$ or $f(n) = O(g(n))$ (computer scientists tend to mean the former, while number theorists usually mean the latter, so we will avoid this notation). There is also a little-omega notation, but the symbol ω already has so many uses in number theory that we will not burden it further (we can always use little-o notation instead). The notation $f(n) = \Theta(g(n))$ means that $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ both hold.

It is easy to see that the complexity of integer addition is $\Theta(n)$, since we have shown it is $O(n)$ and it is clearly $\Omega(n)$ because it takes this long to output n bits (in a Turing machine model one can show that for most inputs the machine will have to write to $\Omega(n)$ cells on the Turing tape, no matter what algorithm it uses).

[2] The Ω-notation originally defined by Hardy and Littlewood had a slightly weaker definition, but modern usage generally follows our convention, which is due to Knuth.

Warning 3.27. Don't confuse a big-O statement with a big-Θ statement; the former implies only an upper bound. If Alice has an algorithm that is $O(2^n)$ this does not mean that Alice's algorithm requires exponential time, and it does not mean that Bob's $O(n^2)$ algorithm is better; Alice's algorithm could be $O(n)$ for all we know. But if Alice's algorithm is $\Omega(2^n)$ then we would definitely prefer to use Bob's algorithm for all sufficiently large n.

Big-O notation can also be used for multi-variable functions: "$f(m,n) = O(g(m,n))$" is shorthand for the statement:

There exist constants c and N such that for all $m, n \geq N$ we have $|f(m,n)| \leq c|g(m,n)|$.

This statement is weaker than it appears. For example, it says nothing about the relationship between $f(m,n)$ and $g(m,n)$ if we fix one of the variables. However, in virtually all of the examples we will see it will actually be true that if we regard $f(m,n) = f_m(n)$ and $g(m,n) = g_m(n)$ as functions of n with a fixed parameter m, we have $f_m(n) = O(g_m(n))$, and similarly, $f_n(m) = O(g_n(m))$. In this situation one says that $f(m,n) = O(g(m,n))$ holds *uniformly* (in m and n).

So far we have spoken only of *time complexity*, but *space complexity* plays a crucial role in many algorithms that we will see in later lectures. Space complexity measures the amount of memory an algorithm requires; this can never be greater than its time complexity (it takes time to use space), but it may be smaller. When we speak of "the complexity" of an algorithm, we should really consider both time and space. An upper bound on the time complexity is also an upper bound on the space complexity but it is often possible (and desirable) to obtain a better bound for the space complexity.

For more information on asymptotic notation and algorithmic complexity, see [5].

Warning 3.28. In this class, unless explicitly stated otherwise, our asymptotic bounds always count bit operations (as opposed to finite field operations, or integer operations). When comparing complexity bounds found in the literature, one must be sure to understand exactly what is being counted. For example, a complexity bound that counts operations in finite fields may need to be converted to a bit complexity to get an accurate comparison, and this conversion is going to depend on exactly which finite field operations are being used and how the finite fields are represented. A lack of care in this regard has led to more than one erroneous claim in the literature.

3.4 Integer multiplication

We now consider the problem of integer multiplication. Unlike addition, this is (still) an open problem; it is widely believed that $O(n \log n)$ is the best possible, and this has even been proved conditionally under various conjectures, but it is not known unconditionally, and it is only very recently that $O(n \log n)$ was established as an upper bound.

Because we do not know the exact complexity of integer multiplication, it is common practice to use the notation $\mathsf{M}(n)$ to denote the time to multiply two n-bit integers; this allows us to state bounds for algorithms that depend on the complexity of integer multiplication in a way that does not depend on whatever the current state of the art is. This convention has proved useful over the past two decades during which upper bounds on $\mathsf{M}(n)$ have improved at least four times.

3.4.1 Schoolbook method

Let us compute $37 \times 43 = 1591$ with the "schoolbook" method, using a binary representation.

$$\begin{array}{r} 101011 \\ \times\ 100101 \\ \hline 101011 \\ 101011 \\ +101011 \\ \hline 11000110111 \end{array}$$

Multiplying individual bits is easy (just use an AND gate), but we need to do n^2 bit multiplications, followed by n additions of n-bit numbers (suitably shifted). The complexity of this algorithm is thus $\Theta(n^2)$. This gives us the upper bound $\mathsf{M}(n) = O(n^2)$. The only lower bound known is the trivial one, $\mathsf{M}(n) = \Omega(n)$, so one might hope to do better than $O(n^2)$, and indeed we can.

3.4.2 Karatsuba's algorithm

Before presenting Karatsuba's algorithm, it is worth making a few remarks regarding its origin. In the first half of the twentieth century it was widely believed that $\mathsf{M}(n) = \Omega(n^2)$; indeed, no less a mathematician than Kolmogorov formally stated this conjecture in a 1956 meeting of the Moscow Mathematical Society [16, §5]. This conjecture was one of the topics at a 1960 seminar led by Kolmogorov, with Karatsuba in attendance. Within the first week of the seminar, Karatsuba was able to disprove the conjecture. Looking back on the event, Karatsuba writes [16, §6]

> *After the next seminar I told Kolmogorov about the new algorithm and about the disproof of the n^2 conjecture. Kolmogorov was very agitated because this contradicted his very plausible conjecture. At the next meeting of the seminar, Kolmogorov himself told the participants about my method and at this point the seminar was terminated.*

Karatsuba's algorithm is based on a divide-and-conquer approach. Rather than representing n-bit integers using n digits in base 2, we may instead write them in base $2^{n/2}$ and may compute their product as follows

$$\begin{aligned} a &= a_0 + 2^{n/2} a_1, \\ b &= b_0 + 2^{n/2} b_1, \\ ab &= a_0 b_0 + 2^{n/2}(a_1 b_0 + b_1 a_0) + 2^n a_1 b_1, \end{aligned}$$

As written, this reduces an n-bit multiplication to four multiplications of $(n/2)$-bit integers and three additions of $O(n)$-bit integers (multiplying an intermediate result by a power of 2 can be achieved by simply writing the binary output "further to the left" and is effectively free). However, as observed by Karatsuba one can use the identity

$$a_0 b_1 + b_0 a_1 = (a_0 + a_1)(b_0 + b_1) - a_0 b_0 - a_1 b_1$$

to compute $a_0 b_1 + b_0 a_1$ using just one multiplication in addition to computing the products $a_0 b_0$ and $a_1 b_1$. By reusing the common subexpressions $a_0 b_0$ and $a_1 b_1$, we can compute ab

using three multiplications and six additions (we count subtractions as additions). We can use the same idea to recursively compute the three products a_0b_0, a_1b_1, and $(a_0+a_1)(b_0+b_1)$; this recursive approach yields Karatsuba's algorithm.

If we let $T(n)$ denote the running time of this algorithm, we have

$$\begin{aligned} T(n) &= 3T(n/2) + O(n) \\ &= O(n^{\lg 3}) \end{aligned}$$

It follows that $\mathsf{M}(n) = O(n^{\lg 3})$, where $\lg 3 := \log_2 3 \approx 1.59$.[3]

3.4.3 The Fast Fourier Transform (FFT)

The fast Fourier transform is widely regarded as one of the top ten algorithms of the twentieth century [6, 10], and has applications throughout applied mathematics. Here we focus on the discrete Fourier transform (DFT), and its application to multiplying integers and polynomials, following the presentation in [9, §8]. It is actually more natural to address the problem of polynomial multiplication first.

Let R be a commutative ring containing a primitive nth root of unity ω, by which we mean that $\omega^n = 1$ and $\omega^i - \omega^j$ is not a zero divisor for $0 \le i < j < n$ (when R is a field this coincides with the usual definition). We shall identify the set of polynomials in $R[x]$ of degree less than n with the set of all n-tuples with entries in R. Thus we represent the polynomial $f(x) = \sum_{i=0}^{n-1} f_i x^i$ by its coefficient vector $(f_0, \dots, f_{n-1}) \in R^n$ and may speak of the polynomial $f \in R[x]$ and the vector $f \in R^n$ interchangeably.

The discrete Fourier transform $\mathrm{DFT}_\omega : R^n \to R^n$ is the R-linear map

$$(f_0, \dots, f_{n-1}) \xrightarrow{\mathrm{DFT}_\omega} (f(\omega^0), \dots, f(\omega^{n-1})).$$

You should think of this map as a conversion between two types of polynomial representations: we take a polynomial of degree less than n represented by n coefficients (its *coefficient-representation*) and convert it to a representation that gives its values at n known points (its *point-representation*).

One can use Lagrange interpolation to recover the coefficient representation from the point representation, but our decision to use values $\omega^0, \dots, \omega^{n-1}$ that are nth roots of unity allows us to do this more efficiently. If we define the Vandermonde matrix

$$V_\omega := \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2n-2} \\ 1 & \omega^3 & \omega^6 & \cdots & \omega^{3n-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \omega^{n-1} & \omega^{2n-2} & \cdots & \omega^{(n-1)^2} \end{pmatrix},$$

then $\mathrm{DFT}_\omega(f) = V_\omega f^{\mathrm{tr}}$. Our assumption that none of the differences $\omega^i - \omega^j$ is a zero divisor in R ensures that the matrix V_ω is invertible, and its inverse is simply $\frac{1}{n} V_{\omega^{-1}}$. It follows that

$$\mathrm{DFT}_\omega^{-1} = \frac{1}{n} \mathrm{DFT}_{\omega^{-1}}.$$

[3] In general we shall use $\lg n$ to denote $\log_2 n$.

Thus if we have an algorithm to compute DFT_ω we can use it to compute DFT_ω^{-1}: just replace ω by $\omega^{-1} = \omega^{n-1}$ and multiply the result by $\frac{1}{n}$.

We now define the *cyclic convolution* $f * g$ of two polynomials $f, g \in R^n$:

$$f * g = fg \bmod (x^n - 1).$$

Reducing the product on the right modulo $x^n - 1$ ensures that $f * g$ is a polynomial of degree less than n, thus we may regard the cyclic convolution as a map $R^n \times R^n \to R^n$. If $h = f * g$, then $h_i = \sum f_j g_k$, where the sum is over $j + k \equiv i \bmod n$. If f and g both have degree less than $n/2$, then $f * g = fg$; thus the cyclic convolution of f and g can be used to compute their product, provided that we make n big enough.

We also define the *pointwise product* $f \cdot g$ of two vectors in $f, g \in R^n$:

$$f \cdot g = (f_0 g_0,\ f_1 g_1,\ \ldots,\ f_{n-1} g_{n-1}).$$

We have now defined three operations on vectors in R^n: the binary operations of convolution and pointwise product, and the unary operation DFT_ω. The following theorem relates these three operations and is the key to the fast Fourier transform.

Theorem 3.29. $\mathrm{DFT}_\omega(f * g) = \mathrm{DFT}_\omega(f) \cdot \mathrm{DFT}_\omega(g)$.

Proof. Since $f * g = fg \bmod (x^n - 1)$, we have

$$f * g = fg + q \cdot (x^n - 1)$$

for some polynomial $q \in R[x]$. For every integer i from 0 to $n - 1$ we then have

$$\begin{aligned}(f * g)(\omega^i) &= f(\omega^i) g(\omega^i) + q(\omega^i)(\omega^{in} - 1)\\ &= f(\omega^i) g(\omega^i),\end{aligned}$$

where we have used $(\omega^{in} - 1) = 0$, since ω is an nth root of unity. □

The theorem implies that if f and g are polynomials of degree less then $n/2$ then

$$fg = f * g = \mathrm{DFT}_\omega^{-1}\big(\mathrm{DFT}_\omega(f) \cdot \mathrm{DFT}_\omega(g)\big). \tag{1}$$

This identity allows us to multiply polynomials using the discrete Fourier transform. In order to put this into practice, we need an efficient way to compute DFT_ω. This is achieved by the following recursive algorithm.

Algorithm: Fast Fourier Transform (FFT)
Input: A positive integer $n = 2^k$, a vector $f \in R^n$, and the vector $(\omega^0, \ldots, \omega^{n-1}) \in R^n$.
Output: $\mathrm{DFT}_\omega(f) \in R^n$.

1. If $n = 1$ then return (f_0) and terminate.
2. Write the polynomial $f(x)$ in the form $f(x) = g(x) + x^{\frac{n}{2}} h(x)$, where $g, h \in R^{\frac{n}{2}}$.
3. Compute the vectors $r = g + h$ and $s = (g - h) \cdot (\omega^0, \ldots, \omega^{\frac{n}{2}-1})$ in $R^{\frac{n}{2}}$.
4. Recursively compute $\mathrm{DFT}_{\omega^2}(r)$ and $\mathrm{DFT}_{\omega^2}(s)$ using $(\omega^0, \omega^2, \ldots, \omega^{n-2})$.
5. Return the vector $(r(\omega^0), s(\omega^0), r(\omega^2), s(\omega^2), \ldots, r(\omega^{n-2}), s(\omega^{n-2}))$

Let $T(n)$ be the number of operations in R used by the FFT algorithm. Then

$$\begin{aligned} T(n) &= 2T(n/2) + O(n) \\ &= O(n \log n). \end{aligned}$$

This shows that the FFT is fast (justifying its name); let us now prove that it is correct.

Theorem 3.30. *The FFT algorithm outputs* $\mathrm{DFT}_\omega(f)$.

Proof. We must verify that the kth entry of the output vector is $f(\omega^k)$, for $0 \le k < n$. For even $k = 2i$ we have:

$$\begin{aligned} f(\omega^{2i}) &= g(\omega^{2i}) + (\omega^{2i})^{n/2} h(\omega^{2i}) \\ &= g(\omega^{2i}) + h(\omega^{2i}) \\ &= r(\omega^{2i}). \end{aligned}$$

For odd $k = 2i + 1$ we have:

$$\begin{aligned} f(\omega^{2i+1}) &= \sum_{0 \le j < n/2} f_j \omega^{(2i+1)j} + \sum_{0 \le j < n/2} f_{n/2+j} \omega^{(2i+1)(n/2+j)} \\ &= \sum_{0 \le j < n/2} g_j \omega^{2ij} \omega^j + \sum_{0 \le j < n/2} h_j \omega^{2ij} \omega^{in} \omega^{n/2} \omega^j \\ &= \sum_{0 \le j < n/2} (g_j - h_j) \omega^j \omega^{2ij} \\ &= \sum_{0 \le j < n/2} s_j \omega^{2ij} \\ &= s(\omega^{2i}), \end{aligned}$$

where we have used the fact that $\omega^{n/2} = -1$. □

Corollary 3.31. *Let R be a commutative ring containing a primitive nth root of unity, with $n = 2^k$, and assume $2 \in R^\times$. We can multiply two polynomials in $R[x]$ of degree less than $n/2$ using $O(n \log n)$ operations in R.*

Proof. From (1) we have

$$fg = \mathrm{DFT}_\omega^{-1}(\mathrm{DFT}_\omega(f) \cdot \mathrm{DFT}_\omega(g)) = \frac{1}{n} \mathrm{DFT}_{\omega^{-1}}(\mathrm{DFT}_\omega(f) \cdot \mathrm{DFT}_\omega(g))$$

and we note that $n = 2^k \in R^\times$ is invertible. We can compute $\omega^0, \dots, \omega^{n-1}$ using $O(n)$ multiplications in R (this also gives us $(\omega^{-1})^0, \dots, (\omega^{-1})^{n-1}$). Computing DFT_ω and $\mathrm{DFT}_{\omega^{-1}}$ via the FFT algorithm uses $O(n \log n)$ operations in R, computing the pointwise product of $\mathrm{DFT}_\omega(f)$ and $\mathrm{DFT}_\omega(g)$ uses $O(n)$ operations in R, and computing $1/n$ and multiplying a polynomial of degree less than n by this scalar uses $O(n)$ operations in R. □

What about rings that do not contain an nth root of unity? By extending R to a new ring $R' := R[\omega]/(\omega^n - 1)$ we can obtain a formal nth root of unity ω, and one can then generalize Corollary 3.31 to multiply polynomials in any ring R in which 2 is invertible using $O(n \log n \log \log n)$ operations in R; see [9, §8.3] for details.

The need for 2 to be invertible can be overcome by considering a 3-adic version of the FFT algorithm that works in rings R in which 3 is invertible. For rings in which neither 2 nor 3 is invertible we instead compute $2^k fg$ and $3^m fg$ (just leave out the multiplication by $1/n$ at the end). Once we know both $2^k fg$ and $3^m fg$ we can recover the coefficients of fg by using the Euclidean algorithm to compute $u, v \in \mathbb{Z}$ such that $u2^k + v3^m = 1$ and applying $u2^k fg + v3^m fg = fg$.

3.5 Integer multiplication

To any positive integer $a = \sum_{i=0}^{n-1} a_i 2^i$ we may associate the polynomial $f_a(x) = \sum_{i=0}^{n} a_i x^i \in \mathbb{Z}[x]$, with $a_i \in \{0, 1\}$, so that $a = f_a(2)$. We can then multiply positive integers a and b via

$$ab = f_{ab}(2) = (f_a f_b)(2).$$

Note that the polynomials $f_a(x)f_b(x)$ and $f_{ab}(x)$ may differ (the former may have coefficients greater than 1), but they take the same value at $x = 2$; in practice one typically uses base 2^{64} rather than base 2 (the a_i and b_i are then integers in $[0, 2^{64} - 1]$).

Applying the generalization of Corollary 3.31 noted above to the ring $\mathbb{Z}$, Schönhage and Strassen [19] obtain an algorithm to multiply two n-bit integers in time $O(n \log n \log\log n)$, which gives us a new upper bound

$$\mathsf{M}(n) = O(n \log n \log\log n).$$

In 2007 Fürer [7] showed that this bound can been improved to

$$\mathsf{M}(n) = O\left(n \log n \, 2^{O(\log^* n)}\right)$$

where $\log^* n$ denotes the iterated logarithm, which counts how many times the log function must be applied to n before the result is less than or equal to 1. In 2016 Harvey, van der Hoeven and Lecerf [15] proved the sharper bound

$$\mathsf{M}(n) = O\left(n \log n \, 8^{\log^* n}\right),$$

and in 2018 Harvey and van der Hoeven [12] further improved this to

$$\mathsf{M}(n) = O\left(n \log n \, 4^{\log^* n}\right).$$

In 2019 Harvey and van der Hoeven [14] announced the spectacular and long awaited result

$$\mathsf{M}(n) = O\left(n \log n\right),$$

which as far as asymptotics go, is almost certainly the final word on the matter.

The algorithms that enabled these improvements and even the original Schönhage–Strassen algorithm are fairly intricate and purely of theoretical interest: in practice one uses the "three primes" algorithm sketched below, which for integers with $n \leq 2^{62}$ bits has a "practical complexity" of $O(n \log n)$; this statement is mathematically meaningless but gives a rough indication of how the running time increases as n varies in this bounded range. But it is a great relief and convenience to know that the theoretical complexity now matches the practical complexity, and that we can dispense with the "$\log\log n$" term you will find in almost any literature that mentions the complexity of integer multiplication prior to 2020.

3.5.1 Three primes FFT for integer multiplication

As noted above, the details of the Schoönhage and Strassen algorithm and its subsequent improvements are rather involved. There is a much simpler approach that is used in practice to multiply integers less than $2^{2^{62}}$; this includes integers that would require 500 petabytes (500,000 terabytes) to write down and is more than enough for any practical application that is likely to arise in the near future. Let us briefly outline this approach.

Write the positive integers $a, b < 2^{2^{62}}$ that we wish to multiply in base 2^{64} as $a = \sum a_i 2^{64i}$ and $b = \sum b_i 2^{64i}$, with $0 \le a_i, b_i < 2^{64}$, and define the polynomials $f_a = \sum a_i x^i \in \mathbb{Z}[x]$ and $f_b = \sum b_i x^i \in \mathbb{Z}[x]$ as above. Our goal is to compute $f_{ab}(2^{64}) = (f_a f_b)(2^{64})$, and we note that the polynomial $f_a f_b \in \mathbb{Z}[x]$ has less than $2^{62}/64 = 2^{56}$ coefficients, each of which is bounded by $2^{56}2^{64}2^{64} < 2^{184}$.

Rather than working over a single ring R we will use three finite fields $\mathbb{F}_p$ of odd characteristic, where p is one of the primes

$$p_1 := 71 \cdot 2^{57} + 1, \qquad p_2 := 75 \cdot 2^{57} + 1, \qquad p_3 := 95 \cdot 2^{57} + 1.$$

Note that if p is any of the primes p_1, p_2, p_3, then $\mathbb{F}_p^\times$ is a cyclic group whose order $p-1$ is divisible by 2^{57}, which implies that $\mathbb{F}_p$ contains a primitive 2^{57}th root of unity ω; indeed, for $p = p_1, p_2, p_3$ we can use $\omega = \omega_1, \omega_2, \omega_3$, respectively, where $\omega_1 = 287, \omega_2 = 149, \omega_3 = 55$.

We can thus use the FFT Algorithm above with $R = \mathbb{F}_p$ to compute $f_a f_b \bmod p$ for each of the primes $p \in \{p_1, p_2, p_3\}$. This gives us the values of the coefficients of $f_a f_b \in \mathbb{Z}[x]$ modulo three primes whose product $p_1 p_2 p_3 > 2^{189}$ is more than large enough to uniquely recover the coefficients via the Chinese Remainder Theorem (CRT); the time to recover the integer coefficients of $f_a f_b$ from their values modulo p_1, p_2, p_3 is negligible compared to the time to apply the FFT algorithm over these three fields. If a and b are significantly smaller, say $a, b \le 2^{2^{44}}$, a "one prime" approach suffices.

3.6 Kronecker substitution

We now note an important converse to the idea of using polynomial multiplication to multiply integers: we can use integer multiplication to multiply polynomials. This is quite useful in practice, as it allows us take advantage of very fast implementations of FFT–based integer multiplication that are now widely available. If f is a polynomial in $\mathbb{F}_p[x]$, we can lift f to $\hat{f} \in \mathbb{Z}[x]$ by representing its coefficients as integers in $[0, p-1]$. If we then consider the integer $\hat{f}(2^m)$, where $m = \lceil 2 \lg p + \lg(\deg f + 1) \rceil$, the coefficients of $\hat{f}$ will appear in the binary representation of $\hat{f}(2^m)$ separated by blocks of $m - \lceil \lg p \rceil$ zeros. If g is a polynomial of similar degree, we can easily recover the coefficients of $\hat{h} = \hat{f}\hat{g} \in \mathbb{Z}[x]$ in the integer product $N = \hat{f}(2^m)\hat{g}(2^m)$; we then reduce the coefficients of $\hat{h}$ modulo p to get $h = fg$. The key is to make m large enough so that the kth block of m binary digits in N contains the binary representation of the kth coefficient of $\hat{h}$.

This technique is known as *Kronecker substitution*, and it allows us to multiply two polynomials of degree d in $\mathbb{F}_p[x]$ in time $O(\mathsf{M}(d(n + \log d)))$, where $n = \log p$. Typically we have $\log d = O(n)$, in which case this simplifies to $O(\mathsf{M}(dn))$ In particular, we can use Kronecker substitution to multiply elements of $\mathbb{F}_q \simeq \mathbb{F}_p[x]/(f)$ in time $O(\mathsf{M}(n))$, where $n = \log q$, provided $\log \deg f = O(\log p)$.

Remark 3.32. When $\log d = O(n)$, if we make the standard assumption that $\mathsf{M}(n)$ grows super-linearly then using Kronecker substitution is strictly faster (by more than any constant

factor) than a layered approach that uses the FFT to multiply polynomials and then recursively uses the FFT for the coefficient multiplications; this is because $\mathsf{M}(dn) = o(\mathsf{M}(d)\,\mathsf{M}(n))$.

3.7 Euclidean division

Given integers $a, b > 0$, we wish to compute the unique integers $q, r \geq 0$ for which

$$a = bq + r \qquad (0 \leq r < b).$$

We have $q = \lfloor a/b \rfloor$ and $r = a \bmod b$. It is enough to compute q, since we can then compute $r = a - bq$. To compute q, we determine a sufficiently precise approximation $c \approx 1/b$ and obtain q by computing ca and rounding down to the nearest integer.

We recall Newton's method for finding the root of a real-valued function $f(x)$. We start with an initial approximation x_0, and at each step, we refine the approximation x_i by computing the x-coordinate x_{i+1} of the point where the tangent line through $(x_i, f(x_i))$ intersects the x-axis, via

$$x_{i+1} := x_i - \frac{f(x_i)}{f'(x_i)}.$$

To compute $c \approx 1/b$, we apply this to $f(x) = 1/x - b$, using the Newton iteration

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} = x_i - \frac{\frac{1}{x_i} - b}{-\frac{1}{x_i^2}} = 2x_i - bx_i^2.$$

As an example, let us approximate $1/b = 1/123456789$. For the sake of illustration we work in base 10, but in an actual implementation would use base 2, or base 2^w, where w is the word size.

$$\begin{aligned}
x_0 &= 1 \times 10^{-8} \\
x_1 &= 2(1 \times 10^{-8}) - (1.2 \times 10^{8})(1 \times 10^{-8})^2 \\
&= 0.80 \times 10^{-8} \\
x_2 &= 2(0.80 \times 10^{-8}) - (1.234 \times 10^{8})(0.80 \times 10^{-8})^2 \\
&= 0.8102 \times 10^{-8} \\
x_3 &= 2(0.8102 \times 10^{-8}) - (1.2345678 \times 10^{8})(0.8102 \times 10^{-8})^2 \\
&= 0.81000002 \times 10^{-8}.
\end{aligned}$$

Note that we double the precision we are using at each step, and each x_i is correct up to an error in its last decimal place. The value x_3 suffices to correctly compute $\lfloor a/b \rfloor$ for $a \leq 10^{15}$.

To analyze the complexity of this approach, let us assume that b has n bits and a has at most $2n$ bits; this is precisely the situation we will encounter when we wish to reduce the product of two integers in $[0, p-1]$ modulo p. During the Newton iteration to compute $c \approx 1/b$, the size of the integers involved doubles with each step, and the cost of the arithmetic operations grows at least linearly. The total cost is thus at most twice the cost of the last step, which is $\mathsf{M}(n) + O(n)$; note that all operations can be performed using integers by shifting the operands appropriately. Thus we can compute $c \approx 1/b$ in time $2\,\mathsf{M}(n) + O(n)$. We can then compute $ca \approx a/b$, round to the nearest integer, and compute $r = a - bq$ using at most $4\,\mathsf{M}(n) + O(n)$ bit operations.

With a slightly more sophisticated version of this approach it is possible to compute r in time $3\,\mathsf{M}(n) + O(n)$, and if we expect to repeatedly perform Euclidean division with the

same denominator we can further reduce this to $2\,\mathsf{M}(n) + O(n)$ by precomputing $c \approx 1/b$. This approach is exploited by two widely used approaches to modular arithmetic, *Barrett reduction* (see [4, Alg. 10.17]) and *Montgomery reduction* (see Problem Set 1). Regardless of the approach taken, we obtain the following bound for multiplication in $\mathbb{F}_p$ using our standard representation as integers in $[0, p-1]$.

Theorem 3.33. *The time to multiply two elements of $\mathbb{F}_p$ is $O(\mathsf{M}(n))$, where $n = \lg p$.*

There is an analogous version of this algorithm above for polynomials that uses the exact same Newton iteration $x_{i+1} = 2x_i - bx_i^2$, where b and the x_i are now polynomials. Rather than working with Laurent polynomials (the polynomial version of approximating a rational number with a truncated decimal expansion), it is simpler to reverse the polynomials and work modulo a sufficiently large power of x, doubling the power of x with each Newton iteration. More precisely, we have the following algorithm, which combines Algorithms 9.3 and 9.5 from [9]. For any polynomial $f(x)$ we write $\operatorname{rev} f$ for the polynomial $x^{\deg f} f(\frac{1}{x})$; this simply reverses the coefficients of f.

Algorithm 3.34 (Fast Euclidean division of polynomials)**.** Given $a, b \in \mathbb{F}_p[x]$ with b monic, compute $q, r \in \mathbb{F}_p[x]$ such that $a = qb + r$ with $\deg r < \deg b$ as follows:

1. If $\deg a < \deg b$ then return $q = 0$ and $r = a$.
2. Let $m = \deg a - \deg b$ and $k = \lceil \lg m + 1 \rceil$.
3. Let $f = \operatorname{rev}(b)$ (reverse the coefficients of b).
4. Compute $g_0 = 1, g_i = (2g_{i-1} - fg_{i-1}^2) \bmod x^{2^i}$ for i from 1 to k.
   (this yields $fg_k \equiv 1 \bmod x^{m+1}$).
5. Set $s = \operatorname{rev}(a) g_k \bmod x^{m+1}$ (now $\operatorname{rev}(b)s \equiv \operatorname{rev}(a) \bmod x^{m+1}$).
6. Return $q = x^{m-\deg s} \operatorname{rev}(s)$ and $r = a - bq$.

As in the integer case, the work is dominated by the last iteration in step 4, which involves multiplying polynomials in $\mathbb{F}_p[x]$. To multiply elements of $\mathbb{F}_q \simeq \mathbb{F}_p[x]/(f)$ represented as polynomials of degree less than $d = \deg f$, we compute the product a in $\mathbb{F}[x]$ and then reduce modulo $b = f$, and the degree of the polynomials involved are all $O(d)$. With Kronecker substitution, we can reduce these polynomial multiplications to integer multiplications, and obtain the following result.

Theorem 3.35. *Let $q = p^e$ be a prime power, and assume that $\log e = O(\log p)$. The time to multiply two elements of $\mathbb{F}_q$ is $O(\mathsf{M}(n)) = O(n \log n)$, where $n = \lg q$.*

Remark 3.36. The constraints on the relative growth rates of p and e in the theorem above are present only so that we can conveniently use Kronecker substitution to bound the complexity in terms of the bound $\mathsf{M}(n)$ for multiplying integers. In fact we fully expect that the $O(n \log n)$ bound implied by Theorem 3.35 holds uniformly. This is known under a widely believed conjecture about the least prime in arithmetic progressions, namely that the least prime in every arithmetic progression $m\mathbb{Z} + a$ with a coprime to m is bounded by $O(m^{1+\epsilon})$ for any $\epsilon > 0$ (in fact any $\epsilon < 2^{-1162}$ would do); see [13].

Before leaving the topic of Euclidean division, we should also mention the standard "schoolbook" algorithm of long division. The classical algorithm works with decimal digits (base 10), but for the sake of simplicity let us work in base 2; in practice one works in base 2^w for some fixed w.

Algorithm 3.37 (Long division)**.** Given positive integers $a = \sum_{i=0}^{m} a_i 2^i$ and $b = \sum_{i=0}^{n} b_i 2^i$, compute $q, r \in \mathbb{Z}$ such that $a = qb + r$ with $0 \le r < b$ as follows:

1. If $b > a$ return $q = 0$ and $r = a$, and if $b = 1$ return $q = a$ and $r = 0$.
2. Set $q \leftarrow 0$, $r \leftarrow 0$, and $k \leftarrow m$.
3. While $k \ge 0$ and $r < b$ set $q \leftarrow 2q$, $r \leftarrow 2r + a_k$, and $k \leftarrow k - 1$.
4. If $r < b$ then return q and r.
5. Set $q \leftarrow q + 1$, $r \leftarrow r - b$, and return to Step 3.

The net effect of all the executions of Step 3 is is to add a to $qb+r$ using double-and-add bitwise addition. The quantity $qb + r$ is initially set to 0 in Step 2 and is unchanged by Step 5, so when the algorithm terminates in Step 4 we have $a = qb + r$ and $0 \le r < b$ as desired. If we are only interested in the remainder r we can omit all operations involving q.

For the complexity analysis we can assume that multiplication by 2 is achieved by bit-shifting and costs $O(1)$ (consider a multi-tape Turing machine, or a bit-addressable RAM). Step 2 costs $O(1)$, the total cost of Step 3 over all iterations is $O(nm)$, as is the total cost of Step 5 (note that q is a multiple of 2 at the start of Step 5, so computing $q \leftarrow q + 1$ is achieved by setting the least significant bit). This yields the following result.

Theorem 3.38. *The long division algorithm uses $O(mn)$ bit operations to perform Euclidean division of an m-bit integer by an n-bit integer.*

Remark 3.39. For $m = O(n)$ the $O(n^2)$ complexity of long division is worse than the $O(\mathsf{M}(n))$ cost of Euclidean division using Newton iteration. But when m is much larger than n, say $n = O(\log m)$ or $n = O(1)$, long division is a better choice. In particular, for any fixed prime p (so $O(1)$ bits) we can reduce n-bit integers modulo p in linear time.

3.8 Extended Euclidean algorithm

We recall the Euclidean algorithm for computing the greatest common divisor of positive integers a and b. For $a > b$ we repeatedly apply

$$\gcd(a, b) = \gcd(b, a \bmod b),$$

where we take $a \bmod b$ to be the unique integer $r \in [0, b-1]$ congruent to a modulo b.

To compute the multiplicative inverse of an integer modulo a prime, we use the extended Euclidean algorithm, which expresses $\gcd(a, b)$ as a linear combination

$$\gcd(a, b) = as + bt,$$

with $|s| \le b/\gcd(a, b)$ and $|t| \le a/\gcd(a, b)$. If a is prime, we obtain $as + bt = 1$, and t is the inverse of b modulo a. To compute the integers s and t we use the following algorithm. First, let

$$R_1 = \begin{bmatrix} a \\ b \end{bmatrix}, \quad S_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad T_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and note that $R_1 = aS_1 + bT_1$. We then compute

$$Q_i = \begin{bmatrix} 0 & 1 \\ 1 & -q_i \end{bmatrix}, \quad R_{i+1} = Q_i R_i, \quad S_{i+1} = Q_i S_i, \quad T_{i+1} = Q_i T_i,$$

where q_i is the quotient $\lfloor R_{i,1}/R_{i,2} \rfloor$ obtained via Euclidean division. Note that applying the linear transformation Q_i to both sides of $R_i = aS_i + bT_i$ ensures $R_{i+1} = aS_{i+1} + bT_{i+1}$. The algorithm terminates in the kth step where $R_{k,2}$ becomes zero, at which point we have

$$R_k = \begin{bmatrix} d \\ 0 \end{bmatrix}, \quad S_k = \begin{bmatrix} s \\ \pm b \end{bmatrix}, \quad T_k = \begin{bmatrix} t \\ \mp a \end{bmatrix},$$

with $\gcd(a, b) = d = sa + tb$. As an example, with $a = 1009$ and $b = 789$ we have

r	q	s	t
1009		1	0
789	1	0	1
220	3	1	−1
129	1	−3	4
91	1	4	−5
38	2	−7	9
15	2	18	−23
8	1	−43	55
7	1	61	−78
1	7	−104	133
0		789	−1009

From the second-to-last line with $s = -104$ and $t = 133$ we see that

$$1 = -104 \cdot 1009 + 133 \cdot 789,$$

and therefore 133 is the inverse of 789 modulo 1009 (and $-104 \equiv 685$ is the inverse of 1009 modulo 789).

It is clear that r is reduced by a factor of at least 2 every two steps, thus the total number of iterations is $O(n)$, and each step involves Euclidean division, whose cost is bounded by $O(\mathsf{M}(n))$. This yields a complexity of $O(n\,\mathsf{M}(n))$, but a more careful analysis shows that it is actually $O(n^2)$, even if schoolbook multiplication is used (the key point is that the total size of all the q_i is $O(n)$ bits).

This can be further improved using the *fast Euclidean algorithm*, which uses a divide-and-conquer approach to compute the product $Q = Q_{k-1} \cdots Q_1$ by splitting the product in half and recursively computing each half using what is known as a *half-gcd* algorithm. One can then compute $R_k = QR_1$, $S_k = QS_1$, and $T_k = QT_1$. The details are somewhat involved (care must be taken when determining how to split the product in a way that balances the work evenly), but this yields a recursive running time of

$$T(n) = 2T(n/2) + O(\mathsf{M}(n)) = O(\mathsf{M}(n) \log n);$$

see [9, §11] for details.

Theorem 3.40. *Let p be a prime. The time to invert an element of $\mathbb{F}_p^\times$ is $O(\mathsf{M}(n) \log n)$, where $n = \lg p$.*

The extended Euclidean algorithm works in any Euclidean ring, that is, a ring with a norm function that allows us to use Euclidean division to write $a = qb + r$ with r of norm strictly less than b (for any nonzero b). This includes polynomial rings, in which the norm of

a polynomial is simply its degree. Thus we can compute the inverse of a polynomial modulo another polynomial, provided the two polynomials are relatively prime.

One issue that arises when working in Euclidean rings other than $\mathbb{Z}$ is that there may be units (invertible elements) other than ± 1, and the gcd is only defined up to a unit. In the case of the polynomial ring $\mathbb{F}_p[x]$, every element of $\mathbb{F}_p^\times$ is a unit, and with the fast Euclidean algorithm in $\mathbb{F}_p[x]$ one typically normalizes the intermediate results by making the polynomials monic at each step; this involves computing the inverse of the leading coefficient in $\mathbb{F}_p$. If $\mathbb{F}_q = \mathbb{F}_p[x]/(f)$ with $\deg f = d$, one can then bound the time to compute an inverse in $\mathbb{F}_q$ by $O(\mathsf{M}(d)\log d)$, operations in $\mathbb{F}_p$, of which $O(d)$ are inversions; see [9, Thm. 11.10(i)]. This gives a bit complexity of

$$O(\mathsf{M}(d)\,\mathsf{M}(\log p)\log d + d\,\mathsf{M}(\log p)\log\log p),$$

but with Kronecker substitution we can sharpen this to

$$O(\mathsf{M}(d(\log p + \log d))\log d + d\,\mathsf{M}(\log p)\log\log p).$$

We will typically assume that either $\log d = O(\log p)$ (large characteristic) or $\log p = O(1)$ (small characteristic); in both cases we can simplify this bound to $O(\mathsf{M}(n)\log n)$, where $n = \lg q = d\lg p$ is the number of bits in q, the same result we obtained for the case where $q = p$ is prime.

Theorem 3.41. *Let $q = p^e$ be a prime power and assume $\log e = O(\log p)$. The time to invert an element of $\mathbb{F}_q^\times$ is $O(\mathsf{M}(n)\log n) = O(n\log^2 n)$, where $n = \lg q$.*

Remark 3.42. As with Theorem 3.35, the assumption $\log e = O(\log p)$ can be removed if one assumes the least prime in every arithmetic progression $m\mathbb{Z} + a$ with a coprime to m is bounded by $O(m^{1+\epsilon})$ for any $\epsilon > 0$.

3.9 Exponentiation (scalar multiplication)

Let a be a positive integer. In a multiplicative group, the computation

$$g^a = \underbrace{gg\cdots g}_{a}$$

is known as *exponentiation*. In an additive group, this is equivalent to

$$ag = \underbrace{g + g + \cdots + g}_{a},$$

and is called *scalar multiplication*. The same algorithms are used in both cases, and most of these algorithms were first developed in a multiplicative setting (the multiplicative group of a finite field) and are called exponentiation algorithms. It is actually more convenient to describe the algorithms using additive notation (fewer superscripts), so we will do so.

The oldest and most commonly used exponentiation algorithm is the "double-and-add" method, also known as left-to-right binary exponentiation. Given an element P of an additive group and a positive integer a with binary representation $a = \sum 2^i a_i$, we compute the scalar multiple $Q = aP$ as follows:

```
def DoubleAndAdd (P,a):
    a=a.digits(2); n=len(a)       # represent a in binary using n bits
    Q=P;                          # start 1 bit below the high bit
    for i in range(n-2,-1,-1):    # for i from n-2 down to 0
        Q += Q                    # double
        if a[i]==1: Q += P        # add
    return Q
```

Alternatively, we may use the "add-and-double" method, also known as right-to-left binary exponentiation.

```
def AddAndDouble (P,a):
    a=a.digits(2); n=len(a)       # represent a in binary using n bits
    Q=0; R=P;                     # start with the low bit
    for i in range(n-1):
        if a[i]==1: Q += R        # add
        R += R                    # double
    Q += R                                              # last add
    return Q
```

The number of group operations required is effectively the same for both algorithms. If we ignore the first addition in the `add_and_double` algorithm (which could be replaced by an assignment, since initially $Q = 0$), both algorithms use precisely

$$n + \mathrm{wt}(a) - 2 \leq 2n - 2 = O(n)$$

group operations, where $\mathrm{wt}(a) = \#\{a_i : a_i = 1\}$ is the *Hamming weight* of a, the number of 1's in its binary representation. Up to the constant factor 2, this is asymptotically optimal, and it implies that exponentiation in a finite field $\mathbb{F}_q$ has complexity $O(n\,\mathsf{M}(n))$ with $n = \lg q$; this assumes the exponent is less than q, but note that we can always reduce the exponent modulo $q - 1$, the order of the cyclic group $\mathbb{F}_q^\times$. Provided the bit-size of the exponent is $O(n^2)$, the $O(\mathsf{M}(n^2))$ time to reduce the exponent modulo $q - 1$ will be majorized by the $O(n\,\mathsf{M}(n))$ time to perform the exponentiation.

Notwithstanding the fact that the simple double-and-add algorithm is within a factor of 2 of the best possible, researchers have gone to great lengths to eliminate this factor of 2, and to take advantage of situations where either the base or the exponent is fixed, and there are a wide variety of optimizations that are used in practice; see [4, Ch. 9] and [11]. Here we give just one example, windowed exponentiation, which is able to reduce the constant factor from 2 to an essentially optimal $1 + o(1)$.

3.9.1 Fixed-window exponentiation

Let the positive integer s be a *window size* and write a as

$$a = \sum a_i 2^{si}, \quad (0 \leq a_i < 2^s).$$

This is equivalent to writing a in base 2^s. With fixed-window exponentiation, one first precomputes multiples dP for each of the "digits" $d \in [0, 2^s - 1]$ that may appear in the base-2^s expansion of a. One then uses a left-to-right approach as in the double-and-add algorithm, except now we double s times and add the appropriate multiple $a_i P$.

```
def FixedWindow (P,a,s):
    a=a.digits(2^s); n=len(a)                    # write a in base 2^s
```

```
R = [0*P,P]
for i in range(2,2^s): R.append(R[-1]+P)    # precompute digits
Q = R[a[-1]]                                # copy the top digit
for i in range(n-2,-1,-1):
    for j in range(0,s): Q += Q             # double s times
    Q += R[a[i]]                            # add the next digit
return Q
```

In the algorithm above we precompute multiples of P for every possible digit that might occur. As an optimization one could examine the base-2^s representation of a and only precompute the multiples of P that are actually needed.

Let n be the number of bits in a and let $m = \lceil n/s \rceil$ be the number of base-2^s digits a_i. The precomputation step uses $2^s - 2$ additions (we get $0P$ and $1P$ for free), there are $m - 1$ additions of multiples of P corresponding to digits a_i (when $a_i = 0$ these cost nothing), and there are a total of $(m-1)s$ doublings. This yields an upper bound of

$$2^s - 2 + m - 1 + (m-1)s \approx 2^s + n/s + n$$

group operations. If we choose $s = \lg n - \lg\lg n$, we obtain the bound

$$n/\lg n + n/(\lg n - \lg\lg n) + n = n + O(n/\log n),$$

which is $(1 + o(1))n$ group operations.

3.9.2 Sliding-window exponentiation

The sliding-window algorithm modifies the fixed-window algorithm by "sliding" over blocks of 0s in the binary representation of a. There is still a window size s, but a is no longer treated as an integer written in a fixed base 2^s. Instead, the algorithm scans the bits of the exponent from left to right, assembling "digits" of at most s bits with both high and low bits set: with a sliding window of size 3 the bit-string 110011010101100 could be broken up as 11|00|11|0|101|0|11|00 with 4 nonzero digits, whereas a fixed window approach would use 110|011|010|101|100 with 5 nonzero digits. This improves the fixed-window approach in two ways: first, it is only necessarily to precompute odd digits, and second, depending on the pattern of bits in a, sliding over the zeros may reduce the number of digits used, as in the example above. In any case, the sliding-window approach is never worse than the fixed-window approach, and for $s > 2$ it is always better.

Example 3.43. Let $a = 26284$ corresponding to the bit-string 110011010101100 above. To compute aP using a sliding window approach with $s = 3$ one would first compute $2P, 3P, 5P$ using 3 additions and then

$$aP = 2^2 \cdot (2^3 \cdot (2^4 \cdot (2^4 \cdot (3P) + 3P)) + 5P) + 3P)$$

using 3 additions and 13 doublings, for a total cost of 19 group operations. A fixed window approach with $s = 3$ would instead compute $2P, 3P, 4P, 5P, 6P$ using 5 additions and

$$aP = 2^3 \cdot (2^3 \cdot (2^3 \cdot (2^3 \cdot 6P + 3P) + 2P) + 5P) + 4P$$

using 4 additions and 12 doublings for a total cost of 21 group operations. Note that in both cases we avoided computing $7P$ since it was not needed.

3.10 Root-finding in finite fields

Let $f(x)$ be a polynomial in $\mathbb{F}_q[x]$ of degree d. We wish to find a solution to $f(x) = 0$ that lies in $\mathbb{F}_q$. As an important special case, this will allow us to compute square roots using $f(x) = x^2 - a$, and, more generally, rth roots.[4]

The algorithm we give here was originally proposed by Berlekamp for prime fields [2], and then refined and extended by Rabin [18], whose presentation we follow here. The algorithm is probabilistic, and is one of the best examples of how randomness can be exploited in a number-theoretic setting. As we will see, it is quite efficient, with an expected running time that is quasi-quadratic in the size of the input. By contrast, no deterministic polynomial-time algorithm for root-finding is known, not even for computing square roots.[5]

3.10.1 Randomized algorithms

Probabilistic algorithms are typically classified as one of two types: *Monte Carlo* or *Las Vegas*. Monte Carlo algorithms are randomized algorithms whose output may be incorrect, depending on random choices that are made, but whose running time is bounded by a function of its input size, independent of any random choices. The probability of error is required to be less than $1/2-\epsilon$, for some $\epsilon > 0$, and can be made arbitrarily small by running the algorithm repeatedly and using the output that occurs most often. In contrast, a Las Vegas algorithm always produces a correct output, but its running time may depend on random choices; we do require that its expected running time is finite. As a trivial example, consider an algorithm to compute $a + b$ that first flips a coin repeatedly until it gets a head and then computes $a + b$ and outputs the result. The running time of this algorithm may be arbitrarily long, even when computing $1 + 1 = 2$, but its *expected* running time is $O(n)$, where n is the size of the inputs.

Las Vegas algorithms are generally preferred, particularly in mathematical applications. Note that any Monte Carlo algorithm whose output can be verified can always be converted to a Las Vegas algorithm (just run the algorithm repeatedly until you get an answer that is verifiably correct). The root-finding algorithm we present here is a Las Vegas algorithm.

3.10.2 Using GCDs to find roots

Recall from the previous lecture that we defined the finite field $\mathbb{F}_q$ to be the splitting field of $x^q - x$ over its prime field $\mathbb{F}_p$; this definition also applies when $q = p$ is prime (since $x^p - x$ splits completely in $\mathbb{F}_p$), and in every case, the elements of $\mathbb{F}_q$ are precisely the roots of $x^q - x$. The roots of f that lie in $\mathbb{F}_q$ are the roots it has in common with the polynomial $x^q - x$. We thus have

$$g(x) := \gcd(f(x), x^q - x) = \prod_i (x - \alpha_i),$$

where the α_i range over all the distinct roots of f that lie in $\mathbb{F}_q$. If f has no roots in $\mathbb{F}_q$ then g will have degree 0 (in which case $g = 1$). We have thus reduced our problem to finding a root of g, where g has distinct roots that are known to lie in $\mathbb{F}_q$.

[4] An entirely different approach to computing rth roots using discrete logarithms is explored in Problem Set 2. It has better constant factors when the r-power torsion subgroup of $\mathbb{F}_q^*$ is small (which is usually the case), but is asymptotically slower than the algorithm presented here in the worst case.

[5] Deterministic polynomial-time bounds for root-finding can be proved in various special cases, including the computation of square-roots, if one assumes a generalization of the Riemann hypothesis.

In order to compute $g = \gcd(f, x^q - x)$ efficiently, we generally do *not* compute $x^q - x$ and then take the gcd with f; this would take time exponential in $n = \log q$.[6] Instead, we compute $x^q \bmod f$ by exponentiating the polynomial x to the qth power in the ring $\mathbb{F}_q[x]/(f)$, whose elements are uniquely represented by polynomials of degree less than $d = \deg f$. Each multiplication in this ring involves the computation of a product in $\mathbb{F}_q[x]$ followed by a reduction modulo f; note that we do not assume $\mathbb{F}_q[x]/(f)$ is a field (indeed for $\deg f > 1$, if f has a root in $\mathbb{F}_q$ then $\mathbb{F}_q[x]/(f)$ is definitely not a field). This reduction is achieved using Euclidean division, and can be accomplished using two polynomial multiplications once an approximation to $1/f$ has been precomputed, see §3.7, and is within a constant factor of the time to multiply two polynomials of degree d in any case. The total cost of each multiplication in $\mathbb{F}_q[x]/(f)$ is thus $O(\mathsf{M}(d(n+\log d)))$, assuming that we use Kronecker substitution to multiply polynomials. The time to compute $x^q \bmod f$ using any of the exponentiation algorithms described in §3.9 is then $O(n\,\mathsf{M}(d(n+\log d)))$.

Once we have computed $x^q \bmod f$, we subtract x and compute $g = \gcd(f, x^q - x)$. Using the fast Euclidean algorithm, this takes $O(\mathsf{M}(d(n+\log d))\log d)$ time. Thus the total time to compute g is $O(\mathsf{M}(d(n+\log d))(n+\log d))$; and in the typical case where $\log d = O(n)$ (e.g. d is fixed and only n is growing) this simplifies to $O(n\,\mathsf{M}(dn))$.

So far we have not used randomness; we have a deterministic algorithm to compute the polynomial $g = (x - r_1)\cdots(x - r_k)$, where $r_1, \ldots, r_k$ are the distinct $\mathbb{F}_q$-rational roots of f. We can thus determine the number of distinct roots f has (this is just the degree of g), and in particular, whether it has any roots, deterministically, but knowledge of g does not imply knowledge of the roots $r_1, \ldots, r_k$ when $k > 1$; for example, if $f(x) = x^2 - a$ has a nonzero square root $r \in \mathbb{F}_q$, then $g(x) = (x - r)(x + r) = f(x)$ tells us nothing beyond the fact that $f(x)$ has a root.

3.11 Randomized GCD splitting

Having computed g, we seek to factor it into two polynomials of lower degree by again applying a gcd, with the goal of eventually obtaining a linear factor, which will yield a root.

Assuming that q is odd (which we do), we may factor the polynomial $x^q - x$ as

$$x^q - x = x(x^s - 1)(x^s + 1),$$

where $s = (q-1)/2$. Ignoring the root 0 (which we can easily check separately), this factorization splits $\mathbb{F}_q^\times$ precisely in half: the roots of $x^s - 1$ are the elements of $\mathbb{F}_q^\times$ that are squares in $\mathbb{F}_q^\times$, and the roots of $x^s + 1$ are the elements of $\mathbb{F}_q^\times$ that are not. Recall that $\mathbb{F}_q^\times$ is a cyclic group of order $q - 1$, and for $\alpha \in \mathbb{F}_q^\times$ we have $\alpha^s = \pm 1$ with $\alpha^s = 1$ precisely when α is a square in $\mathbb{F}_q^\times$. If we compute

$$h(x) = \gcd(g(x), x^s - 1),$$

we obtain a divisor of g whose roots are precisely the roots of g that are squares in $\mathbb{F}_q^\times$. If we suppose that the roots of g are as likely to be squares as not, we should expect the degree of h to be approximately half the degree of g. And so long as the degree of h is strictly between 0 and $\deg g$, one of h or g/h is a polynomial of degree at most half the degree of g, whose roots are all roots of our original polynomial f.

[6] The exception is when $d > q$, but in this case computing $\gcd(f(x), x^q - x)$ takes $O(\mathsf{M}(d(n+\log d)\log d))$ time, which turns out to be the same bound that we get for computing $x^q \bmod f(x)$ in any case.

To make further progress, and to obtain an algorithm that is guaranteed to work no matter how the roots of g are distributed in $\mathbb{F}_q$, we take a probabilistic approach. Rather than using the fixed polynomial $x^s - 1$, we consider random polynomials of the form

$$(x + \delta)^s - 1,$$

where δ is uniformly distributed over $\mathbb{F}_q$.

We claim that if α and β are any two nonzero roots of g, then with probability $1/2$, exactly one of these is a root $(x + \delta)^s - 1$. It follows from this claim that so long as g has at least 2 distinct nonzero roots, the probability that the polynomial $h(x) = \gcd(g(x), (x + \delta)^s - 1)$ is a proper divisor of g is at least $1/2$.

Let us say that two elements $\alpha, \beta \in \mathbb{F}_q$ are of *different type* if they are both nonzero and $\alpha^s \neq \beta^s$ (in which case $\alpha^s = \pm 1$ and $\beta^s = \mp 1$). Our claim is an immediate consequence of the following theorem from [18].

Theorem 3.44 (Rabin 1980)**.** *For every pair of distinct $\alpha, \beta \in \mathbb{F}_q$ we have*

$$\#\{\delta \in \mathbb{F}_q : \alpha + \delta \text{ and } \beta + \delta \text{ are of different type}\} = \frac{q-1}{2}.$$

Proof. Consider the map $\phi(\delta) = \frac{\alpha+\delta}{\beta+\delta}$, defined for $\delta \neq -\beta$. We claim that ϕ is a bijection from the set $\mathbb{F}_q - \{-\beta\}$ to the set $\mathbb{F}_q - \{1\}$. The sets are the same size, so we just need to show surjectivity. Let $\gamma \in \mathbb{F}_q - \{1\}$, then we wish to find a solution $\sigma \neq -\beta$ to $\gamma = \frac{\alpha+\sigma}{\beta+\sigma}$. We have $\gamma(\beta + \sigma) = \alpha + \sigma$ which means $\sigma - \gamma\sigma = \gamma\beta - \alpha$. This yields $\sigma = \frac{\gamma\beta-\alpha}{1-\gamma}$; we have $\gamma \neq 1$, and $\sigma \neq -\beta$, because $\alpha \neq \beta$. Thus ϕ is surjective.

We now note that

$$\phi(\delta)^s = \frac{(\alpha + \delta)^s}{(\beta + \delta)^s}$$

is -1 if and only if $\alpha + \delta$ and $\beta + \delta$ are of different type. The elements $\gamma = \phi(\delta)$ for which $\gamma^s = -1$ are precisely the non-residues in $\mathbb{F}_q \backslash \{1\}$, of which there are exactly $(q-1)/2$. □

We now give the algorithm, which assumes that its input $f \in \mathbb{F}_q[x]$ is monic (has leading coefficient 1). If f is not monic we can make it so by dividing f by its leading coefficient, which does not change its roots or the complexity of finding them. You can find an implementation of the algorithm below in this Jupyter notebook.

Algorithm 3.45. Given a monic polynomial $f \in \mathbb{F}_q[x]$, output an element $r \in \mathbb{F}_q$ such that $f(r) = 0$, or `null` if no such r exists.

1. If $f(0) = 0$ then return 0.
2. Compute $g = \gcd(f, x^q - x)$.
3. If $\deg g = 0$ then return `null`.
4. While $\deg g > 1$:
   a. Pick a random $\delta \in \mathbb{F}_q$.
   b. Compute $h = \gcd(g, (x + \delta)^s - 1)$.
   c. If $0 < \deg h < \deg g$ then replace g by h or g/h, whichever has lower degree.
5. Return r, where $g(x) = x - r$.

It is clear that the output of the algorithm is always correct: either it outputs a root of f in step 1, proves that f has no roots in $\mathbb{F}_q$ and outputs `null` in step 3, or outputs a root of g that is also a root of f in step 5 (note that whenever g is updated it replaced with a proper divisor). We now consider its complexity.

3.11.1 Complexity analysis

It follows from Theorem 3.44 that the polynomial h computed in step 4b is a proper divisor of g with probability at least $1/2$, since g has at least two distinct nonzero roots $\alpha, \beta \in \mathbb{F}_q$. Thus the expected number of iterations needed to obtain a proper factor h of g is bounded by 2, and the expected cost of obtaining such an h is $O(\mathsf{M}(e(n+\log e))(n+\log e))$, where $n = \log q$ and $e = \deg g$, and this dominates the cost of the division in step 4c.

Each time g is updated in step 4c its degree is reduced by at least a factor of 2. It follows that the expected total cost of step 4 is within a constant factor of the expected time to compute the initial value of $g = \gcd(f, x^q - x)$, which is $O(\mathsf{M}(d(n+\log d))(n+\log d))$, where $d = \deg f$; this simplifies to $O(n\,\mathsf{M}(dn))$ in the typical case that $\log d = O(n)$, which holds in all the applications we shall be interested in.

3.11.2 Finding all roots

We modify our algorithm to find all the distinct roots of f, by modifying step 4c to recursively find the roots of both h and g/h. In this case the amount of work done at each level of the recursion tree is bounded by $O(\mathsf{M}(d(n+\log d))(n+\log d))$. Bounding the depth of the recursion is somewhat more involved, but one can show that with very high probability the degrees of h and g/h are approximately equal and that the expected depth of the recursion is $O(\log d)$. Thus we can find all the distinct roots of f in

$$O(\mathsf{M}(d(n+\log d))(n+\log d)\log d) \tag{2}$$

expected time. When $\log d = O(n)$ this simplifies to $O(n\,\mathsf{M}(dn)\log d)$.

Once we know the distinct roots of f we can determine their multiplicity by repeated division, but this is not the most efficient approach. By taking GCDs with derivatives one can first compute the *squarefree factorization* of f, which for a monic nonconstant polynomial f is defined as the unique sequence $g_1, \ldots, g_m \in \mathbb{F}_q[x]$ of monic squarefree coprime polynomials with $g_m \neq 1$ such that

$$f = g_1 g_2^2 \cdots g_m^m.$$

When the degree of f is less than the characteristic p of $\mathbb{F}_q$, this can be done directly via Yun's algorithm [21]; see Exercise 14.30 in [9] for the necessary modifications to handle $\deg f \geq p$, which simply involves taking pth roots of known pth powers at suitable points and does not change the complexity.

Algorithm 3.46. Given a monic polynomial $f \in \mathbb{F}_q[x]$ with $\deg f < \mathrm{char}(\mathbb{F}_q)$, compute squarefree coprime polynomials $g_1, \ldots, g_m \in \mathbb{F}_q[x]$ with $g_m \neq 1$ such that $f = g_1 g_2^2 \cdots g_m^m$.

1. Compute $u = \gcd(f, f')$, $v_1 = f/u$, $w_1 = f'/u$, and set $i = 1$.
2. Compute $g_1 = \gcd(v_1, w_1 - v_1')$
3. While $v_i \neq g_i$:
   a. Compute v_{i+1} with v_i/g_i and $w_{i+1} = (w_i - v_i')/g_i$.
   b. Increment i and compute $g_i = \gcd(v_i, w_i - v_i')$
4. Set $m = i$ and return $g_1, \ldots, g_m$.

The key fact that Yun's algorithm exploits is that if $g \in \mathbb{F}_q[x]$ is irreducible then $g^2 | f$ if and only if $g | \gcd(f, f')$. This is true because $\mathbb{F}_q$ is a perfect field, (by Theorem 3.22): if

$f = gh$ then $f' = g'h + gh'$ is divisible by g if and only if $g'h$ is divisible by g, which occurs if and only if $g|h$ (in which case $g^2|f$), since $g' \neq 0$ for any irreducible g in a perfect field.

Yun's algorithm begins with $u = \gcd(f, f') = f/(g_1 \cdots g_m) = g_2 g_3^2 \cdots g_m^{m-1}$, which yields $v_1 = g_1 \cdots g_m$ and $w_1 = \sum_j j g_j' v_1/g_j$, since $f' = u \sum_{j=1}^m j g_j' v_1/g_j$. One can show by induction that we always have

$$v_i = g_i \cdots g_m \qquad \text{and} \qquad w_i = \sum_{j=i}^{m} (j - i + 1) g_j' v_i/g_j,$$

which implies $\gcd(v_i, w_i - v_i') = g_i$, since $w_i - v_i' = \sum_{j=i+1}^m (j-i) g_j' v_i/g_j$; see [9, Thm. 14.23].

Yun's algorithm uses $O(\mathsf{M}(d) \log d)$ ring operations in $\mathbb{F}_q$, which is $O(\mathsf{M}(n)\, \mathsf{M}(d) \log d)$ bit operations and strictly dominated by the complexity bound (2) for finding the distinct roots of f. The cost of finding the distinct roots of each g_i separately is no greater than the cost of finding the distinct roots of f, since the complexity of root-finding is superlinear in the degree, and with this approach we know *a priori* the multiplicity of each root of f.

It follows that we can determine all the roots of f and their multiplicities, with the same time complexity as finding the distinct roots of f (with the same leading constant, the extra time to determine the multiplicities is not only asymptotically negligible, when f is not squarefree it is actually faster to compute the squarefree factorization first).

3.12 Computing a complete factorization

Factoring a polynomial $f \in \mathbb{F}_q[x]$ into irreducibles can effectively be reduced to finding roots of f in extensions of $\mathbb{F}_q$. Linear factors of f correspond to the roots of f in $\mathbb{F}_q$, irreducible quadratic factors of f correspond to roots of f that lie in $\mathbb{F}_{q^2}$ but do not lie in $\mathbb{F}_q$; recall from Corollary 3.10 that *every* quadratic polynomial $\mathbb{F}_q[x]$ splits completely in $\mathbb{F}_{q^2}[x]$. More generally, each irreducible degree d-factor g of f is the minimal polynomial of a root α of f that lies in $\mathbb{F}_{q^d}$ but none of its proper subfields; note that if α is a root of $f \in \mathbb{F}_q[x]$, then so are all of its Galois conjugates, and these are precisely the roots of its minimal polynomial.

One can thus compute the complete factorization of f by applying the root-finding algorithm of the previous section over extensions of $\mathbb{F}_q$. But note that this involves picking random $\delta \in \mathbb{F}_{q^n}$ and performing polynomial arithmetic in $\mathbb{F}_{q^n}[x]$. As observed by Cantor and Zassenhaus shortly after Rabin's probabilistic root-finding algorithm appeared, rather than using random linear polynomials $x + \delta \in \mathbb{F}_{q^n}[x]$, it is better to use random degree n polynomials in $\mathbb{F}_q[x]$, and one can show that this works just as well.

To state this more precisely, let us first note that by computing the squarefree factorization of f and successively computing gcds with $x^{q^j} - x$ we can deterministically compute a factorization of f into squarefree polynomials each of which is a product of irreducible polynomials of the same known degree; this is called the *distinct-degree factorization* of f. It then only remains to consider the case where f is a product of distinct irreducible polynomials $f_1, \ldots, f_r$ of degree j. If $r = 1$ then f is irreducible and we are done, so let us assume $r > 1$. By the Chinese Remainder Theorem (CRT) we have a ring isomorphism

$$\frac{\mathbb{F}_q[x]}{(f)} \simeq \frac{\mathbb{F}_q[x]}{(f_1)} \times \cdots \times \frac{\mathbb{F}_q[x]}{(f_r)} \simeq \mathbb{F}_{q^j}^r$$

that sends $h \bmod f$ to $(h \bmod f_1, \ldots, h \bmod f_r)$. We can represent $\mathbb{F}_q[x]/(f)$ as the set of all polynomials $u \in \mathbb{F}_q[x]$ of degree strictly less than $\deg f = rj$. If we pick u uniformly at random, the polynomials $u \bmod f_i$ will also be uniformly random, and independent, by the

CRT (because the f_i are coprime). In other words, picking u at random amounts to picking an element $(u_1,\dots,u_r)$ of $\mathbb{F}_{q^j}^r$ at random. Moreover, if we pick a random u coprime to f we get a random $(u_1,\dots,u_r) \in (\mathbb{F}_{q^j}^\times)^r$.

Now let $s = (q^j-1)/2$. The ring isomorphism $u \mapsto (u_1,\dots,u_r) \in \mathbb{F}_{q^j}^r$ sends u^s to $(u_1^s,\dots,u_r^s) \in \{0,\pm 1\}^r$, and if we restrict to u that are coprime to f we will have $(u_1^s,\dots,u_r^s) \in \{\pm 1\}^r$ and $\gcd(f,u^s-1)$ will be non-trivial whenever we have $u_i^s = 1$ for at least one but not all of the u_i. Exactly half the elements of $\mathbb{F}_{q^j}^\times$ are roots of x^s-1 and half are not, so this probability is

$$1 - 2^{-r} - 2^{-r} = 1 - 2^{1-r} \geq 1/2.$$

We thus have at least a fifty-fifty chance of splitting f with each random u coprime to f. We now give the complete Cantor-Zassenhaus algorithm, as described in [9, §14]; you can find a basic implementation of the algorithm below in this Jupyter notebook.

Algorithm 3.47. Given a monic polynomial $f \in \mathbb{F}_q[x]$, compute its irreducible factorization as follows:

1. Compute the squarefree factorization of $f = g_1 g_2^2 \cdots g_m^m$ using Yun's algorithm.
2. By successively computing $g_{ij} = \gcd(g_i, x^{q^j} - x)$ and replacing g_i with g_i/g_{ij} for $j = 1,2,3,\dots,\deg g_i$, factor each g_i into polynomials g_{ij} that are each (possibly trivial) products of distinct irreducible polynomials of degree j; note that once $j > (\deg g_i)/2$ we know g_i must be irreducible and can immediately determine all the remaining g_{ij}.
3. Factor each nontrivial g_{ij} into irreducible polynomials g_{ijk} of degree j as follows: while $\deg g_{ij} > j$ generate random polynomials $u \in \mathbb{F}_q[x]$ with $\deg u < \deg g_{ij}$ until either $h = \gcd(g_{ij}, u)$ or $h := \gcd(g_{ij}, u^{(q^j-1)/2} - 1)$ properly divides g_{ij}, then recursively factor h and g_{ij}/h (note that $j | \deg h$ and $j | \deg(g_{ij}/h)$).
4. Output each g_{ijk} with multiplicity i.

In step 3, for $j > 1$ one computes $h_j := x^{q^j} \bmod g_{ij}$ via $h_j = h_{j-1}^q \bmod g_{ij}$. The expected cost of computing the g_{ij} for a given g_i of degree d is then bounded by

$$O(\mathsf{M}(d(n+\log d))d(n+\log d)),$$

which simplifies to $O(dn\,\mathsf{M}(dn))$ when $\log d = O(n)$ and is in any case quasi-quadratic in both d and n. The cost of factoring a particular g_{ij} satisfies the same bound with d replaced by j; the fact that this bound is superlinear and $\deg g_i = \sum_j \deg g_{ij}$ implies that the cost of factoring all the g_{ij} for a particular g_i is bounded by the cost of computing them, and superlinearity also implies that simply putting $d = \deg f$ gives us a bound on the cost of computing the g_{ij} for all the g_i, and this bound also dominates the $O(\mathsf{M}(d)(\log d)\,\mathsf{M}(n))$ complexity of step 1.

Notice that the first three steps of Algorithm 3.47, which compute the squarefree and distinct degree factorizations of f without making any random choices, yield a deterministic algorithm for computing the factorization pattern of f (the degrees and multiplicities of its irreducible factors), and in particular, can function as an irreducibility test.

There are faster algorithms for polynomial factorization that use linear algebra in $\mathbb{F}_q$; see [9, 14.8]. These are of interest primarily when the degree d is large relative to $n = \log q$.

The asymptotically fastest algorithm due to Kedlaya and Umans [17] uses recursive modular composition to obtain an expected running time of

$$O(d^{1.5+o(1)}n^{1+o(1)} + d^{1+o(1)}n^{2+o(1)}),$$

but this algorithm is primarily of theoretical interest.

There are also algorithms for $d = 2, 3, 4$ that use specialized methods for computing square-roots and cube-roots and then solve by radicals; these achieve a significant constant factor improvement for for most values of q, but will be slower in the rare worst case (the worst-case is slower by a $\log n/\log\log n$ factor [20], but one can easily detect this and switch algorithms if the slowdown outweighs the constant factor improvement).

For general purpose factoring of polynomials over finite fields, the Cantor-Zassenhaus algorithm is the algorithm of choice; it is implemented in virtually every computer algebra system that supports finite fields.

Remark 3.48. We should emphasize that all provably efficient algorithms known for root-finding and factoring polynomials over finite fields are probabilistic algorithms (of Las Vegas type). Even for the simplest non-trivial case, computing square roots, no deterministic polynomial-time algorithm is known. There are deterministic algorithms that can be shown to run in polynomial-time under the generalized Riemann hypothesis, but even these have worst-case running times that are asymptotically worse than the expected running time of the fastest probabilistic algorithms by at least a linear factor in $n = \log q$.

3.13 Summary

The table below summarizes the bit-complexity of the various arithmetic operations we have considered, both in the integer ring $\mathbb{Z}$ and in a finite field $\mathbb{F}_q$ of characteristic p with $q = p^e$; in both cases n denotes the bit-size of elements of the base ring (so $n = \log q$ for $\mathbb{F}_q$). Here we use $\mathsf{M}_q(n)$ to denote the time to multiply elements of $\mathbb{F}_q$. As noted in Remarks 3.36 and 3.42, if one is willing to assume a widely believed conjecture about the least prime in arithmetic progressions, we can take $\mathsf{M}_q(n) = O(n\log n)$ in the bounds below, and for $\log e = O(\log p)$ then this applies unconditionally.

	integers $\mathbb{Z}$	finite field $\mathbb{F}_q$
addition/subtraction	$O(n)$	$O(n)$
multiplication	$O(n\log n)$	$\mathsf{M}_q(n)$
Euclidean division (reduction)	$O(n\log n)$	$O(\mathsf{M}_q(n))$
extended gcd (inversion)	$O(n\log^2 n)$	$O(\mathsf{M}_q(n)\log n)$
exponentiation		$O(n\,\mathsf{M}_q(n))$
square-roots (probabilistic)		$O(n\,\mathsf{M}_q(n))$
root-finding (probabilistic)		$O(\mathsf{M}_q(d(n+\log d))(n+\log d))$
factoring (probabilistic)		$O(\mathsf{M}_q(d(n+\log d))d(n+\log d))$
irreducibility testing		$O(\mathsf{M}_q(d(n+\log d))d(n+\log d))$

In the case of root-finding, factorization, and irreducibility testing, d is the degree of the polynomial, and for probabilistic algorithms these are bounds on the expected running time of a Las Vegas algorithm. The bound for exponentiation assumes that the bit-length of the exponent is $O(n^2)$. Unless d is very large (super-exponential in n) one can ignore the $\log d$ terms in the last three complexity bounds, and we note that for large d there are faster approaches to factoring and irreducibility testing that are sub-quadratic in d.

References

[1] Razvan Barbulescu, Pierrick Gaudry, Antoine Joux, Emmanuel Thomé, *A heuristic quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic*, in *Advances in Cryptology — EUROCRYPT 2014*, LNCS **8441** (2014), 1–16.

[2] Elwyn R. Berlekamp, *Factoring polynomials over large finite fields*, Mathematics of Computation **24** (1970), 713–735.

[3] David G. Cantor and Hans Zassenhaus, *A new algorithm for factoring polynomials over finite fields*, Math. Comp. **36** (1981), 587–592.

[4] Henri Cohen et al., *Handbook of elliptic and hyperelliptic curve cryptography*, CRC Press, 2006.

[5] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to algorithms*, third edition, MIT Press, 2009.

[6] Jack Dongarra, Francis Sullivan, *Top ten algorithms of the century*, Computing in Science and Engineering **2** (2000), 22–23.

[7] Martin Fürer, *Faster integer multiplication*, Proceedings of the thirty-ninth annual ACM Symposium on the Theory of Computing (STOC), 2007.

[8] Joachim von zur Gathen, *Irreducible trinomials over finite fields*, Mathematics of Computation **72** (2003), 1787–2000.

[9] Joachim von zur Gathen and Jürgen Gerhard, *Modern computer algebra*, third edition, Cambridge University Press, 2013.

[10] Dan Givoli, *The top 10 computational methods of the 20th century*, IACM Expressions **11** (2001), 5–9.

[11] Daniel M. Gordon, *A survey of fast exponentiation methods*, Journal of Algorithms **27** (1998), 129–146.

[12] David Harvey and Joris van der Hoeven, *Faster integer multiplication using short lattice vectors*, Thirteenth Algorithmic Number Theory Symposium (ANTS XIII), Open Book Series **2** (2018), 293–310.

[13] David Harvey and Joris van der Hoeven, *Polynomial multiplication over finite fields in time $O(n \log n)$*, J. ACM **69** (2022), 1–40.

[14] David Harvey and Joris van der Hoeven, *Integer multiplication in time $O(n \log n)$*, Annals of Math. **193** (2021), 563–617.

[15] David Harvey, Joris van der Hoeven, and Grégoire Lecerf, *Even faster integer multiplication*, J. Complexity **36** (2016), 1–30.

[16] A. A. Karatsuba, *The complexity of computations*, Proceedings of the Steklov Institute of Mathematics **211** (1995), 169–193 (English translation of Russian article).

[17] Kiran S. Kedlaya and Christopher Umans, *Fast polynomial factorization and modular composition*, SIAM J. Comput. **40** (2011), 1767–1802.

[18] Michael O. Rabin, *Probabilistic algorithms in finite fields*, SIAM Journal of Computing **9** (1980), 273–280.

[19] Arnold Schönhage and Volker Strassen, *Schnelle Multiplikation großer Zahlen*, Computing, **7** (1971), 281–292.

[20] Andrew V. Sutherland, *Structure computation and discrete logarithms in finite abelian p-groups*, Math. Comp. **80** (2011), 815–831.

[21] David Y.Y. Yun, *On square-free decomposition algorithms*, in *Proceedings of the third ACM symposium on symbolic and algebraic computation (SYMSAC '76)*, R.D. Jenks (ed.), ACM Press, 1976, 26–35.

4 Isogenies

In almost every branch of mathematics, when considering a category of mathematical objects with a particular structure, the maps between objects that preserve this structure (morphisms) play a crucial role. For groups and rings we have homomorphisms, for vector spaces we have linear transformations, and for topological spaces we have continuous functions. For elliptic curves (and more generally, abelian varieties), the structure-preserving maps are called *isogenies*.[1]

Remark 4.1. I have included some general background on field extensions and algebraic sets at the end of these notes (see §4.6 and §4.7) for those who have not seen this material before (or would just like a refresher).

4.1 Morphisms of projective curves

As abelian varieties, elliptic curves have both an algebraic structure (as an abelian group), and a geometric structure (as a smooth projective curve). We are all familiar with morphisms of groups (these are group homomorphisms), but we have not formally defined a morphism of projective curves. To do so we need to define a few notions from algebraic geometry. Since algebraic geometry is not a prerequisite for this course, we will take a brief detour to define the terms we need.

To keep things as simple and concrete as possible, we will focus on plane projective curves with a few remarks along the way about how to generalize these definitions for those who are interested (those who are not can safely ignore the remarks). As usual, we use $\bar{k}$ to denote a fixed algebraic closure of our base field k that contains any and all algebraic extensions of k that we may consider (see §4.6 for more on algebraic closures).

Definition 4.2. Let C/k be a plane projective curve $f(x,y,z) = 0$ with f a nonconstant homogeneous polynomial in $k[x,y,z]$ that is irreducible in $\bar{k}[x,y,z]$. The *function field* $k(C)$ is the set of equivalence classes of rational functions g/h such that:

(i) g and h are homogeneous polynomials in $k[x,y,z]$ of the same degree;
(ii) h is not divisible by f, equivalently, h is not an element of the ideal (f);
(iii) g_1/h_1 and g_2/h_2 are considered equivalent whenever $g_1h_2 - g_2h_1 \in (f)$.

If L is any algebraic extension of k (including $L = \bar{k}$), the function field $L(C)$ is similarly defined with $g, h \in L[x,y,z]$.

Remark 4.3. The function field $k(X)$ of an irreducible projective variety X/k given by homogeneous polynomials $f_1, \dots, f_m \in k[x_0, \dots, x_n]$ is defined similarly: just replace the homogeneous ideal (f) with the homogeneous ideal $(f_1, \dots, f_m)$ (homogeneous ideal means an ideal of $k[x_0, \dots, x_n]$ generated by homogeneous polynomials).

Remark 4.4. Be sure not to confuse the notation $k(C)$ with $C(k)$; the latter denotes the set of k-rational points on C, not its function field.

[1] The word *isogeny* literally means "equal origins". It comes from biology, where the terms *isogenous*, *isogenic*, and *isogenetic* refer to different tissues derived from the same progenitor cell. The prefix "iso" means equal and the root "gene" means origin (as in the word *genesis*).

We claim that $k(C)$ is a ring under addition and multiplication of rational functions. To see this, first note that if $h_1, h_2 \notin (f)$ then $h_1 h_2 \notin (f)$ because f is irreducible and $k[x,y,z]$ is a unique factorization domain (in particular, (f) is a prime ideal). Thus for any $g_1/h_1, g_2/h_2 \in k(C)$ we have

$$\frac{g_1}{h_1} + \frac{g_2}{h_2} = \frac{g_1 h_2 + g_2 h_1}{h_1 h_2} \in k(C) \qquad \text{and} \qquad \frac{g_1}{h_1} \cdot \frac{g_2}{h_2} = \frac{g_1 g_2}{h_1 h_2} \in k(C).$$

We can compute the inverse of g/h as h/g except when $g \in (f)$, but in this case g/h is equivalent to $0/1 = 0$, since $g \cdot 1 - 0 \cdot h = g \in (f)$; thus every nonzero element of $k(C)$ is invertible, hence the ring $k(C)$ is a field.

Remark 4.5. The field $k(C)$ contains k as a subfield (take g and h with degree 0), but it is not an algebraic extension of k, it is transcendental. Indeed, it has transcendence degree 1, consistent with the fact that C is a projective variety of dimension 1 (this is one way to define the dimension of an algebraic variety). See §4.6 for more on transcendental field extensions.

The fact that g and h have the same degree allows us to meaningfully assign a value to the function g/h at a projective point $P = (x_0 : y_0 : z_0)$ on C, so long as $h(P) \neq 0$, since

(a) we get the same result for any projectively equivalent $P = (\lambda x_0 : \lambda y_0 : \lambda z_0)$ with $\lambda \in k^\times$, because g and h are homogeneous of the same degree (say d):

$$\frac{g(\lambda x, \lambda y, \lambda z)}{h(\lambda x, \lambda y, \lambda z)} = \frac{\lambda^d g(x,y,z)}{\lambda^d h(x,y,z)} = \frac{g(x,y,z)}{h(x,y,z)}.$$

(b) if g_1/h_1 and g_2/h_2 are equivalent and $h_1(P), h_2(P) \neq 0$, then $g_1(P)h_2(P) - g_2(P)h_1(P)$ is a multiple of $f(P) = 0$, so $(g_1/h_1)(P) = (g_2/h_2)(P)$.

Thus assuming the denominators involved are all nonzero, for $\alpha \in k(C)$ the value of $\alpha(P)$ does not depend on how we choose to represent either α or P. If $\alpha = g_1/h_1$ with $h_1(P) = 0$, it may happen that g_1/h_1 is equivalent to some g_2/h_2 with $h_2(P) \neq 0$. This is a slightly subtle point. It may not be immediately obvious whether or not such a g_2/h_2 exists, since it depends on equivalence modulo f; in general there may be no canonical way to write g/h in "lowest terms", because the ring $k[x,y,z]/(f)$ is typically *not* a unique factorization domain.

Example 4.6. Suppose C/k is defined by $f(x,y,z) = zy^2 - x^3 - z^2x = 0$, and consider the point $P = (0:0:1) \in C(k)$. We can't evaluate $\alpha = 3xz/y^2 \in k(C)$ at P as written since its denominator vanishes at P, but we can use the equivalence relation in $k(C)$ to write

$$\alpha = \frac{3xz}{y^2} = \frac{3xz^2}{x^3 + z^2x} = \frac{3z^2}{x^2 + z^2},$$

and we then see that $\alpha(P) = 3$.

Definition 4.7. Let C/k be a projective curve with $\alpha \in k(C)$. We say that α is *defined* (or *regular*) at a point $P \in C(\bar{k})$ if α can be represented as g/h for some $g, h \in k[x,y,z]$ with $h(P) \neq 0$.

Remark 4.8. If C is the projective closure of an affine curve $f(x,y)=0$, one can equivalently define $k(C)$ as the fraction field of $k[x,y]/(f)$; this ring is known as the *coordinate ring* of C, denoted $k[C]$, and it is an integral domain provided that (f) is a prime ideal (which holds in our setting because we assume f is irreducible). In this case one needs to homogenize rational functions $r(x,y)=g(x,y)/h(x,y)$ in order to view them as functions defined on projective space. This is done by introducing powers of z so that the numerator and denominator are homogeneous polynomials of the same degree. The same remark applies to (irreducible) varieties of higher dimension.

Recall that for any field F (including $F=k(C)$), we use $\mathbb{P}^2(F)$ to denote the set of projective triples $(x:y:z)$, with $x,y,z\in F$ not all zero, modulo the equivalence relation $(x:y:z)\sim(\lambda x:\lambda y:\lambda z)$ for $\lambda\in F^\times$.

Definition 4.9. Let C_1 and C_2 be plane projective curves defined over k. A *rational map* $\phi\colon C_1\to C_2$ is a projective triple $(\phi_x:\phi_y:\phi_z)\in\mathbb{P}^2(k(C_1))$, such that for every $P\in C_1(\bar{k})$ where ϕ_x,ϕ_y,ϕ_z are defined and not all zero, the projective point $(\phi_x(P):\phi_y(P):\phi_z(P))$ lies in $C_2(\bar{k})$. The map ϕ is *defined* (or *regular*) at P if there exists $\lambda\in k(C_1)^\times$ such that $\lambda\phi_x,\lambda\phi_y,\lambda\phi_z$ are all defined at P and not all zero at P.

Remark 4.10. This definition generalizes to projective varieties in $\mathbb{P}^n$ in the obvious way.

We should note that a rational map is not simply a function from $C_1(k)$ to $C_2(k)$ defined by rational functions, for two reasons. First, it might not be defined everywhere (although for smooth projective curves this does not happen, by Theorem 4.15 below). Second, it is required to map $C_1(\bar{k})$ to $C_2(\bar{k})$, which does not automatically hold for every rational map the carries $C_1(k)$ to $C_2(k)$; indeed, in general $C_1(k)$ could be the empty set (if C_1 is an elliptic curve then $C_1(k)$ is nonempty, but it could contain just a single point).

Remark 4.11. This is a general feature of classical algebraic geometry. In order for the definitions to work properly, one must consider the situation over an algebraic closure. An alternative and much more general approach is to use *schemes*, but this requires more material than we have time to develop in this course (take 18.725/6 to learn about schemes).

It is important to remember that a rational map $\phi=(\phi_x:\phi_y:\phi_z)$ is defined only up to scalar equivalence by functions in $k(C)^\times$. There may be points $P\in C_1(\bar{k})$ where one of $\phi_x(P),\phi_y(P),\phi_z(P)$ is not defined or all three are zero, but it may still possible to evaluate $\phi(P)$ after rescaling by $\lambda\in k(C)^\times$; we will see an example of this shortly.

The value of $\phi(P)$ is unchanged if we clear denominators in $(\phi_x:\phi_y:\phi_z)$ by multiplying through by an appropriate homogeneous polynomial (note: this is not the same as rescaling by an element of $\lambda\in k(C)^\times$). This yields a triple $(\psi_x:\psi_y:\psi_z)$ of homogeneous polynomials of equal degree that we view as a representing any of the three equivalent rational maps

$$(\psi_x/\psi_z:\psi_y/\psi_z:1),\qquad(\psi_x/\psi_y:1:\psi_z/\psi_y),\qquad(1:\psi_y/\psi_x:\psi_z/\psi_x),$$

all of which are equivalent to ϕ. We then have $\phi(P)=(\psi_x(P):\psi_y(P):\psi_z(P))$ whenever any of ψ_x,ψ_y,ψ_z is nonzero at P. Of course it can still happen that ψ_x,ψ_y,ψ_z all vanish at P, in which case we might need to look for an equivalent tuple of homogeneous polynomials that represents ϕ. The tuples $(\psi_x:\psi_y:\psi_z)$ and $(\psi'_x:\psi'_y:\psi'_z)$ represent the same rational map whenever the polynomials $\psi_x\psi'_y-\psi'_x\psi_y$ and $\psi_x\psi'_z-\psi'_x\psi_z$ and $\psi_y\psi'_z-\psi'_y\psi_z$ all lie in the ideal (f_1) defining C_1.

This defines an equivalence relation on set of triples $(\psi_x : \psi_y : \psi_z)$ of nonzero homogeneous polynomials of the same degree that satisfy $f_2(\psi_x, \psi_y, \psi_2) \in (f_1)$, where (f_2) is the ideal defining C_2. Each equivalence class corresponds to a rational map $C_1 \to C_2$ and every rational map has a corresponding equivalence class.

Remark 4.12. This set of equivalence classes of tuples defining rational maps $\psi\colon V_1 \to V_2$ of projective varieties also generalizes: replace (f_1) with the homogeneous ideal I_1 defining V_1 and require $f_2(\psi) \in I_1$ for every generator f_2 of the homogeneous ideal I_2 defining V_2.

This leads to the following equivalent definition of a rational map.

Definition 4.13. Let C_1 and C_2 be plane projective curves over k defined by $f_1(x, y, z) = 0$ and $f_2(x, y, z) = 0$, respectively. A *rational map* $\psi\colon C_1 \to C_2$ is an equivalence class of triples $(\psi_x : \psi_y : \psi_z)$ of homogeneous polynomials in $k[x, y, z]$ of the same degree, not all of which lie in (f_1), such that $f_2(\psi_x, \psi_y, \psi_z) \in (f_1)$. Triples $(\psi_x : \psi_y : \psi_z)$ and $(\psi'_x : \psi'_y : \psi'_z)$ are equivalent whenever $\psi_x\psi'_y - \psi'_x\psi_y$ and $\psi_x\psi'_z - \psi'_x\psi_z$ and $\psi_y\psi'_z - \psi'_y\psi_z$ all lie in (f_1).

The rational map ϕ is *defined* at $P \in C_1(\bar{k})$ if any of $\psi_x(P), \psi_y(P), \psi_z(P)$ is nonzero, in which case $(\psi_x(P) : \psi_y(P) : \psi_z(P) \in C_2(\bar{k}))$.

The equivalence of Definitions 4.9 and 4.13 follows from Corollary 4.52 (see §4.7).

Definition 4.14. A rational map that is defined everywhere is called a *morphism*.

For elliptic curves, distinguishing rational maps from morphisms is unnecessary; every rational map between elliptic curves is a morphism. More generally, we have the following.

Theorem 4.15. *If C_1 is a smooth projective curve then every rational map from C_1 to a projective curve C_2 is a morphism.*

The proof of this theorem is straightforward (see [6, II.2.1]), but requires a bit of commutative algebra that is outside the scope of this course.[2]

Remark 4.16. Theorem 4.15 is specific to smooth curves; it is not true more generally.

Two projective curves C_1 and C_2 are *isomorphic* if they are related by an invertible morphism ϕ; this means that there is a morphism ϕ^{-1} such that $\phi^{-1} \circ \phi$ and $\phi \circ \phi^{-1}$ are the identity maps on $C_1(\bar{k})$ and $C_2(\bar{k})$, respectively. An isomorphism $\phi\colon C_1 \to C_2$ is necessarily a morphism that defines a bijection from $C_1(\bar{k})$ to $C_2(\bar{k})$, but the converse is not true, in general, because the inverse map of sets from $C_2(\bar{k})$ to $C_1(\bar{k})$ might not be a morphism (because it can't be defined by rational functions); we will see an example of this shortly.

Before leaving the topic of morphisms of curves, we note one more useful fact.

Theorem 4.17. *A morphism of projective curves is either surjective or constant.*

This theorem is a consequence of the fact that projective varieties are *complete* (or *proper*), which implies that the image of a morphism of projective varieties is itself a projective variety. This is a standard result that is proved in most introductory algebraic geometry textbooks, see [2, II.4.9], for example. In the case of projective curves the image of a morphism $\phi\colon C_1 \to C_2$ of curves either has dimension 1, in which case ϕ is surjective (our curves are irreducible, by definition, and therefore cannot properly contain another curve), or dimension 0, in which case the image is a single point and ϕ is constant.

[2]The key point is that the coordinate ring of a smooth curve is a Dedekind domain. Thus its localization at every point P is a DVR, and after choosing a uniformizer we can rescale any rational map ϕ by a suitable λ (which will typically vary with P) so that all the components of ϕ have non-negative valuation at P and at least one has valuation zero and is therefore nonvanishing at P.

4.2 Isogenies of elliptic curves

We can now define the structure-preserving maps between elliptic curves that will play a key role in this course.

Definition 4.18. An *isogeny* $\phi\colon E_1 \to E_2$ of elliptic curves defined over k is a surjective morphism of curves that induces a group homomorphism $E_1(\bar{k}) \to E_2(\bar{k})$. The elliptic curves E_1 and E_2 are then said to be *isogenous*.

Remark 4.19. Unless otherwise stated, we assume that the isogeny ϕ is itself defined over k (meaning that it can be represented by a rational map whose coefficients lie in k). In general, if L/k is an algebraic extension, we say that two elliptic curves defined over k are "isogenous over L" if they are related by an isogeny that is defined over L. Strictly speaking, in this situation we are really referring to the "base change" of the elliptic curves to L (same equations, different field of definition), but we won't be pedantic about this.

This definition is stronger than is actually necessary, for three reasons. First, any morphism of abelian varieties that preserves the identity element (the distinguished point that is the zero element of the group) induces a group homomorphism; we won't bother to prove this (see [6, Theorem III.4.8] for a proof), since for all the isogenies we are interested in it will be obvious that they are group homomorphisms. Second, by Theorem 4.17, any non-constant morphism of curves is surjective, and third, by Theorem 4.15, a rational map whose domain is a smooth projective curve is automatically a morphism. This leads to the following equivalent definition which is commonly used.

Definition 4.20. An *isogeny* $\phi\colon E_1 \to E_2$ of elliptic curves defined over k is a non-constant rational map that sends the distinguished point of E_1 to the distinguished point of E_2.

Warning 4.21. Under our definitions the zero morphism, which maps every point on E_1 to the zero point of E_2, is *not* an isogeny. This follows the standard convention for general abelian varieties which requires isogenies to preserve dimension (so they must be surjective and have finite kernel). In the case of elliptic curves this convention is not always followed (notably, Silverman [6, III.4] includes the zero morphism in his definition of an isogeny), but it simplifies the statement of many theorems and is consistent with the more general usage you may see in later courses, so we will use it (we will still have occasion to refer to the zero morphism, we just won't call it an isogeny).

Definition 4.22. Elliptic curves E_1 and E_2 defined over a field k are *isomorphic* if there exist isogenies $\phi_1\colon E_1 \to E_2$ and $\phi_2\colon E_2 \to E_1$ whose composition is the identity; the isogenies ϕ_1 and ϕ_2 are then *isomorphisms*.

Definition 4.23. A morphism from an elliptic curve E/k to itself that fixes the distinguished point is called an *endomorphism*. An endomorphism that is also an isomorphism is an *automorphism*.

Except for the zero morphism, every endomorphism is an isogeny. As we shall see in the next lecture, the endomorphisms of an elliptic curve have a natural ring structure.

4.3 Examples of isogenies

We now give three examples of isogenies that are endomorphisms of an elliptic curve E/k defined by a short Weierstrass equation $y^2 = x^3 + Ax + b$ (we assume $\mathrm{char}(k) \neq 2, 3$).

4.3.1 The negation map

In projective coordinates the map $P \mapsto -P$ is given by

$$(x : y : z) \mapsto (x : -y : z),$$

which is evidently a rational map. It is defined at every projective point, and in particular, at every $P \in E(\bar{k})$, so it is a morphism (as it must be, since it is a rational map defined on a smooth curve). It fixes $0 = (0 : 1 : 0)$ and is not constant, thus it is an isogeny. It is also an endomorphism, since it is a morphism from E to E that fixes 0, and moreover an isomorphism (it is its own inverse), and therefore an automorphism.

4.3.2 The multiplication-by-2 map

Let E/k be the elliptic curve defined by $y^2 = x^3 + Ax + B$, and let $\phi\colon E \to E$ be defined by $P \mapsto 2P$. This is obviously a non-trivial group homomorphism (at least over $\bar{k}$), and we will now show that it is a morphism of projective curves. Recall that the formula for doubling an affine point $P = (x, y)$ on E is given by the rational functions

$$\phi_x(x, y) = m(x, y)^2 - 2x = \frac{(3x^2 + A)^2 - 8xy^2}{4y^2},$$
$$\phi_y(x, y) = m(x, y)(x - \phi_x(x, y)) - y = \frac{12xy^2(3x^2 + A) - (3x^2 + A)^3 - 8y^4}{8y^3},$$

where $m(x, y) := (3x^2 + A)/(2y)$ is the slope of the tangent line at P. Homogenizing these and clearing denominators yields the rational map $\phi := (\psi_x/\psi_z : \psi_y/\psi_z : 1)$, where

$$\psi_x(x, y, z) = 2yz\big((3x^2 + Az^2)^2 - 8xy^2z\big),$$
$$\psi_y(x, y, z) = 12xy^2z(3x^2 + Az^2) - (3x^2 + Az^2)^3 - 8y^4z^2,$$
$$\psi_z(x, y, z) = 8y^3z^3.$$

If $y = 0$ then $3x^2 + Az^2 \neq 0$ (because $y^2z = x^3 + Axz^2 + Bz^3$ is non-singular), and it follows that the only point in $E(\bar{k})$ where ψ_x, ψ_y, ψ_z simultaneously vanish is the point $0 = (0 : 1 : 0)$. As a rational map of smooth projective curves, we know that ϕ is a morphism, hence defined everywhere, so there must be an alternative representation of ϕ that we can evaluate at the point 0. Now in fact we know *a priori* that $\phi(0)$ must be 0, since $2 \cdot 0 = 0$ but let's verify this explicitly.

In projective coordinates the curve equation is $f(x, y, z) := y^2z - x^3 - Axz^2 - Bz^3 = 0$. We are free to add any multiple of f in $k[x, y, z]$ of the correct degree (in this case 6) to any of ψ_x, ψ_y, ψ_z without changing the rational function ϕ they define. Let us replace ψ_x with $\psi_x + 18xyzf$ and ψ_y with $\psi_y + (27f - 18y^2z)f$, and remove the common factor z^2 to obtain

$$\psi_x(x, y, z) = 2y\big(xy^2 - 9Bxz^2 + A^2z^3 - 3Ax^2z\big),$$
$$\psi_y(x, y, z) = y^4 - 12y^2z(2Ax + 3Bz) - A^3z^4$$
$$\qquad + 27Bz(2x^3 + 2Axz^2 + Bz^3) + 9Ax^2(3x^2 + 2Az^2),$$
$$\psi_z(x, y, z) = 8y^3z.$$

This is another representation of the rational map ϕ, and we can use this representation of ϕ to evaluate $\phi(0) = (\psi_x(0, 1, 0) : \psi_y(0, 1, 0) : \psi_z(0, 1, 0)) = (0 : 1 : 0) = 0$, as expected.

Having seen how messy things can get even with the relatively simple isogeny $P \mapsto 2P$, in the future we will be happy to omit such verifications and rely on the fact that if we have a rational map that we know represents an isogeny ϕ, then $\phi(0) = 0$ must hold. For elliptic curves in Weierstrass form, this means we only have to worry about evaluating isogenies at affine points, which allows us to simplify the equations by fixing $z = 1$.

4.3.3 The Frobenius endomorphism

Let $\mathbb{F}_p$ be a finite field of prime order p. The *Frobenius automorphism* $\pi\colon \overline{\mathbb{F}}_p \to \overline{\mathbb{F}}_p$ is the map $x \mapsto x^p$. It is easy to check that π is a field automorphism: $0^p = 0$, $1^p = 1$, $(-a)^p = -a^p$, $(a^{-1})^p = (a^p)^{-1}$, $(ab)^p = a^p b^p$, and $(a+b)^p = \sum \binom{p}{k} a^k b^{p-k} = a^p + b^p$. If $f(x_1, \ldots, x_k)$ is any rational function with coefficients in $\mathbb{F}_p$, then

$$f(x_1, \ldots, x_k)^p = f(x_1^p, \ldots, x_k^p),$$

since the coefficients of f are all fixed by π, which acts trivially on $\mathbb{F}_p$.

Every power π^n of π is also an automorphism of $\overline{\mathbb{F}}_p$; the fixed field of π^n is the finite field $\mathbb{F}_{p^n}$ with p^n elements. For a finite field $\mathbb{F}_q = \mathbb{F}_{p^n}$ the map $x \mapsto x^q$ is called the *q-power Frobenius map*, which we may denote by π_q.

Definition 4.24. Let E be an elliptic curve over a finite field $\mathbb{F}_q$. The *Frobenius endomorphism* of E is the map $\pi_E\colon (x : y : z) \mapsto (x^q : y^q : z^q)$.

To see that this defines a morphism from E to E, for any point $P = (x, y, z) \in E(\overline{\mathbb{F}}_q)$, if we raise both sides of the curve equation

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

to the qth power, we get

$$\begin{aligned}(y^2 z)^q &= (x^3 + Axz^2 + Bz^3)^q \\ (y^q)^2 z^q &= (x^q)^3 + Ax^q (z^q)^2 + B(z^q)^3,\end{aligned}$$

thus $(x^q : y^q : z^q) \in E(\overline{\mathbb{F}}_q)$; we have $A^q = A$ and $B^q = B$ because $A, B \in \mathbb{F}_q$. Note that when $q \neq p$ applying the p-power Frobenius yields a point on the elliptic curve $y^2 = x^3 + A^p x + B^p$, and unless $A, B \in \mathbb{F}_p$ this won't be the same curve as E (or even isomorphic to E, in general).

To see that π_E is also a group homomorphism, note that the group law on E is defined by rational functions whose coefficients lie in $\mathbb{F}_q$; these coefficients are invariant under the q-power map, so $\pi_E(P + Q) = \pi_E(P) + \pi_E(Q)$ for all $P, Q \in E(\overline{\mathbb{F}}_q)$.

These facts hold regardless of the equation used to define E and the formulas for the group law, including curves defined by a general Weierstrass equation (which is needed in characteristic 2 and 3).

Remark 4.25. The Frobenius endomorphism induces a group isomorphism from $E(\overline{\mathbb{F}}_q)$ to $E(\overline{\mathbb{F}}_q)$, since over the algebraic closure we can take qth roots of coordinates of points, and doing so still fixes elements of $\mathbb{F}_q$ (in other words, the inverse of π_q in $\mathrm{Gal}(\overline{\mathbb{F}}_q/\mathbb{F}_q)$ also commutes with the group operation). But as an *isogeny* the Frobenius endomorphism is *not* an isomorphism because there is no rational map from $E \to E$ that acts as its inverse (why this is so will become obvious in later lectures).

4.4 A standard form for isogenies

To facilitate our work with isogenies, it will be convenient to put them in a standard form. In order to do so we will assume throughout that we are working with elliptic curves of the form $y^2 = f(x)$, and when it is convenient we will further assume $f(x) = x^3 + Ax + B$ so that our curves are in short Weierstrass form. Implicit in this assumption is that our elliptic curves are defined over a field k whose characteristic is not 2, and when we assume $f(x) = x^3 + Ax + B$ we eliminate some elliptic curves in characteristic 3.

Lemma 4.26. *Let $E_1\colon y^2 = f_1(x)$ and $E_2\colon y^2 = f_2(x)$ be elliptic curves over k, and let $\alpha\colon E_1 \to E_2$ be an isogeny. Then α can be defined by an affine rational map of the form*

$$\alpha(x,y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right),$$

where $u, v, s, t \in k[x]$ are polynomials in x with $u \perp v$ and $s \perp t$.

The notation $u \perp v$ indicates that the polynomials u and v are coprime ($\gcd(u,v) = 1$).

Proof. Suppose α is defined by the rational map $(\alpha_x : \alpha_y : \alpha_z)$. Then for any affine point $(x : y : 1) \in E_1(\bar{k})$ we can write

$$\alpha(x,y) = \Big(r_1(x,y), r_2(x,y)\Big),$$

with $r_1(x,y) := \alpha_x(x,y,1)/\alpha_z(x,y,1)$ and $r_2(x,y) := \alpha_y(x,y,1)/\alpha_z(x,y,1)$. By repeatedly using the curve equation $y^2 = f_1(x)$ for E_1 to replace y^2 with $f_1(x)$, we can assume that both r_1 and r_2 have degree at most 1 in y. We then have

$$r_1(x,y) = \frac{p_1(x) + p_2(x)y}{p_3(x) + p_4(x)y},$$

for some $p_1, p_2, p_3, p_4 \in k[x]$. We now multiply the numerator and denominator of $r_1(x,y)$ by $p_3(x) - p_4(x)y$, and use the curve equation for E_1 to replace the y^2 in the denominator with $f_1(x)$, putting r_1 in the form

$$r_1(x,y) = \frac{q_1(x) + q_2(x)y}{q_3(x)},$$

for some $q_1, q_2, q_3 \in k[x]$.

We now use the fact that α is a group homomorphism and must therefore satisfy $\alpha(-P) = -\alpha(P)$ for any $P \in E_1(\bar{k})$. Recall that the inverse of an affine point (x,y) on a curve in short Weierstrass form is $(x,-y)$. Thus $\alpha(x,-y) = -\alpha(x,y)$ and we have

$$\Big(r_1(x,-y), r_2(x,-y)\Big) = \Big(r_1(x,y), -r_2(x,y)\Big)$$

Thus $r_1(x,y) = r_1(x,-y)$, and this implies that q_2 is the zero polynomial. After eliminating any common factors from q_1 and q_3, we obtain $r_1(x,y) = \frac{u(x)}{v(x)}$ for some $u, v \in k[x]$ with $u \perp v$, as desired. The argument for $r_2(x,y)$ is similar, except now we use $r_2(x,-y) = -r_2(x,y)$ to show that q_1 must be zero, yielding $r_2(x,y) = \frac{s(x)}{t(x)}y$ for some $s, t \in k[x]$ with $s \perp t$. □

We shall refer to the expression $\alpha(x,y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$ given by Lemma 4.26 as the *standard form* of an isogeny $\alpha\colon E_1 \to E_2$. This expression represents the isogeny $\alpha\colon E_1 \to E_2$ as an affine point in $\mathbb{P}^2(k(E_1))$ whose z-coordinate α_z is the constant function 1, which means that α_x and α_y are uniquely determined as elements of $k(E_1)$. The rational functions representing α_x and α_y as elements of $k(E_1)$ are uniquely determined if we also require that

- the numerator and denominator of α_x should be homogeneous coprime polynomials in $k[x,z]$ of the same degree.
- the numerator and denominator of α_y should be homogeneous coprime polynomials in $k[x,y,z]$ of the same degree, such that the numerator has the form $g(x,z)y$.

If we additionally require that u and s be monic, the polynomials $u,v,s,t \in k[x]$ that appear in the standard form of $\alpha(x,y)$ are uniquely determined by α.

Lemma 4.27. *Let $E_1\colon y^2 = f_1(x)$ and $E_2\colon y^2 = f_2(x)$ be elliptic curves over k and let $\alpha(x,y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$ be an isogeny from E_1 to E_2 in standard form. Then v^3 divides t^2 and t^2 divides $v^3 f_1$. Moreover, $v(x)$ and $t(x)$ have the same set of roots in $\bar{k}$.*

Proof. Substituting $\left(\frac{u}{v}, \frac{s}{t}y\right)$ for (x,y) in the equation for E_2 gives $((s/t)y)^2 = f_2(u/v)$, and using the equation for E_1 to replace y^2 with $f_1(x)$ yields

$$(s/t)^2 f_1 = f_2(u/v)$$

as an identity involving polynomials $f_1, f_2, s, t, u, v \in k[x]$. If we put $w = v^3 f_2(u/v) \in k[x]$ and clear denominators we obtain

$$v^3 s^2 f_1 = t^2 w. \tag{1}$$

Note that $u \perp v$ implies $v \perp w$, since any common factor of v and w must divide u. It follows that $v^3 | t^2$ and $t^2 | v^3 f_1$. This implies that v and t have the same roots in $\bar{k}$: every root of v is clearly a root of t (since $v^3|t^2$), and every root x_0 of t is a double root of $t^2|v^3 f_1$, and since f_1 has no double roots (because E_1 is not singular), x_0 must be a root of v (and possibly also a root of f_1). □

Corollary 4.28. *Let $\alpha(x,y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$ be an isogeny $E_1 \to E_2$ in standard form. The affine points $(x_0 : y_0 : 1) \in E_1(\bar{k})$ in the kernel of α are precisely those for which $v(x_0) = 0$.*

Proof. If $v(x_0) \neq 0$, then $t(x_0) \neq 0$, and $\alpha(x_0, y_0) = \left(\frac{u(x_0)}{v(x_0)}, \frac{s(x_0)}{t(x_0)}y\right)$ is an affine point and therefore not 0 (the point at infinity), hence not in the kernel of α.

By homogenizing and putting α into projective form, we can write α as

$$\alpha = (ut : vsy : vt),$$

where ut, vsy, and vt are now homogeneous polynomials of equal degree ($s,t,u,v \in k[x,z]$).

Suppose $y_0 \neq 0$. By the previous lemma, if $v(x_0,1) = 0$, then $t(x_0,1) = 0$, and since $v^3|t^2$, the multiplicity of $(x_0,1)$ as a root of t is strictly greater than its multiplicity as a root of v. This implies that, working over $\bar{k}$, we can renormalize α by dividing by a suitable power of $x - x_0 z$ so that α_y does not vanish at $(x_0 : y_0 : 1)$ but α_x and α_z both do. Then $\alpha(x_0 : y_0 : 1) = (0 : 1 : 0) = 0$, and $(x_0 : y_0 : 1)$ lies in the kernel of α as claimed.

If $y_0 = 0$, then x_0 is a root of the cubic $f(x)$ in the equation $y^2 = f_1(x)$ for E_1, and it is not a double root, since E_1 is not singular. In this case we renormalize α by multiplying

by yz and then replacing y^2z with $f_1(x,z)$. Because $(x_0,1)$ only has multiplicity 1 as a root of $f_1(x,z)$, its multiplicity as a root of vf_1 is no greater than its multiplicity as a root of t (here again we use $v^3|t^2$), and we can again renormalize α by dividing by a suitable power of $x - x_0z$ so that α_y does not vanish at $(x_0 : y_0 : 1)$, but α_x and α_z do (since they are now both divisible by $y_0 = 0$). Thus $(x_0 : y_0 : 1)$ is again in the kernel of α. □

The corollary implies that if we have an isogeny $\alpha\colon E_1 \to E_2$ in standard form, we know exactly what to do whenever we get a zero in the denominator when we try to compute $\alpha(P)$: we must have $\alpha(P) = 0$. This allows us to avoid in all cases the messy process that we went through earlier with the multiplication-by-2 map. We also obtain the following.

Corollary 4.29. *Let $\alpha\colon E_1 \to E_2$ be an isogeny of elliptic curves defined over a field k. The kernel of α is a finite subgroup of $E_1(\bar{k})$*

This corollary is true in general, but we will prove it under the assumption that we can put the isogeny α in our standard form (so $\operatorname{char}(k) \neq 2$).

Proof. If we put α in standard form $\left(\frac{u}{v}, \frac{s}{t}y\right)$ then the polynomial $v(x)$ has at most $\deg v$ distinct roots in $\bar{k}$, each of which can occur as the x-coordinate of at most two points on the elliptic curve E_1. □

Remark 4.30. Note that this corollary would not be true if we included the zero morphism in our definition of an isogeny.

One can also use the standard form of an isogeny $\alpha\colon E_1 \to E_2$ to show that α is surjective as a map from $E_1(\bar{k})$ to $E_2(\bar{k})$; see [7, Thm. 2.22].[3] But we already know that this applies to any non-constant morphism of curves (and even included surjectivity in our original definition of an isogeny), so we won't bother to prove this.

4.5 Degree and separability

We now define two important invariants of an isogeny that can be easily determined when it is in standard form.

Definition 4.31. Let $\alpha(x,y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$ be an isogeny in standard form. The *degree* of α is $\deg\alpha := \max\{\deg u, \deg v\}$, and we say that α is *separable* if the derivative of $\frac{u(x)}{v(x)}$ is nonzero; otherwise we say that α is *inseparable*.

As noted earlier, the polynomials u, v, s, t are uniquely determined up to a scalar factor, so the degree and separability of α are intrinsic properties that do not depend on its representation as a rational map.

Remark 4.32. The degree and separability of an isogeny can be defined in a way that is more obviously intrinsic using function fields. If $\alpha\colon E_1 \to E_2$ is an isogeny of elliptic curves defined over k then it induces an injection of function fields

$$\alpha^*\colon k(E_2) \to k(E_1)$$

that sends f to $f \circ \alpha$ (notice the direction of this map; the categorical equivalence between smooth projective curves and their function fields is contravariant). The degree of α is then

[3] The theorem in [7] assumes that α is an endomorphism but the proof works for any isogeny.

the degree of $k(E_1)$ as an extension of the subfield $\alpha^*(k(E_2))$; this degree is finite because both are finite extensions of a purely transcendental extension of k. The isogeny α is then said to be separable if this field extension is separable (and is inseparable otherwise). This approach has the virtue of generality, but it is not as easy to apply explicitly. Our definition is equivalent, but we won't prove this.

Let us now return to the three examples that we saw earlier.

- The standard form of the negation map is $\alpha(x, y) = (x, -y)$. It is separable and has degree 1.
- The standard form of the multiplication-by-2 isogeny on $y^2 = x^3 + Ax + B$ is

$$\alpha(x,y) = \left(\frac{x^4 - 2Ax^2 - 8Bx + A^2}{4(x^3 + Ax + B)}, \frac{x^6 + 5Ax^4 + 20Bx^3 - 5A^2x^2 - 4ABx - A^3 - 8B^2}{8(x^3 + Ax + B)^2} y\right).$$

  It is separable and has degree 4.
- The standard form of the Frobenius endomorphism of $E\colon y^2 = f(x)$ over $\mathbb{F}_q$ is

$$\pi_E(x,y) = \left(x^q, f(x)^{(q-1)/2} y\right).$$

  We have used the curve equation to replace y^q with $f(x)^{(q-1)/2}y$; note that q is odd because we are not in characteristic 2. The Frobenius endomorphism is inseparable, because $(x^q)' = qx^{q-1} = 0$ in $\mathbb{F}_q$ (since q is a multiple of the characteristic p), and it has degree q.

4.6 Field extensions

Most of the material in this section can be found in any standard introductory algebra text, such as [1, 3]. We will occasionally need results in slightly greater generality than you may have seen before, and here we may reference [4, 5].

We start in the general setting of an arbitrary field extension L/k with no restrictions on k or L. The fields k and L necessarily have the same prime field (the subfield of k generated by the multiplicative identity), and therefore the same characteristic. The *degree* of the extension L/k, denoted $[L\colon k]$, is the dimension of L as a k-vector space; this is a cardinal number, which need not be finite. If we have a tower of fields $k \subseteq L \subseteq M$, then

$$[M : k] = [M : L][L : k],$$

where the RHS is a product of cardinals.[4] When $[L : k]$ is finite we say that L/k is a *finite extension*.

An element $\alpha \in L$ is said to be *algebraic* over k if it is the root of a polynomial in $k[x]$, and otherwise it is *transcendental* over k. The extension L/k is *algebraic* if every element of L is algebraic over k, and otherwise it is transcendental. If M/L and L/k are both algebraic extensions, so is M/k. A necessary and sufficient condition for L/k to be algebraic is that L be equal to the union of all finite extensions of k contained in L; in particular, every finite extension is algebraic.

[4] Recall that a cardinal number is an equivalence class of equipotent sets (sets that can be put in bijection). The product of $n_1 = \#S_1$ and $n_2 = \#S_2$ is $n_1 n_2 = \#(S_1 \times S_2)$ and the sum is the cardinality of the disjoint union: $n_1 + n_2 = \#(S_1 \sqcup S_2)$. But we shall be primarily interested in finite cardinals (natural numbers).

The subset of L consisting of the elements that are algebraic over k forms a field called the *algebraic closure* of k in L. A field k is *algebraically closed* if every non-constant polynomial in $k[x]$ has a root in k; equivalently, k has no non-trivial algebraic extensions. For every field k there exists an extension $\bar{k}/k$ with $\bar{k}$ algebraically closed; such a $\bar{k}$ is called an *algebraic closure* of k, and all such $\bar{k}$ are isomorphic (but this isomorphism is not unique in general). Any algebraic extension L/k can be embedded into any algebraic closure of k, since every algebraic closure of L is also an algebraic closure of k.

Remark 4.33. When working with algebraic extensions of k it is convenient to view them all as subfields of some fixed algebraic closure $\bar{k}$ (there is in general no canonical choice). The key point is that we can always (not necessarily uniquely) embed any algebraic extension of L/k in our chosen $\bar{k}$, and if we have another extension M/L, our embedding of L into $\bar{k}$ can always be extended to an embedding of M into $\bar{k}$.

A set $S \subseteq L$ is said to be *algebraically independent* (over k) if for every finite subset $\{s_1, \dots, s_n\}$ of S and every nonzero polynomial $f \in k[x_1, \dots, x_n]$ we have

$$f(s_1, \dots, s_n) \neq 0.$$

Note that this means the empty set is algebraically independent (just as the empty set is linearly independent in any vector space). An algebraically independent set $S \subseteq L$ for which $L/k(S)$ is algebraic is called a *transcendence basis* for the extension L/k.

Theorem 4.34. *Every transcendence basis for L/k has the same cardinality.*

Proof. We will only prove this in the case that L/k has a finite transcendence basis (which includes all extensions of interest to us); see [4, Theorem 7.9] for the general case. Let $S = \{s_1, \dots, s_m\}$ be a smallest transcendence basis and let $T = \{t_1, \dots, t_n\}$ be any other transcendence basis, with $n \geq m$. The set $\{t_1, s_1, \dots, s_m\}$ must then algebraically dependent, since $t_1 \in L$ is algebraic over $k(S)$, and since t_1 is transcendental over k, some s_i, say s_1, must be algebraic over $k(t_1, s_2, \dots, s_m)$. It follows that L is algebraic over $k(t_1, s_2, \dots, s_m)$, and the set $T_1 = \{t_1, s_2, \dots, s_m\}$ must be algebraically independent, otherwise it would contain a transcendence basis for L/k smaller than S. So T_1 is a transcendence basis for L/k of cardinality m that contains t_1.

Continuing in this fashion, for $i = 2, \dots, m$ we can iteratively construct transcendence bases T_i of cardinality m that contain $\{t_1, \dots, t_i\}$, until $T_m \subseteq T$ is a transcendence basis of cardinality m; but then we must have $T_m = T$, so $n = m$. □

Definition 4.35. The *transcendence degree* of a field extension L/K is the cardinality of any (hence every) transcendence basis for L/k.

Unlike extension degrees, which multiply in towers, transcendence degrees add in towers: for any fields $k \subseteq L \subseteq M$, the transcendence degree of M/k is the sum (as cardinals) of the transcendence degrees of M/L and L/k.

We say that the extension L/k is *purely transcendental* if $L = k(S)$ for some transcendence basis S for L/k. All purely transcendental extensions of k with the same transcendence degree are isomorphic. Every field extension L/k can be viewed as an algebraic extension of a purely transcendental extension: if S is a transcendence basis of L/k then $L/k(S)$ is an algebraic extension of the purely transcendental extension $k(S)/k$.

Remark 4.36. It is not the case that every field extension is a purely transcendental extension of an algebraic extension; indeed, most function fields are counterexamples.

The field extension L/k is said to be *simple* if $L = k(x)$ for some $x \in L$. A purely transcendental extension of transcendence degree 1 is obviously simple, but, less trivially, so is any finite separable extension (see below for the definition of separable); this is known as the primitive element theorem.

Remark 4.37. The notation $k(x)$ can be slightly confusing. If $x \in L$ is transcendental over k then $k(x)$ is isomorphic to the field of rational functions over k, in which case we may as well regard x as a variable. But if $x \in L$ is algebraic over k, then every rational expression $r(x)$ with nonzero denominator can be simplified to a polynomial in x of degree less than $n = [k(x) : k]$ by reducing modulo the minimal polynomial f of x (note that we can invert nonzero denominators modulo f); indeed, this follows from the fact that $\{1, x, \dots, x^{n-1}\}$ is a basis for the n-dimensional k-vector space $k(x)$.

4.6.1 Algebraic extensions

We now assume that L/k is algebraic and fix $\bar{k}$ so that $L \in \bar{k}$. The extension L/k is *normal* if it satisfies either of the equivalent conditions:

- every irreducible polynomial in $k[x]$ with a root in L splits completely in L;
- $\sigma(L) = L$ for all $\sigma \in \mathrm{Aut}(\bar{k}/k)$ (every automorphism of $\bar{k}$ that fixes k also fixes L).[5]

Even if L/k is not normal, there is always an algebraic extension M/L for which M/k is normal. The minimal such extension is called the *normal closure* of L/k; it exists because intersections of normal extensions are normal. It is not true in general that if L/k and M/L are normal extensions then so is M/k, but if $k \subseteq L \subseteq M$ is a tower of fields with M/k normal, then M/L is normal (but L/k need not be).

A polynomial $f \in k[x]$ is *separable* if any of the following equivalent conditions hold:

- the factors of f in $\bar{k}[x]$ are all distinct;
- f and f' have no common root in $\bar{k}$;
- $\gcd(f, f') = 1$ in $k[x]$.

An element $\alpha \in L$ is separable over k if any of the following equivalent conditions hold:

- α is a root of a separable polynomial $f \in k[x]$;
- the minimal polynomial of α is separable;
- $\mathrm{char}(k) = 0$ or $\mathrm{char}(k) = p > 0$ and the minimal polynomial of α is not of the form $g(x^p)$ for some $g \in k[x]$.

The elements of L that are separable over k form a field called the *separable closure* of k in L. The separable closure of k in its algebraic closure $\bar{k}$ is denoted k^{sep} and is simply called the separable closure of k. If $k \subseteq L \subseteq M$ then M/k is separable if and only if both M/L and L/k are separable.

Definition 4.38. A field k is *perfect* if any of the following equivalent conditions hold:

- $\mathrm{char}(k) = 0$ or $\mathrm{char}(k) = p > 0$ and $k = \{x^p : x \in k\}$ (k is fixed by Frobenius);
- every finite extension of k is separable over k;
- every algebraic extension of k is separable over k.

[5] Some authors write $\mathrm{Gal}(L/k)$ for $\mathrm{Aut}(L/k)$, others only use $\mathrm{Gal}(L/k)$ when L/k is known to be Galois; we will use the latter convention.

It is clear from the definition that finite fields and all fields of characteristic 0 are perfect, which includes most of the fields of interest to us in this course.

Example 4.39. The rational function field $k = \mathbb{F}_p(t)$ is not perfect. If we consider the finite extension $L = k(t^{1/p})$ obtained by adjoining a pth root of t to k, the minimal polynomial of $t^{1/p}$ is $x^p - t$, which is irreducible over k but not separable (its derivative is 0).

Definition 4.40. An algebraic extension L/k is *Galois* if it is both normal and separable, in which case we call $\mathrm{Gal}(L/k) = \mathrm{Aut}(L/k)$ the *Galois group* of L/k.

The extension k^{sep}/k is always normal: if an irreducible polynomial $f \in k[x]$ has a root α in k^{sep}, then (up to scalars) f is the minimal polynomial of α over k, hence separable over k, so all its roots lie in k^{sep}. Thus k^{sep}/k is a Galois extension and its Galois group

$$G_k \coloneqq \mathrm{Gal}(k^{\mathrm{sep}}/k)$$

is the *absolute Galois group* of k (we could also define G_k as $\mathrm{Aut}(\bar{k}/k)$, since the restriction map from $\mathrm{Aut}(\bar{k}/k)$ to $\mathrm{Gal}(k^{\mathrm{sep}}/k)$ is an isomorphism).

The *splitting field* of a polynomial $f \in k[x]$ is the extension of k obtained by adjoining all the roots of f (which lie in $\bar{k}$). Every splitting field is normal, and every finite normal extension of k is the splitting field of some polynomial over k; when k is a perfect field we can go further and say that L/k is a finite Galois extension if and only if it is the splitting field of some polynomial over k.

For finite Galois extensions M/k we always have $\#\mathrm{Gal}(M/k) = [M : k]$, and the fundamental theorem of Galois theory gives an inclusion-reversing bijection between subgroups $H \subseteq \mathrm{Gal}(M/k)$ and intermediate fields $k \subseteq L \subseteq M$ in which $L = M^H$ and $H = \mathrm{Gal}(M/L)$ (note that M/L is necessarily Galois). Beware that none of the statements in this paragraph necessarily apply to infinite Galois extensions; modifications are required.[6]

4.7 Algebraic sets

Let k be a perfect field and fix an algebraic closure $\bar{k}$.

Definition 4.41. The n-dimensional *affine space* $\mathbb{A}^n = \mathbb{A}^n_k$ over k is the set

$$\mathbb{A}^n \coloneqq \{(x_1, \dots, x_n) \in \bar{k}^n\},$$

equivalently, $\mathbb{A}^n$ is the vector space $\bar{k}^n$ regarded as a set. When k is clear from context we may just write $\mathbb{A}^n$. If $k \subseteq L \subseteq \bar{k}$, the set of *$L$-rational points* (or just L-points) in $\mathbb{A}^n$ is

$$\mathbb{A}^n(L) = \{(x_1, \dots, x_n) \in L^n\} = \mathbb{A}^n(\bar{k})^{G_L},$$

where $\mathbb{A}^n(\bar{k})^{G_L}$ denotes the set of points in $\mathbb{A}^n(\bar{k})$ fixed by $G_L \coloneqq \mathrm{Gal}(L^{\mathrm{sep}}/L)$. In particular, $\mathbb{A}^n(k) = \mathbb{A}^n(\bar{k})^{G_k}$.

Definition 4.42. If S is a set of polynomials in $\bar{k}[x_1, \dots, x_n]$, the set of points

$$Z_S \coloneqq \{P \in \mathbb{A}^n : f(P) = 0 \text{ for all } f \in S\},$$

is called an (affine) *algebraic set*. If $k \subseteq L \subseteq \bar{k}$, the set of L-rational points in Z_S is

$$Z_S(L) = Z_S \cap \mathbb{A}^n(L).$$

When S is a singleton $\{f\}$ we may write Z_f in place of $Z_{\{f\}}$.

[6] See Section 26.3 in the 18.785 Lecture notes for more details on infinite Galois extensions.

Note that if I is the $\bar{k}[x_1,\dots,x_n]$-ideal generated by S, then $Z_I = Z_S$, since $f(P) = g(P) = 0$ implies $(f+g)(P) = 0$ and $f(P) = 0$ implies $(fg)(P) = 0$. Thus we can always replace S by the ideal (S) that it generates, or by any set of generators for (S).

Example 4.43. We have $Z_\emptyset = Z_{(0)} = \mathbb{A}^n$ and $Z_{\{1\}} = Z_{(1)} = \emptyset$.

For any $S, T \subseteq A$ we have

$$S \subseteq T \quad \Longrightarrow \quad Z_T \subseteq Z_S,$$

but the converse need not hold, even if S and T are ideals: consider $T = (x_1)$ and $S = (x_1^2)$.

We now recall the notion of a noetherian ring and the Hilbert basis theorem.

Definition 4.44. A commutative ring R is *noetherian* if every R-ideal is finitely generated.[7] Equivalently, every infinite ascending chain of R-ideals

$$I_1 \subseteq I_2 \subseteq \cdots$$

eventually stabilizes, that is, $I_{n+1} = I_n$ for all sufficiently large n.

Theorem 4.45 (Hilbert basis theorem)**.** *If R is a noetherian ring, then so is $R[x]$.*

Proof. See [1, Theorem 14.6.7] or [3, Theorem 8.32]. □

Note that we can apply the Hilbert basis theorem repeatedly: if R is noetherian then so is $R[x_1]$, and so is $(R[x_1])[x_2] = R[x_1, x_2]$, ..., and so is $R[x_1,\dots,x_n]$. Like every field, $\bar{k}$ is a noetherian ring (it has just two ideals, so it certainly satisfies the ascending chain condition). Thus $A = \bar{k}[x_1,\dots,x_n]$ is noetherian, so every A-ideal is finitely generated. It follows that every algebraic set can be written in the form Z_S with S finite.

Definition 4.46. For an algebraic set $Z \subseteq \mathbb{A}^n$, the *ideal of* Z is the set

$$I(Z) = \{f \in \bar{k}[x_1,\dots,x_n] : f(P) = 0 \text{ for all } P \in Z\}.$$

The set $I(Z)$ is clearly an ideal, since it is closed under addition and under multiplication by elements of $\bar{k}[x_1,\dots,x_n]$, and we note that

$$Y \subseteq Z \implies I(Z) \subseteq I(Y)$$

and

$$I(Y \cup Z) = I(Y) \cap I(Z)$$

(both statements are immediate from the definition).

We have $Z = Z_{I(Z)}$ for every algebraic set Z, but it is not true that $I = I(Z_I)$ for every ideal I. As a counterexample, consider $I = (f^2)$ for some polynomial $f \in A$. In this case

$$I(Z_{(f^2)}) = (f) \neq (f^2).$$

In order to avoid this situation, we want to restrict our attention to *radical* ideals.

Definition 4.47. Let R be a commutative ring. For any R-ideal I we define

$$\sqrt{I} = \{x \in R : x^r \in I \text{ for some integer } r > 0\},$$

and say that I is a *radical ideal* if $I = \sqrt{I}$.

[7] The term "noetherian" refers to the German mathematician Emmy Noether.

Lemma 4.48. *For any ideal I in a commutative ring R, the set $\sqrt{I}$ is an ideal.*

Proof. Let $x \in \sqrt{I}$ with $x^r \in I$. For any $y \in R$ we have $y^r x^r = (xy)^r \in I$, so $xy \in \sqrt{I}$. If $y \in \sqrt{I}$ with $y^s \in I$, then every term in the sum

$$(x+y)^{r+s} = \sum_i \binom{r+s}{i} x^i y^{r+s-i}$$

is a multiple of either $x^r \in I$ or $y^s \in I$, hence lies in I, so $(x+y)^{r+s} \in I$ and $(x+y) \in \sqrt{I}$. □

Theorem 4.49 (Hilbert's *Nullstellensatz*)**.** *For every ideal $I \subseteq \bar{k}[x_1, \ldots, x_n]$ we have*

$$I(Z_I) = \sqrt{I}.$$

Proof. See [4, Theorem 7.1]. □

Nullstellensatz literally means "zero locus theorem". Theorem 4.49 is the strong form of the *Nullstellensatz*; it implies the *weak Nullstellensatz*.

Theorem 4.50 (*weak Nullstellensatz*)**.** *For any ideal $I \subsetneq \bar{k}[x_1, \ldots, x_n]$, the variety Z_I is nonempty.*

Proof. Suppose I is an ideal for which Z_I is the empty set. Then $I(Z_I) = (1)$, and by the strong *Nullstellensatz*, $\sqrt{I} = (1)$. But then $1^r = 1 \in I$, so $I = \bar{k}[x_1, \ldots, x_n]$. □

Note the importance of working over the algebraic closure $\bar{k}$. It is easy to find proper ideals I for which $Z_I(k) = \emptyset$ when k is not algebraically closed; consider $Z_{(x^2+y^2+1)}(\mathbb{Q})$ in $\mathbb{A}^2$. A useful corollary of the weak *Nullstellensatz* is the following.

Corollary 4.51. *The maximal ideals of the ring $\bar{k}[x_1, \ldots, x_n]$ are all of the form*

$$m_P = (x_1 - P_1, \ldots, x_n - P_n)$$

for some point $P = (P_1, \ldots, P_n)$ in $\mathbb{A}^n(\bar{k})$.

Proof. The evaluation map that sends $f \in \bar{k}[x_1, \ldots, x_n]$ to $f(P) \in \bar{k}$ is a surjective ring homomorphism with kernel m_P. Thus $\bar{k}[x_1, \ldots, x_n]/m_P \simeq \bar{k}$ is a field, hence m_P is a maximal ideal. If m is any maximal ideal in $\bar{k}[x_1, \ldots, x_n]$, then it is a proper ideal, and by the weak *Nullstellensatz* the algebraic set Z_m is nonempty and contains a point $P \in \mathbb{A}^n$. So $I(Z_m) \subseteq m_P$, but $m \subseteq I(Z_m) \subseteq m_P$ is maximal, so $m = m_P$. □

We also have the following corollary of Hilbert's *Nullstellensatz*.

Corollary 4.52. *There is a one-to-one inclusion-reversing correspondence between radical ideals $I \subseteq \bar{k}[x_1, \ldots, x_n]$ and algebraic sets $Z \subseteq \mathbb{A}^n(\bar{k})$ in which $I = I(Z)$ and $Z = Z_I$.*

Remark 4.53. It is hard to overstate the importance of Corollary 4.52; it is the basic fact that underlies nearly all of algebraic geometry. It tells us that the study of algebraic sets (geometric objects) is the same thing as the study of radical ideals (algebraic objects). It also suggests ways in which we might generalize our notion of an algebraic set: there is no reason to restrict ourselves to radical ideals in the ring $\bar{k}[x_1, \ldots, x_n]$, there are many other rings we might consider. This approach eventually leads to the more general notion of a *scheme*, which is the fundamental object in modern algebraic geometry.

Definition 4.54. An algebraic set is *irreducible* if it is nonempty and not the union of two smaller algebraic sets.

Theorem 4.55. *An algebraic set is irreducible if and only if its ideal is prime.*

Proof. ($\Rightarrow$) Let Y be an irreducible algebraic set and suppose $fg \in I(Y)$ for some $f, g \in A$. We will show that either $f \in I(Y)$ or $g \in I(Y)$ (and therefore $I(Y)$ is prime).

$$\begin{aligned} Y &\subseteq Z_{fg} = Z_f \cup Z_g \\ &= (Y \cap Z_f) \cup (Y \cap Z_g), \end{aligned}$$

and since Y is irreducible we must have either $Y = (Y \cap Z_f) = Z_f$ or $Y = (Y \cap Z_g) = Z_g$, hence either $f \in I(Y)$ or $g \in I(Y)$. Therefore $I(Y)$ is a prime ideal.

($\Leftarrow$) Now suppose $I(Y)$ is prime and that $Y = Y_1 \cup Y_2$. We will show that either $Y = Y_1$ or $Y = Y_2$. This will show that Y is irreducible, since Y must be nonempty ($I(Y) \neq A$ because $I(Y)$ is prime). We have

$$I(Y) = I(Y_1 \cup Y_2) = I(Y_1) \cap I(Y_2) \supseteq I(Y_1)I(Y_2),$$

and therefore $I(Y)$ divides/contains either $I(Y_1)$ or $I(Y_2)$, since $I(Y)$ is a prime ideal, but it is also contained in both $I(Y_1)$ and $I(Y_2)$, so either $I(Y) = I(Y_1)$ or $I(Y) = I(Y_2)$. Thus either $Y = Y_1$ or $Y = Y_2$, since algebraic sets with the same ideal must be equal. □

References

[1] Michael Artin, *Algebra*, 2nd edition, Pearson Education, 2011.

[2] Robin Hartshorne, *Algebraic geometry*, Graduate Texts in Mathematics **52**, Springer, 1977.

[3] Anthony W. Knapp, *Basic algebra*, Springer, 2006.

[4] Anthony W. Knapp, *Advanced algebra*, Springer, 2007.

[5] J. S. Milne, *Fields and Galois Theory*, 2012.

[6] J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics **106**, second edition, Springer 2009.

[7] Lawrence C. Washington, *Elliptic curves: Number theory and cryptography*, second edition, Chapman and Hall/CRC, 2008.

5 Isogeny kernels and division polynomials

In this lecture we continue our study of isogenies of elliptic curves. Recall that an isogeny is a surjective morphism that is also a group homomorphism, equivalently, a non-constant rational map that fixes the identity. In the previous lecture we showed that every nonzero isogeny $\alpha\colon E_1 \to E_2$ between elliptic curves of the form $y^2 = f(x)$ can be written in the standard affine form

$$\alpha(x,y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right),$$

where $u \perp v$ and $s \perp t$ are pairs of relatively prime polynomials in $k[x]$.[1] For any affine point $(x_0, y_0) \in E_1(\bar{k})$, we have $\alpha(x_0, y_0) = 0$ if and only if $v(x_0) = 0$ (equivalently, if and only if $t(x_0) = 0$; see Lemma 4.27 and Corollary 4.28). This follows from the fact that $\ker \alpha$ is a subgroup, so if $P = (x_0, y_0) \in \ker \alpha$ then so is $-P = (x_0, -y_0)$, and this accounts for every point in $E_1(\bar{k})$ with x-coordinate x_0. It follows that

$$\ker \alpha = \{(x_0, y_0) \in E_1(\bar{k})\colon v(x_0) = 0\} \cup \{0\}$$

is determined by the polynomial $v(x)$ (here $0 := (0:1:0)$ is the point at infinity).

When α is the multiplication-by-n map $P \mapsto nP = P + \cdots + P$ (which is an isogeny because it is a group homomorphism defined by a non-constant rational map), the kernel of α is the *n-torsion subgroup*

$$E[n] := \{P \in E(\bar{k}) : nP = 0\}.$$

Torsion subgroups play a key role in the theory of elliptic curves. In particular, when $k = \mathbb{F}_q$ is a finite field, the finite abelian group $E(\mathbb{F}_q)$ is completely determined by its intersection with the n-torsion subgroups $E[n]$. Understanding the structure of $E[n]$ will allow us to understand the structure of $E(\mathbb{F}_q)$, and will also turn out to be the key to efficiently computing $\#E(\mathbb{F}_q)$.

5.1 Kernels of isogenies

Recall that the *degree* of an isogeny α in standard form is defined to be $\max\{\deg u, \deg v\}$, and α is *separable* whenever $\left(\frac{u}{v}\right)' \neq 0$. We are going to prove that for separable isogenies, the order of its kernel is equal to its degree. But we will first dispose of the inseparable case by showing that every isogeny can be decomposed into the composition of a separable isogeny and a power of the p-power Frobenius morphism (which has trivial kernel).

Lemma 5.1. *Let u and v be relatively prime polynomials in $k[x]$.*

$$\left(\frac{u}{v}\right)' = 0 \quad\iff\quad u' = v' = 0 \quad\iff\quad u = f(x^p) \text{ and } v = g(x^p),$$

where f and g are polynomials in $k[x]$ and p is the characteristic of k (which may be zero).

[1] The assumption that E_1 and E_2 are defined by equations of the form $y^2 = f(x)$ implies we are not in characteristic 2. Most of the results we will prove can easily be extended to curves in general Weierstrass form and thus apply to all elliptic curves. When this is the case we will state our theorems generally, but in our proofs we will restrict to elliptic curves $y^2 = f(x)$.

Proof. Suppose $\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v^2} = 0$. Then

$$u'v = v'u.$$

The polynomials u and v have no common roots in $\bar{k}$, therefore every root of u in $\bar{k}$ must also be a root of u', with at least the same multiplicity. But $\deg u' < \deg u$, so this is possible only if $u' = 0$, and by the same argument we must also have $v' = 0$. Conversely, if $u' = v' = 0$ then $u'v = v'u$. This proves the first equivalence.

Now let $u(x) = \sum_n a_n x^n$. If $u'(x) = \sum n a_n x^{n-1} = 0$, then $na_n = 0$ for every n, which means that n must be a multiple of p for every nonzero a_n (if $p = 0$ this means $u' = 0$). In this case we can write u as

$$u(x) = \sum_m a_{pm}(x^p)^m = f(x^p),$$

where $f = \sum_m a_m x^m$. Similarly, if $v'(x) = 0$ then $v(x) = g(x^p)$ for some $g \in k[x]$. Conversely, if $u(x) = f(x^p)$ then $u'(x) = px^{p-1}f'(x^p) = 0$, and similarly for $v(x)$. □

Corollary 5.2. *Over a field of characteristic zero, every isogeny is separable.*

We now show that every inseparable isogeny arises as the composition of a separable isogeny with some power of the p-power Frobenius map $\pi\colon (x, y, z) \mapsto (x^p, y^p, z^p)$.

Lemma 5.3. *Let $\alpha\colon E_1 \to E_2$ be an inseparable isogeny of elliptic curves $E_1\colon y^2 = f_1(x)$ and $E_2\colon y^2 = f_2(x)$ over a field k of characteristic $p > 0$. Then α can be written in the form*

$$\alpha = (a(x^p), b(x^p)y^p)$$

for some rational functions $a, b \in k(x)$.

Proof. Let $\alpha(x, y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$ be in standard form. It follows from Lemma 5.1 that $\frac{u(x)}{v(x)} = a(x^p)$ for some $a \in k(x)$, we only need to show that $\frac{s(x)}{t(x)}y$ can be put in the form $b(x^p)y^p$ for some $b \in k(x)$. As in the proof of Lemma 4.27, substituting u/v and s/t into the equation for E_2 and using the equation for E_1 to eliminate y^2 yields the equality

$$v^3 s^2 f_1 = t^2 w,$$

where $w = v^3 f_2(u/v) \in k[x]$. Since α is inseparable, we have $u' = v' = 0$, hence $w' = 0$, and therefore $\left(w/v^3\right)' = \left(s^2 f_1/t^2\right)' = 0$. Thus $s(x)^2 f_1(x) = g(x^p)$ and $t(x)^2 = h(x^p)$, for some polynomials g and h. If $x_0 \in \bar{k}$ is a root of f_1, then x_0^p is a root of g, so $(x - x_0^p)$ divides g and $(x^p - x_0^p) = (x - x_0)^p$ divides $g(x^p)$.[2] The roots of f_1 are distinct, so $f_1(x)^p$ divides $g(x^p)$ and $g(x^p) = g_1(x^p)f_1(x)^p$ for some $g_1 \in k[x]$.[3]

We have $s(x)^2 f_1(x) = g_1(x)f_1(x)^p$, so $s(x)^2 = g_1(x^p)f_1(x)^{p-1}$. Now p is odd, so $g_1(x^p)$ is a square; indeed, $g_1(x^p) = h_1(x)^2$ where $h_1 = s/f_1^{(p-1)/2} \in k[x]$. We have $(h_1^2)' = 2h_1h_1' = 0$, since $g(x^p)' = 0$, which implies $h_1' = 0$, since h_1 cannot be zero (s is not) and $p \neq 2$. So $h_1(x) = g_2(x^p)$ for some $g_2 \in k[x]$ and $s(x)^2 f_1(x) = g_2(x^p)^2 f_1(x)^p$. We now note that

$$(s(x)y)^2 \equiv s(x)^2 f_1(x) = g_2(x^p)^2 f_1(x)^p \equiv (g_2(x^p)y^p)^2,$$

[2]We are not assuming k is perfect, this argument applies to any k. The key point is that even though $\bar{k}$ may contain inseparable elements, the roots of f_1 are separable (because $\operatorname{disc} f_1 \neq 0$).

[3]Note that $f_1(x)^p$ is not necessarily equal to $f_1(x^p)$, but it is a polynomial in x^p.

where the equivalences are modulo the curve equation for E_1. Therefore

$$\left(\frac{s(x)}{t(x)}y\right)^2 \equiv \left(\frac{g_2(x^p)y^p}{h(x^p)}\right) = (r(x^p)y^p)^2,$$

where $r(x) = g_2(x)/h(x)$. It follows that $\frac{s(x)}{t(x)}y \equiv b(x^p)y^p$ with $b = \pm r$; two rational functions that agree up to sign at infinitely many $\bar{k}$-points can differ only in sign. □

Corollary 5.4. *Let α be an isogeny of elliptic curves over a field k of characteristic $p > 0$. Then*

$$\alpha = \alpha_{\text{sep}} \circ \pi^n$$

for some separable isogeny α_{sep} and integer $n \geq 0$, where π is the p-power Frobenius morphism $(x : y : z) \mapsto (x^p : y^p : z^p)$. We then have $\deg \alpha = p^n \deg \alpha_{\text{sep}}$.

Proof. This holds in general, but we will only prove it for $p > 3$. If α is separable then $\alpha_{\text{sep}} = \alpha$ and $n = 0$, so we now assume α is inseparable. By Lemma 5.3 we may write $\alpha = (r_1(x^p), r_2(x^p)y^p)$ for some $r_1, r_2 \in k(x)$. Then $\alpha = \alpha_1 \circ \pi$ with $\alpha_1 = (r_1(x), r_2(x)y)$. If α_1 is inseparable we apply the same procedure to α_1 (recursively) and eventually obtain $\alpha = \alpha_n \circ \pi^n$ where α_n is a separable isogeny (this process terminates because each step reduces the degree of α_n by a factor of p). We may then take $\alpha_{\text{sep}} = \alpha_n$. If $\alpha_{\text{sep}} = (\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y)$ is in standard form, composing with π^n replaces $u(x)$ by $u(x^{p^n})$ and $v(x)$ by $v(x^{p^n})$, and then $\deg \alpha = \max(p^n \deg u, p^n \deg v) = p^n \max(\deg u, \deg v) = p^n \deg \alpha_{\text{sep}}$. □

Remark 5.5. The isogeny α_{sep} does not necessarily have the same domain as $\alpha\colon E_1 \to E_2$, since the image of π^n is not necessarily E_1 (but π^n will map E_1 to E_1 whenever E_1 is defined over $\mathbb{F}_{p^n}$). We also note that when k is a perfect field (including all finite fields), we can also decompose α as $\alpha = \pi^n \circ \tilde{\alpha}_{\text{sep}}$, where $\tilde{\alpha}_{\text{sep}}$ is separable and has the same degree as α_{sep} (indeed, α_{sep} is just $\tilde{\alpha}_{\text{sep}}$ with each coefficient replaced by its pth power).

Definition 5.6. For an isogeny $\alpha = \alpha_{\text{sep}} \circ \pi^n$ decomposed as in Corollary 5.4, we define the *separable degree* $\deg_s \alpha$ and *inseparable degree* $\deg_i \alpha$ of α as

$$\deg_s \alpha := \deg \alpha_{\text{sep}}, \qquad \deg_i \alpha = p^n,$$

and we always have

$$\deg \alpha = (\deg_s \alpha)(\deg_i \alpha).$$

The inseparable isogeny π^n has separable degree 1; such isogenies are said to be *purely inseparable*. The degree of a purely inseparable isogeny is always a power of p, but the converse does not hold (as we shall see in the next lecture).

Remark 5.7. Note that isogenies of degree 1 (isomorphisms) are both separable and purely inseparable. We are primarily interested in purely inseparable isogenies of degree greater than 1.

We can now prove our first main result.

Theorem 5.8. *The order of the kernel of an isogeny is equal to its separable degree.*

Proof. Let $\alpha = \alpha_{\text{sep}} \circ \pi^n$. Then $\#\ker\alpha = \#\ker\alpha_{\text{sep}}$, since the kernel of π (and hence π^n) is trivial: we can have $(x^p : y^p : z^p) = (0 : 1 : 0)$ if and only if $(x : y : z) = (0 : 1 : 0)$. It thus suffices to consider the case $\alpha = \alpha_{\text{sep}}$, which we now assume.

Let $\alpha(x, y) = (\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y)$ be in standard form and pick a point (a, b) in $\alpha(E_1(\bar{k}))$ with $a, b \neq 0$ and such that a is not equal to the ratio of the leading coefficients of u and v (such a point (a, b) certainly exists, since $\alpha(E_1(\bar{k}))$ is infinite). We now consider the set

$$S(a, b) = \{(x_0, y_0) \in E_1(\bar{k}) : \alpha(x_0, y_0) = (a, b)\}$$

of points in the pre-image of (a, b). Since α is a group homomorphism, $\#S(a, b) = \#\ker\alpha$.

If $(x_0, y_0) \in S(a, b)$ then

$$\frac{u(x_0)}{v(x_0)} = a, \qquad \frac{s(x_0)}{t(x_0)}y_0 = b.$$

We must have $t(x_0) \neq 0$, since α is defined at (x_0, y_0), and $b \neq 0$ implies $s(x_0) \neq 0$. It follows that $y_0 = \frac{t(x_0)}{s(x_0)}b$ is uniquely determined by x_0. Thus to compute $\#S(a, b)$ it suffices to count the number of distinct values of x_0 that occur among the points in $S(a, b)$.

We now let $g = u - av$ so that $\alpha(x_0, y_0) = (a, b)$ if and only if $g(x_0) = 0$. We must have $\deg g = \deg\alpha$, since a is not equal to the ratio of the leading coefficients of u and v (so the leading terms of u and av do not cancel when we subtract them). The cardinality of $S(a, b)$ is then equal to the number of *distinct* roots of g.

Any $x_0 \in \bar{k}$ is a multiple root of g if and only if $g(x_0) = g'(x_0) = 0$, equivalently, if and only if $av(x_0) = u(x_0)$ and $av'(x_0) = u'(x_0)$. If we multiply opposite sides of these equations and cancel the a's we get

$$u'(x_0)v(x_0) = v'(x_0)u(x_0). \tag{1}$$

Now α is separable, so $u'v - v'u \neq 0$ has only a finite number of roots. Since $\alpha(E_1(\bar{k}))$ is infinite and $\#S(a, b) = \#\ker\alpha$ is finite, we may assume that (a, b) was chosen so that (1) is not satisfied for any (x_0, y_0) in $S(a, b)$. Then every root x_0 of g is distinct and we have

$$\#\ker\alpha = \#S(a, b) = \deg g = \deg\alpha,$$

as desired. □

Corollary 5.9. *Every purely inseparable isogeny has trivial kernel.*

Corollary 5.10. *For any composition of isogenies $\alpha = \beta \circ \gamma$ we have*

$$\deg\alpha = (\deg\beta)(\deg\gamma), \qquad \deg_s\alpha = (\deg_s\beta)(\deg_s\gamma), \qquad \deg_i\alpha = (\deg_i\beta)(\deg_i\gamma).$$

Proof. It suffices to prove the last two equalities. The fact that γ is surjective group homomorphism implies

$$\#(\ker\alpha) = \#(\ker\beta)\#(\ker\gamma),$$

since $\ker\alpha$ is the preimage of $\ker\beta$ under γ, which is a union of $\#\ker\beta$ cosets of $\ker\alpha$; Theorem 5.8 implies $\deg_s\alpha = (\deg_s\beta)(\deg_s\gamma)$. Applying Corollary 5.4 to α, β, γ yields

$$\alpha_{\text{sep}} \circ \pi^a = \beta_{\text{sep}} \circ \pi^b \circ \gamma_{\text{sep}} \circ \pi^c.$$

The isogeny $\delta = \pi^b \circ \gamma_{\text{sep}}$ has the same kernel, hence the same separable degree, as γ_{sep}, and we can apply Corollary 5.4 to write it as $\delta = \delta_{\text{sep}} \circ \pi^b$. We then have

$$\alpha_{\text{sep}} \circ \pi^a = \beta_{\text{sep}} \circ \delta_{\text{sep}} \circ \pi^{bc},$$

so $\deg_s \alpha = \deg_s(\beta_{\text{sep}} \circ \delta_{\text{sep}}) = (\deg_s \beta)(\deg_s \delta) = (\deg_s \beta)(\deg_s \gamma)$. We must have $a = bc$ and therefore $\deg_i \alpha = (\deg_i \beta)(\deg_i \gamma)$, since $\beta_{\text{sep}} \circ \delta_{\text{sep}}$ is separable (this follows from the chain rule, the derivative of a composition of functions with nonzero derivative is nonzero). □

5.2 Isogenies from kernels

We have seen that for each isogeny $\alpha\colon E_1 \to E_2$ the kernel of α is a finite subgroup of $E_1(\bar{k})$. It is reasonable to ask whether the converse holds, that is, given a finite subgroup G of $E_1(\bar{k})$, is there an isogeny α from E_1 to some elliptic curve E_2 that has G as its kernel?

The answer is yes. Moreover, if we restrict our attention to separable isogenies (which we should, since if $\alpha = \alpha_{\text{sep}} \circ \pi^n$ then the purely inseparable isogeny π^n has trivial kernel), the isogeny α and the elliptic curve E_2 are uniquely determined up to isomorphism.

The proof of this theorem relies on some standard facts from algebraic geometry that are slightly outside the scope of this course (such as the Hurwitz genus formula), but the theorem is so striking and useful that we will take a moment to sketch the proof. We will then present explicit formulas for constructing α and E_2 from G due to Vélu [2].

Theorem 5.11. *Let E/k be an elliptic curve and let G be a finite subgroup of $E(\bar{k})$. There exists an elliptic curve E' and a separable isogeny $\phi\colon E \to E'$ with $\ker \phi = G$. The curve E' and the isogeny ϕ are defined over a finite extension of k and unique up to isomorphism.*

We can be more precise about the field over which the elliptic curve E' and the isogeny ϕ are defined; it is the minimal extension L/k for which G is invariant under the action of $\mathrm{Gal}(\bar{k}/L)$ (each field automorphism in $\mathrm{Gal}(\bar{k}/k)$ acts on points $P \in E(\bar{k})$ via its action on the coordinates of P); we then say that G is *defined* over L. To say that G is invariant under the action of $\mathrm{Gal}(\bar{k}/L)$ means that the image of G under each $\sigma \in \mathrm{Gal}(\bar{k}/L)$ is G; it does *not* mean that every point in G is necessarily fixed by $\mathrm{Gal}(\bar{k}/L)$, which is a stronger condition (G may be defined over L even when it contains points that are not).

Proof sketch. Given any smooth projective curve C and a finite group G of automorphisms of the curve (invertible morphisms from the curve to itself), there is a smooth projective curve C/G and a surjective morphism $\phi\colon C \to C/G$ that maps each G-orbit $\{\sigma(P) : \sigma \in G\}$ of points $P \in C(\bar{k})$ to a distinct point in C/G. The curve C/G is called the *quotient* of C by G. The standard way to prove this is to use the categorical equivalence of smooth projective curves and their function fields to derive C/G and ϕ from the field embedding

$$k(C)^{G^*} \hookrightarrow k(C),$$

where G^* denotes the group of automorphisms $\sigma^*\colon k(C) \to k(C))$ induced by the automorphisms $\sigma\colon C \to C$ in G (so $\sigma^*(f) = f \circ \sigma$), and $k(C)^{G^*}$ is the subfield of $k(C)$ fixed by G^*. The morphism ϕ is separable because $k(C)/k(C)^{G^*}$ is separable, and provided that the group G is defined over k, both ϕ and C/G are defined over k (if not we can always base change E to the minimal field over which G is defined).

In our situation the curve C is an elliptic curve, and we can associate to each point $P \in E(\bar{k})$ the automorphism $\tau_P\colon Q \mapsto Q + P$, the *translation-by-$P$ map*. Note that τ_P is not an isogeny because it does not fix the point 0 (unless $P = 0$), but it is a morphism $E \to E$, and it has an inverse τ_{-P}, so it is an automorphism. We can thus associate a group of automorphisms G to any finite subgroup of $E(\bar{k})$, consisting of translation-by-P maps τ_P for each $P \in G$, and we obtain a morphism $\phi\colon E \to E/G$ from E to its quotient by G.

It is not immediately clear that the smooth projective curve E/G is actually an elliptic curve, but this is indeed the case. This follows from the Hurwitz genus formula [1, II.2.7], which implies that for any *unramified* morphism $\phi\colon C_1 \to C_2$ we must have

$$(2g_1 - 2) = (\deg \phi)(2g_2 - 2).$$

Here g_i denotes the genus of C_i, and ϕ is unramified if its fibers $\phi^{-1}(P) \subseteq C_1(\bar{k})$ have the same cardinality for every point $P \in C_2(\bar{k})$.

In our situation $\phi\colon E \to E/G$ is unramified because the G-orbits of $E(\bar{k})$ are cosets, which necessarily all have the same size, and the Hurwitz genus formula then implies that E/G must have genus 1 (since E has genus 1), no matter what the degree of ϕ is.[4] Assuming G is defined over k, the point $\phi(0)$ will be rational and we can take it as our distinguished rational point (in any case $\phi(0)$ will be defined over the field of definition of E/G). So E/G is an elliptic curve, and $\phi\colon E \to E/G$ is a surjective morphism that fixes the identity, hence an isogeny, and as noted above, it is separable. The kernel of ϕ is the G-orbit of 0 in $E(\bar{k})$, which is precisely the subgroup G of $E(\bar{k})$ that we started with.

Moreover, if we have another separable isogeny $\phi'\colon E \to E'$ with the same kernel G, then we can view $k(E')$ as a subfield of $k(E)$ via the induced embedding $\phi'^*\colon k(E') \to k(E)$, and $k(E')$ is then fixed by every automorphism in G, hence a subgroup of $k(E)^G$. Since ϕ' is separable, we have $\deg \phi = [k(E)\colon k(E')] = \#G$, so $k(E')$ must be (isomorphic to) the fixed field $k(E)^G$. It follows that there exists an isomorphism $\iota\colon E/G \xrightarrow{\sim} E'$ for which $\phi' = \iota \circ \phi$; the curve E/G and the isogeny ϕ are thus unique up to such an isomorphism. □

Corollary 5.12. *An isogeny of composite degree can always be decomposed into a sequence of isogenies of prime degree.*

Proof. Let $\alpha\colon E_1 \to E_2$ be an isogeny. If we are working in a field of characteristic $p > 0$, by writing α as $\alpha = \alpha_{\text{sep}} \circ \pi^n$ we can decompose $\pi^n = \pi \circ \cdots \circ \pi$ as a sequence of isogenies of prime degree p. Thus it suffices to consider the case where α is separable. As a non-trivial abelian group, $G = \ker \alpha$ contains a subgroup H of prime order. By Theorem 5.11, there exists a separable isogeny $\alpha_1\colon E_1 \to E_3$ with H as its kernel. Then $\alpha_1(G)$ is a finite subgroup of $E_3(\bar{k})$ isomorphic to G/H, and (applying Theorem 5.11 again), there exists a separable isogeny $\alpha_2\colon E_3 \to E_4$ with $\alpha_1(G)$ as its kernel. The kernel of the composition $\alpha_2 \circ \alpha_1$ is $G = \ker \alpha$, so there exists an isomorphism $\iota\colon E_4 \to E_2$ such that $\alpha = \iota \circ \alpha_2 \circ \alpha_1$.

We can now proceed by induction and apply the same decomposition to $\iota \circ \alpha_2$, which has smaller degree than α. We eventually obtain a sequence of separable isogenies of prime degree whose composition is equal to α. □

This is all very nice from an abstract point of view, but it is not immediately useful for practical applications. We would really like to have an explicit description of the elliptic curve E/G and the isogeny ϕ. So let $E\colon y^2 = x^3 + Ax + B$ be an elliptic curve and let G be a finite subgroup of $E(\bar{k})$. Let $G_{\neq 0}$ denote the set of nonzero points in G, all of which are affine points $Q = (x_Q, y_Q)$, and for each point $P = (x_P, y_P)$ in $E(\bar{k})$ that is not in G, let us define

$$\phi(P) := \left(x_P + \sum_{Q \in G_{\neq 0}} (x_{P+Q} - x_Q)\,,\ y_P + \sum_{Q \in G_{\neq 0}} (y_{P+Q} - y_Q) \right).$$

[4] This is yet another remarkable property of elliptic curves; isogenies $\phi\colon E \to E'$ are necessarily unramified and we always get zero on both sides of the Hurwitz genus formula (allowing ϕ to have any degree); this phenomenon does not occur for curves of any other genus.

Here x_P and y_P are variables, x_Q and y_Q are fixed elements of $\bar{k}$, and x_{P+Q} and y_{P+Q} are the affine coordinates of $P+Q$, which we can view as rational functions of x_P and y_P by plugging the coordinates of P and Q into the formulas for the group law.

It's not immediately obvious what the image of this map is, but it is clearly a non-constant rational map, so it defines a morphism from E to some smooth projective curve E'. Moreover, we can see that the group law on E induces a group law on E' that is defined by rational maps, thus E' is an abelian variety (of dimension one), hence an elliptic curve. For any $P \notin G$ we have $\phi(P) = \phi(P+Q)$ if and only if $Q \in G$, so the kernel of ϕ must be G.

Thus, assuming it is separable, ϕ is the isogeny we are looking for (up to isomorphism). By using the group law to write x_{P+Q} and y_{P+Q} as rational functions in terms of x_P and y_P (and the coordinates of the points in G, which we regard as constants), we can get explicit equations for ϕ and determine an equation for its image E'. The details are somewhat involved (see [3, Thm. 12.16]), so we will just give the formulas. To simplify the expressions we will assume that the order of G is either 2 or odd; this covers all separable isogenies of prime degree, and by the corollary above, we can obtain any isogeny by composing separable isogenies of prime degree and copies of the Frobenius morphism (if necessary).

Theorem 5.13 (Vélu)**.** *Let $E\colon y^2 = x^3 + Ax + B$ be an elliptic curve over k and let $x_0 \in \bar{k}$ be a root of $x^3 + Ax + B$. Define $t := 3x_0^2 + A$ and $w := x_0 t$. The rational map*

$$\phi(x,y) := \left(\frac{x^2 - x_0 x + t}{x - x_0},\ \frac{(x-x_0)^2 - t}{(x-x_0)^2} y \right)$$

is a separable isogeny from $E_{\bar{k}}$ to $E'_{\bar{k}}\colon y^2 = x^3 + A'x + B'$, where $A' := A - 5t$ and $B' := B - 7w$. The kernel of ϕ is the group of order 2 generated by $(x_0, 0)$.

Proof. It is clear that ϕ is a separable isogeny of degree 2 with $(x_0, 0)$ in its kernel. The only thing to check is that E' is its image, which is an easy verification (just plug the formulas for $\phi(x,y)$ into the equation for E'). □

Remark 5.14. If $x_0 \in k$ then ϕ and E' will both be defined over k, but in general they will be defined over the extension field $k(x_0)$ which contains A' and B'.

Theorem 5.15 (Vélu)**.** *Let $E\colon y^2 = x^3 + Ax + B$ be an elliptic curve over k and let G be a finite subgroup of $E(\bar{k})$ of odd order. For each nonzero $Q = (x_Q, y_Q)$ in G define*

$$t_Q := 3x_Q^2 + A, \qquad u_Q := 2y_Q^2, \qquad w_Q := u_Q + t_Q x_Q,$$

and let

$$t := \sum_{Q \in G_{\neq 0}} t_Q, \qquad w := \sum_{Q \in G_{\neq 0}} w_Q, \qquad r(x) := x + \sum_{Q \in G_{\neq 0}} \left(\frac{t_Q}{x - x_Q} + \frac{u_Q}{(x - x_Q)^2} \right).$$

The rational map

$$\phi(x,y) := \big(r(x),\ r'(x)y\big)$$

defines a separable isogeny from $E_{\bar{k}}$ to $E'_{\bar{k}}\colon y^2 = x^3 + A'x + B'$, where $A' := A - 5t$ and $B' := B - 7w$, with $\ker \phi = G$.

Proof. This is a special case of [3, Thm. 12.16]. □

Remark 5.16. The formulas for $t, w, r(x)$ sum over all the nonzero points in G but notice that they depend only on the x-coordinates x_Q. Since $|G|$ is odd and $Q = (x_Q, y_Q) \in G$ if and only if $-Q = (x_Q, -y_Q) \in G$, it suffices to sum over just half the points in $G_{\neq 0}$ (representatives of $G/\{\pm 1\}$), and double the result. The elliptic curve E' and ϕ are defined over any extension L/k where G is defined.

Remark 5.17. Theorem 5.15 implies that (possibly after composing with an isomorphism) we can put any separable isogeny α of odd degree in the form

$$\alpha(x, y) = \left(\frac{u}{w^2}, \left(\frac{u}{w^2}\right)' y\right) = \left(\frac{u}{w^2}, \frac{u'w - 2w'u}{w^3} y\right),$$

for some relatively prime polynomials u and w in $k[x]$.

5.3 Jacobian coordinates

We now turn to the multiplication-by-n map $P \mapsto nP$, which we will denote by $[n]$. We want to write the isogeny $[n]$ in standard form. To do this, it turns out to be more convenient to work with *Jacobian coordinates*, which we now define.

Recall that points in standard projective coordinates are nonzero triples $(x : y : z)$ subject to the equivalence relation

$$(x : y : z) \sim (\lambda x : \lambda y : \lambda z),$$

for any $\lambda \in k^\times$. We will instead work with the equivalence relation

$$(x : y : z) \sim (\lambda^2 x : \lambda^3 y : \lambda z),$$

which corresponds to assigning *weights* 2, 3, 1 to the variables x, y, z, respectively. Projective coordinates with these weights are called *Jacobian coordinates*. The homogeneous curve equation for E in Jacobian coordinates then has the form

$$y^2 = x^3 + Axz^4 + Bz^6,$$

which makes visible the motivation for giving x weight 2 and y weight 3: the leading terms for x and y do not involve z. In Jacobian coordinates, each point $(x : y : z)$ with $z \neq 0$ corresponds to the affine point $(x/z^2, y/z^3)$, and the point at infinity is now $(1 : 1 : 0)$.

Remark 5.18. As an aside, the general Weierstrass form of an elliptic curve in Jacobian coordinates is

$$y^2 + a_1xyz + a_3yz^3 = x^3 + a_2x^2z^2 + a_4xz^4 + a_6z^6,$$

which is a weighted homogeneous equation of degree 6. Each a_i is the coefficient of a term with degree i in z. This explains the otherwise mysterious fact that there is no Weierstrass coefficient a_5.

5.4 The group law in Jacobian coordinates

We now compute formulas for the elliptic curve group law in Jacobian coordinates, beginning with addition. Recall that in affine coordinates, to compute the sum $P_3 = (x_3, y_3)$ of two affine points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ with $P_1 \neq \pm P_2$ we use the formulas

$$x_3 = m^2 - (x_1 + x_2) \quad \text{and} \quad y_3 = m(x_1 - x_3) - y_1,$$

where $m = \frac{y_1-y_2}{x_1-x_2}$ is the slope of the line through P_1 and P_2. In Jacobian coordinates we have $P_i = (x_i/z_i^2, y_i/z_i^3)$ and the formula for the x-coordinate becomes

$$\frac{x_3}{z_3^2} = \left(\frac{y_1/z_1^3 - y_2/z_2^3}{x_1/z_1^2 - x_2/z_2^2}\right)^2 - \left(\frac{x_1}{z_1^2} + \frac{x_2}{z_2^2}\right) = \frac{(y_1z_2^3 - y_2z_1^3)^2 - (x_1z_2^2 + x_2z_1^2)(x_1z_2^2 - x_2z_1^2)^2}{(x_1z_2^2 - x_2z_1^2)^2 z_1^2 z_2^2}.$$

This formula can be simplified by using $y_i^2 - x_i^3 = Ax_iz_i^4 + Bz_i^6$ to get rid of the terms in the numerator containing y_i^2 or x_i^3. This makes the numerator divisible by $z_1^2z_2^2$ allowing us to cancel this with the corresponding factor in the denominator. We have

$$\begin{aligned}\frac{x_3}{z_3^2} &= \frac{(y_1^2z_2^6 - x_1^3z_2^6) + (y_2^2z_1^6 - x_2^3z_1^6) + x_1^2x_2z_1^2z_2^4 + x_1x_2^2z_1^4z_2^2 - 2y_1y_2z_1^3z_2^3}{(x_1z_2^2 - x_2z_1^2)^2 z_1^2 z_2^2} \\ &= \frac{(Ax_1z_1^4 + Bz_1^6)z_2^6 + (Ax_2z_2^4 + Bz_2^6)z_1^6 + x_1^2x_2z_1^2z_2^4 + x_1x_2^2z_1^4z_2^2 - 2y_1y_2z_1^3z_2^3}{(x_1z_2^2 - x_2z_1^2)^2 z_1^2 z_2^2} \\ &= \frac{A(x_1z_2^2 + x_2z_1^2)z_1^2z_2^2 + 2Bz_1^4z_2^4 - 2y_1y_2z_1z_2}{(x_1z_2^2 - x_2z_1^2)^2}.\end{aligned}$$

For the y-coordinate, using $y_3 = m(x_1 - x_3) - y_1 = m(2x_1 + x_2) - m^3 - y_1$ we have

$$\begin{aligned}\frac{y_3}{z_3^3} &= \left(\frac{y_1/z_1^3 - y_2/z_2^3}{x_1/z_1^2 - x_2/z_2^2}\right)\left(\frac{2x_1}{z_1^2} + \frac{x_2}{z_2^2}\right) - \left(\frac{y_1/z_1^3 - y_2/z_2^3}{x_1/z_1^2 - x_2/z_2^2}\right)^3 - \frac{y_1}{z_1^3} \\ &= \frac{(y_1z_2^3 - y_2z_1^3)(2x_1z_2^2 + x_2z_1^2)(x_1z_2^2 - x_2z_1^2)^2 - (y_1z_2^3 - y_2z_1^3)^3 - y_1z_2^3(x_1z_2^2 - x_2z_1^2)^3}{(x_1z_2^2 - x_2z_1^2)^3 z_1^3 z_2^3} \\ &= \frac{\cdots}{(x_1z_2^2 - x_2z_1^2)^3}\end{aligned}$$

where the missing numerator is some complicated polynomial in $x_1, y_1, z_1, x_2, y_2, z_2, A, B$. These formulas look horrible, but the key point is in Jacobian coordinates we now have

$$z_3 = x_1z_2^2 - x_2z_1^2, \tag{2}$$

which is actually a lot simpler than it would have otherwise been; note that the z-coordinate is the most interesting to us, because it will determine the kernel we are interested in.

The doubling formulas are simpler. In affine coordinates the slope of the tangent line is $m = (3x_1^2 + A)/(2y_1)$. For the x-coordinate we have

$$\frac{x_3}{z_3^2} = \left(\frac{3(x_1/z_1^2)^2 + A}{2y_1/z_1^3}\right)^2 - 2\frac{x_1}{z_1^2} = \frac{(3x_1^2 + Az_1^4)^2 - 8x_1y_1^2}{(2y_1z_1)^2} = \frac{x_1^4 - 2Ax_1^2z_1^4 - 8Bx_1z_1^6 + A^2z_1^8}{(2y_1z_1)^2}$$

and for the y-coordinate we get

$$\begin{aligned}\frac{y_3}{z_3^3} &= \left(\frac{3(x_1/z_1^2)^2 + A}{2y_1/z_1^3}\right)\frac{3x_1}{z_1^2} - \left(\frac{3(x_1/z_1^2)^2 + A}{2y_1/z_1^3}\right)^3 - \frac{y_1}{z_1^3} \\ &= \frac{12x_1y_1^2(3x_1^2 + Az_1^4) - (3x_1^2 + Az_1^4)^3 - 8y_1^4}{(2y_1z_1)^3} \\ &= \frac{x_1^6 + 5Ax_1^4z_1^4 + 20Bx_1^3z_1^6 - 5A^2x_1^2z_1^8 - 4ABx_1z_1^{10} - (A^3 + 8B^2)z_1^{12}}{(2y_1z_1)^3}.\end{aligned}$$

Thus

$$z_3 = 2y_1z_1. \tag{3}$$

5.5 Division polynomials

We now wish to apply our addition formulas to a "generic" point $P = (x : y : 1)$ on the elliptic curve E defined by $y^2 = x^3 + Ax + B$, and use them to compute $2P, 3P, 4P, \ldots, nP$. In Jacobian coordinates, the point nP has the form $(\phi_n : \omega_n : \psi_n)$, where ϕ_n, ω_n, and ψ_n are integer polynomials in x, y, A, B that we reduce modulo the curve equation so that the degree in y is at most 1. In affine coordinates we then have

$$nP = \left(\frac{\phi_n}{\psi_n^2}, \frac{\omega_n}{\psi_n^3}\right). \tag{4}$$

We will see that ϕ_n and ψ_n^2 do not depend on y, so for fixed A and B they are univariate polynomials in x, and exactly one of ω_n and ψ_n^3 depends on an odd power of y, so this will give us $[n]$ in standard form. This Sage notebook computes the polynomials ϕ_n, ω_n, ψ_n for the first several values of n.

Remark 5.19. Another way to think of division polynomials is to view E as an elliptic curve over $k(E)$. In concrete terms, let F be the fraction field of the ring $k[x, y]/(y^2 - x^3 - Ax - B)$, and let P be the affine point $(x, y) \in E(F)$, which is by construction a point on E of infinite order. Equation (4) then gives the coordinates of the point $nP \in E(F)$.

The polynomial ψ_n is known as the nth *division polynomial*. So far we have really only defined the ratios ϕ_n/ψ_n^2 and ω_n/ψ_n^3, since we have been working in projective coordinates. In order to nail down ϕ_n ω_n and ψ_n precisely, we make the following recursive definition. Let $\psi_0 = 0$, and define $\psi_1, \psi_2, \psi_3, \psi_4$ via the formulas:

$$\begin{aligned}
\psi_1 &= 1,\\
\psi_2 &= 2y,\\
\psi_3 &= 3x^4 + 6Ax^2 + 12Bx - A^2,\\
\psi_4 &= 4y(x^6 + 5Ax^4 + 20Bx^3 - 5A^2x^2 - 4ABx - A^3 - 8B^2).
\end{aligned}$$

These are exactly the same polynomials computed in the Sage worksheet linked to above (up to a sign). We then define the division polynomials ψ_n for integers $n > 4$ via the recurrences

$$\begin{aligned}
\psi_{2n+1} &= \psi_{n+2}\psi_n^3 - \psi_{n-1}\psi_{n+1}^3,\\
\psi_{2n} &= \frac{1}{2y}\psi_n(\psi_{n+2}\psi_{n-1}^2 - \psi_{n-2}\psi_{n+1}^2),
\end{aligned}$$

where we reduce the result modulo the curve equation so that ψ_n is at most linear in y. It is not difficult to show that $\psi_n(\psi_{n+2}\psi_{n-1}^2 - \psi_{n-2}\psi_{n+1}^2)$ is always divisible by $2y$, so that ψ_{2n} is in fact a polynomial; see Lemma 5.20 below. If we define $\psi_{-n} := -\psi_n$, one can check that these recurrences hold for all integers n.

We then define ϕ_n and ω_n via

$$\begin{aligned}
\phi_n &:= x\psi_n^2 - \psi_{n+1}\psi_{n-1},\\
\omega_n &:= \frac{1}{4y}(\psi_{n+2}\psi_{n-1}^2 - \psi_{n-2}\psi_{n+1}^2).
\end{aligned}$$

These equations hold for all integers n, and one finds that $\phi_n = \phi_{-n}$ and $\omega_n = \omega_{-n}$. As above, we reduce ϕ_n and ω_n modulo the curve equation to make them at most linear in y.

Lemma 5.20. *For every integer n,*

$$\begin{aligned}
\psi_n \text{ lies in } &\begin{cases} \mathbb{Z}[x,A,B] & n \text{ odd} \\ 2y\mathbb{Z}[x,A,B] & n \text{ even,} \end{cases} \\
\phi_n \text{ lies in } &\quad \mathbb{Z}[x,A,B] \qquad \text{for all } n, \\
\omega_n \text{ lies in } &\begin{cases} \mathbb{Z}[x,A,B] & n \text{ even} \\ y\mathbb{Z}[x,A,B] & n \text{ odd.} \end{cases}
\end{aligned}$$

Proof. These are easy inductions; see [3, Lemmas 3.3 and 3.4]. □

It follows from the lemma that, after replacing y^2 with x^3+Ax+B if necessary, ψ_n^2 lies in $\mathbb{Z}[x,A,B]$ for all positive n, so we view ϕ_n and ψ_n^2 as polynomials in x, while exactly one of ω_n and ψ_n^3 depends on y. In the latter case we can multiply the numerator and denominator of ω_n/ψ_n^3 by y and then replace y^2 in the denominator with x^3+Ax+B so that $\omega_n/\psi_n \in y\mathbb{Z}(x,A,B)$. With this understanding, we can view

$$\left(\frac{\phi_n(x)}{\psi_n^2(x)}, \frac{\omega_n(x,y)}{\psi_n^3(x,y)}\right)$$

as an isogeny in standard form provided that the numerators and denominators are relatively prime (which we will verify below).

5.6 Multiplication-by-n maps

At this point it is not at all obvious that the polynomials ϕ_n, ω_n,ψ_n defined by our recursive equations actually satisfy equation (4) for nP, but this is indeed the case.

Theorem 5.21. *Let E/k be an elliptic curve defined by the equation $y^2 = x^3+Ax+B$ and let n be a nonzero integer. The rational map*

$$[n](x,y) = \left(\frac{\phi_n(x)}{\psi_n^2(x)}, \frac{\omega_n(x,y)}{\psi_n^3(x,y)}\right)$$

sends each point $P \in E(\bar{k})$ to nP.

Proof. We have

$$[-n](x,y) = \left(\frac{\phi_{-n}(x)}{\psi_{-n}^2(x)}, \frac{\omega_{-n}(x,y)}{\psi_{-n}^3(x,y)}\right) = \left(\frac{\phi_n(x)}{\psi_n^2(x)}, \frac{\omega_n(x,y)}{-\psi_n^3(x,y)}\right) = -\left(\frac{\phi_n(x)}{\psi_n^2(x)}, \frac{\omega_n(x,y)}{\psi_n^3(x,y)}\right),$$

so it suffices to consider positive n. The proof given in [3, Thm. 9.33] uses complex analysis and the Weierstrass $\wp$-function, which we will see later in the course. However, as noted in [1, Ex. 3.7], one can give a purely algebraic proof by induction, using the formulas for the group law. This approach has the virtue of being completely elementary and works over any field, but it is computationally intensive (and really should be done with a computer algebra system).[5] Here we will just verify that the formulas for ψ_n are correct; the verifications for ϕ_n and ω_n are similar.

[5] If k has characteristic 2 or 3 one needs to modify the formulas to use a general Weierstrass equation; this changes ψ_2, ψ_3, ψ_4, and the recurrence for ω_n, but the recurrences for ϕ_n and ψ_n are unaffected. Be aware that there are a few typos in the formulas given in [1, Ex. 3.7] on page 105 that are corrected in the errata.

For $1 \le n \le 4$ the formulas given for ψ_n match our computations in Sage using the group law. To verify the formula for ψ_n when $n = 2m+1 > 4$ is odd, we let P_m be the point $(\phi_m, \omega_m, \psi_m)$ in Jacobian coordinates and compute $P_m + P_{m+1}$ using the group law. The z-coordinate of the sum is given by the formula $z_3 = x_1 z_2^2 - x_2 z_1^2$ from (2). Substituting ϕ_m for x_1, ψ_m for z_1, ϕ_{m+1} for x_2, and ψ_{m+1} for z_2 yields

$$\phi_m \psi_{m+1}^2 - \phi_{m+1}\psi_m^2,$$

which we wish to show is equal to ψ_{2m+1}. Applying the formulas for ϕ_m and ϕ_{m+1} gives

$$\begin{aligned}\phi_m \psi_{m+1}^2 - \phi_{m+1}\psi_m^2 &= (x\psi_m^2 - \psi_{m+1}\psi_{m-1})\psi_{m+1}^2 - (x\psi_{m+1}^2 - \psi_{m+2}\psi_m)\psi_m^2 \\ &= \psi_{m+2}\psi_m^3 - \psi_{m-1}\psi_{m+1}^3 \\ &= \psi_{2m+1},\end{aligned}$$

To verify the formula for ψ_n when $n = 2m > 4$ is even, we now compute $P_m + P_m$. The z-coordinate of the sum is given by the formula $z_3 = 2y_1 z_1$ from (3). We then have

$$\begin{aligned}2\omega_m\psi_m &= 2 \cdot \frac{1}{4y}(\psi_{m+2}\psi_{m-1}^2 - \psi_{m-2}\psi_{m+1}^2)\psi_m \\ &= \psi_{2m}.\end{aligned}$$

as desired. This completes the verification for ψ_n. To complete the proof one performs a similar verification for ϕ_n and ω_n using the group law formulas for x_3 and y_3 in Jacobian coordinates that we derived earlier. □

To compute the degree of $[n]\colon E \to E$, we need to know the degrees of the polynomials $\phi_n(x)$ and $\psi_n^2(x)$, and we need to verify that they are relatively prime.

Lemma 5.22. *For every positive integer n the polynomials ϕ_n and ψ_n satisfy*

$$\begin{aligned}\phi_n(x) &= x^{n^2} + \cdots, \\ \psi_n(x) &= \begin{cases} nx^{\frac{n^2-1}{2}} + \cdots, & n \text{ odd} \\ y\left(nx^{\frac{n^2-4}{2}} + \cdots\right), & n \text{ even.}\end{cases}\end{aligned}$$

where each ellipsis hides terms of lower degree in x.

Proof. We first prove the formula for ψ_n by induction on n. By inspection, the formulas hold for $n = 1, 2, 3, 4$. There are then four cases to consider, depending on the value of n mod 4. For any polynomial $f(x, y)$ we let $\mathrm{lt}_x f$ denote the leading term of f as a polynomial in x.

Case 0: $n \equiv 0 \bmod 4$. Let $n = 2m$, with m even. We have

$$\begin{aligned}\mathrm{lt}_x \psi_{2m} &= \mathrm{lt}_x\left(\frac{1}{2y}\psi_m(\psi_{m+2}\psi_{m-1}^2 - \psi_{m-2}\psi_{m+1}^2)\right) \\ &= \frac{1}{2y} \cdot ymx^{\frac{m^2-4}{2}}\left(y(m+2)x^{\frac{(m+2)^2-4}{2}}(m-1)^2 x^{\frac{2(m-1)^2-2}{2}} - y(m-2)x^{\frac{(m-2)^2-4}{2}}(m+1)^2 x^{\frac{2(m+1)^2-2}{2}}\right) \\ &= \frac{ym}{2}\left((m-1)^2(m+2)x^{\frac{m^2-4+m^2+4m+4-4+2m^2-4m}{2}} - (m-2)(m+1)^2 x^{\frac{m^2-4+m^2-4m+4-4+2m^2+4m}{2}}\right) \\ &= \frac{ym}{2}\left((m-1)^2(m+2) - (m-2)(m+1)^2\right)x^{\frac{4m^2-4}{2}} \\ &= y(2m)x^{\frac{4m^2-4}{2}} = ynx^{\frac{n^2-4}{2}}.\end{aligned}$$

Case 1: $n \equiv 1 \bmod 4$. Let $n = 2m+1$, with m even. We have

$$\begin{aligned}
\mathrm{lt}_x\psi_{2m+1} &= \mathrm{lt}_x\left(\psi_{m+2}\psi_m^3 - \psi_{m-1}\psi_{m+1}^3\right)\\
&= \mathrm{lt}_x\left(y(m+2)x^{\frac{(m+2)^2-4}{2}}y^3m^3x^{\frac{3m^2-12}{2}} - (m-1)x^{\frac{(m-1)^2-1}{2}}(m+1)^3x^{\frac{3(m+1)^2-3}{2}}\right)\\
&= (m+2)m^3x^6x^{\frac{m^2+4m+3m^2-12}{2}} - (m-1)(m+1)^3x^{\frac{m^2-2m+3m^2+6m}{2}}\\
&= (2m+1)x^{\frac{4m^2+4m}{2}} = nx^{\frac{n^2-1}{2}}.
\end{aligned}$$

Here we used the curve equation to replace y^4 with x^6, the leading term of $(x^3+Ax+B)^2$.

Case 2: $n \equiv 2 \bmod 4$. Let $n = 2m$, with m odd. We have

$$\begin{aligned}
\mathrm{lt}_x\psi_{2m} &= \mathrm{lt}_x\left(\frac{1}{2y}\psi_m(\psi_{m+2}\psi_{m-1}^2 - \psi_{m-2}\psi_{m+1}^2)\right)\\
&= \frac{1}{2y}mx^{\frac{m^2-1}{2}}\left((m+2)x^{\frac{(m+2)^2-1}{2}}y^2(m-1)^2x^{\frac{2(m-1)^2-8}{2}} - (m-2)x^{\frac{(m-2)^2-1}{2}}y^2(m+1)^2x^{\frac{2(m+1)^2-8}{2}}\right)\\
&= \frac{y}{2}m\left((m+2)(m-1)^2x^{\frac{m^2-1+(m+2)^2-1+2(m-1)^2-8}{2}} - (m-2)(m+1)^2x^{\frac{m^2-1+(m-2)^2-1+2(m+1)^2-8}{2}}\right)\\
&= \frac{y}{2}m\left((m+2)(m-1)^2 - (m-2)(m+1)^2\right)x^{\frac{4m^2-4}{2}}\\
&= y(2m)x^{\frac{4m^2-4}{2}} = ynx^{\frac{n^2-4}{2}}.
\end{aligned}$$

Case 3: $n \equiv 3 \bmod 4$. Let $n = 2m+1$, with m odd. We have

$$\begin{aligned}
\mathrm{lt}_x\psi_{2m+1} &= \mathrm{lt}_x\left(\psi_{m+2}\psi_m^3 - \psi_{m-1}\psi_{m+1}^3\right)\\
&= \mathrm{lt}_x\left((m+2)x^{\frac{(m+2)^2-1}{2}}m^3x^{\frac{3m^2-3}{2}} - y(m-1)x^{\frac{(m-1)^2-4}{2}}y^3(m+1)^3x^{\frac{3(m+1)^2-12}{2}}\right)\\
&= (2m+1)x^{\frac{4m^2+4m}{2}}\\
&= nx^{\frac{n^2-1}{2}}.
\end{aligned}$$

Here we have again used the curve equation to replace y^4 with x^6.

Now that we have verified the formulas for ψ_n, we need to check ϕ_n. There are two cases, depending on the parity of n. If n is even we have

$$\begin{aligned}
\mathrm{lt}_x\phi_n &= \mathrm{lt}_x\left(x\psi_n^2 - \psi_{n+1}\psi_{n-1}\right)\\
&= \mathrm{lt}_x\left(xy^2n^2x^{\frac{2n^2-8}{2}} - (n+1)x^{\frac{(n+1)^2-1}{2}}(n-1)x^{\frac{(n-1)^2-1}{2}}\right)\\
&= n^2x^{n^2} - (n^2-1)x^{n^2}\\
&= x^{n^2},
\end{aligned}$$

and if n is odd we have

$$\begin{aligned}
\mathrm{lt}_x\phi_n &= \mathrm{lt}_x\left(x\psi_n^2 - \psi_{n+1}\psi_{n-1}\right)\\
&= \mathrm{lt}_x\left(xn^2x^{n^2-1} - y(n+1)x^{\frac{(n+1)^2-4}{2}}y(n-1)x^{\frac{(n-1)^2-4}{2}}\right)\\
&= n^2x^{n^2} - (n^2-1)x^{n^2}\\
&= x^{n^2},
\end{aligned}$$

where we have used the curve equation to replace y^2 with x^3. □

Corollary 5.23. *For all positive integers n, we have $\psi_n^2(x) = n^2 x^{n^2-1} + \cdots$, where the ellipsis denotes terms of degree less than $n^2 - 1$.*

Lemma 5.24. *Let E/k be an elliptic curve defined by $y^2 = x^3 + Ax + B$. The polynomials $\phi_n(x)$ and $\psi_n^2(x)$ are relatively prime.*

Proof. Suppose not. Let $x_0 \in \bar{k}$ be a common root of $\phi_n(x)$ and $\psi_n^2(x)$, and let $P = (x_0, y_0)$ be a nonzero point in $E(\bar{k})$. Then $nP = 0$, since $\psi_n^2(x_0) = 0$, and we also have

$$\begin{aligned} \phi_n(x_0) &= x_0\psi_n^2(x_0) - \psi_{n+1}(x_0, y_0)\psi_{n-1}(x_0, y_0) \\ 0 &= 0 - \psi_{n+1}(x_0, y_0)\psi_{n-1}(x_0, y_0), \end{aligned}$$

so at least one of $\psi_{n+1}(x_0, y_0)$ and $\psi_{n-1}(x_0, y_0)$ is zero. But then either $(n-1)P = 0$ or $(n+1)P = 0$, and after subtracting $nP = 0$ we see that either $-P = 0$ or $P = 0$, which is a contradiction. □

Theorem 5.25. *Let E/k be an elliptic curve. The multiplication-by-n map $[n]\colon E \to E$ has degree n^2. It is separable if and only if n is not divisible by the characteristic of k.*

Proof. From Lemma 5.22, we have $\deg \phi_n = n^2$ and $\deg \psi_n^2 \le n^2 - 1$, and from Lemma 5.24 we know that $\phi_n \perp \psi_n^2$. It follows that $\deg[n] = n^2$. If n is not divisible by the characteristic of k, then the leading term $n^2 x^{n^2-1}$ of $\phi_n'(x)$ is nonzero and therefore

$$\left(\frac{\phi_n(x)}{\psi_n^2(x)}\right)' \neq 0,$$

which implies that $[n]$ is separable. If n is divisible by the characteristic of k then the $n^2 x^{n^2-1}$ leading term in ψ_n^2 vanishes and $\deg \psi_n^2$ is less than $n^2 - 1$. This implies that the kernel of $[n]$, which consists of 0 and the affine points (x_0, y_0) for which $\psi_n(x_0) = 0$, is strictly smaller than its degree n^2, in which case $[n]$ must be inseparable, by Theorem 5.8. □

References

[1] J. H. Silverman, *The Arithmetic of Elliptic Curves*, Graduate Texts in Mathematics **106**, second edition, Springer 2009.

[2] J. Vélu, *Isogénies entre courbe elliptiques*, C. R. Acad. Sci. Paris Séries A **273** (1971), 238–241, English translation by Alex Ghitza.

[3] L. C. Washington, *Elliptic Curves: Number Theory and Cryptography*, second edition, Chapman and Hall/CRC, 2008.

6 Torsion subgroups and endomorphism rings

6.1 The n-torsion subgroup $E[n]$

Having determined the degree and separability of the multiplication-by-n map $[n]$ in the previous lecture, we now want to determine the structure of its kernel, the n-torsion subgroup $E[n]$, as a finite abelian group. Recall that any finite abelian group G can be written as a direct sum of cyclic groups of prime power order (unique up to ordering). Since $\#E[n]$ always divides $\deg[n] = n^2$, to determine the structure of $E[n]$ it suffices to determine the structure of $E[\ell^e]$ for each prime power ℓ^e dividing n.

Theorem 6.1. *Let E/k be an elliptic curve and let $p := \mathrm{char}(k)$. For each prime ℓ:*

$$E[\ell^e] \simeq \begin{cases} \mathbb{Z}/\ell^e\mathbb{Z} \oplus \mathbb{Z}/\ell^e\mathbb{Z} & \text{if } \ell \neq p, \\ \mathbb{Z}/\ell^e\mathbb{Z} \text{ or } \{0\} & \text{if } \ell = p. \end{cases}$$

Proof. We first suppose $\ell \neq p$. The multiplication-by-ℓ map $[\ell]$ is then separable, and we may apply Theorem 5.8 to compute $\#E[\ell] = \#\ker[\ell] = \deg[\ell] = \ell^2$. Every nonzero element of $E[\ell]$ has order ℓ, so we must have $E[\ell] \simeq \mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z}$. If $E[\ell^e] \simeq \langle P_1\rangle \oplus \cdots \oplus \langle P_r\rangle$ with each $P_i \in E(\bar{k})$ of order $\ell^{e_i} > 1$, then

$$E[\ell] \simeq \langle \ell^{e_1-1}P_1\rangle \oplus \cdots \oplus \langle \ell^{e_r-1}P_r\rangle \simeq (\mathbb{Z}/\ell\mathbb{Z})^r,$$

and we must have $r = 2$; more generally, for any abelian group G the ℓ-rank r of $G[\ell^e]$ is the same as the ℓ-rank of $G[\ell]$. It follows that $E[\ell^e] \simeq \mathbb{Z}/\ell^e\mathbb{Z} \oplus \mathbb{Z}/\ell^e\mathbb{Z}$, since we have $\#E[\ell^e] = \#\ker[\ell^e] = \deg[\ell^e] = \ell^{2e}$ and $E[\ell^e]$ contains no elements of order greater than ℓ^e.

We now suppose $\ell = p$. We have $\deg[\ell] = \deg_s[\ell]\deg_i[\ell] = \ell^2$ with $\deg_i[\ell] > 1$, so $\deg_s[\ell]$ is either ℓ or 1, which means that $E[\ell]$ must be isomorphic to $\mathbb{Z}/\ell\mathbb{Z}$ or $\{0\}$. In the latter case we clearly have $E[\ell^e] = \{0\}$ and the theorem holds, so we assume $E[\ell] \simeq \mathbb{Z}/\ell\mathbb{Z}$. Then $[\ell]$ is an ℓ-to-1 map on $E(\bar{k})$ and it follows that $\#E[\ell^e] = \ell \cdot \#E[\ell^{e-1}] = \cdots = \ell^{e-1}\#E[\ell] = \ell^e$ for all $e > 1$. The group $E[\ell^e]$ has the same ℓ-rank as $E[\ell]$, so $E[\ell^e] \simeq \mathbb{Z}/\ell^e\mathbb{Z}$. □

The two possibilities for $E[p]$ admitted by the theorem lead to the following definitions. We do not need this terminology today, but it will be important in the weeks that follow.

Definition 6.2. Let E be an elliptic curve defined over a field of characteristic $p > 0$. If $E[p] \simeq \mathbb{Z}/p\mathbb{Z}$ then E is said to be *ordinary*, and if $E[p] \simeq \{0\}$, we say that E is *supersingular*.

Remark 6.3. The term "supersingular" is unrelated to the term "singular" (recall that an elliptic curve is nonsingular by definition). Supersingular refers to the fact that such elliptic curves are exceptional.

Corollary 6.4. *Let E/k be an elliptic curve. Every finite subgroup of $E(\bar{k})$ can be written as the direct sum of two (possibly trivial) cyclic groups, at most one of which has order divisible by the characteristic of k. If $k = \mathbb{F}_q$ is a finite field of characteristic p we have*

$$E(\mathbb{F}_q) \simeq \mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$$

for some positive integers m, n with $m|n$ and $p \nmid m$.

Proof. Let T be a finite subgroup of $E(\bar{k})$. As a finite abelian group, T is the direct sum of its ℓ-Sylow subgroups T_ℓ, each of which is a subgroup of $E[\ell^e]$ for some e, hence a product of at most two cyclic groups by Theorem 6.1, and we can write $T_\ell \simeq T_{\ell,1} \oplus T_{\ell,2}$ with $T_{\ell,1}$ and $T_{\ell,2}$ groups of ℓ-power order, with $T_{\ell,2}$ is trivial if $\ell = p$. The groups $T_1 := \bigoplus_\ell T_{\ell,1}$ and $T_2 := \bigoplus_\ell T_{\ell,2}$ are cyclic, with $p \nmid \#T_2$, and $T \simeq T_1 \oplus T_2$. □

Now that we know what the structure of $E(\mathbb{F}_q)$ looks like, our next goal is to bound its cardinality. In the next lecture we will prove Hasse's Theorem, which states that

$$\#E(\mathbb{F}_q) = q + 1 - t,$$

where $|t| \leq 2\sqrt{q}$, but we first need to study the endomorphism ring of E.

6.2 Groups of homomorphisms

For any pair of elliptic curves E_1/k and E_2/k, the set $\mathrm{Hom}(E_1, E_2)$ of homomorphisms from E_1 to E_2 (defined over k) consists of all morphisms of curves $E_1 \to E_2$ that are also group homomorphisms $E_1(\bar{k}) \to E_2(\bar{k})$; since a morphism of curves is either surjective or constant, this is just the set of all isogenies from E_1 to E_2 plus the zero morphism. For any algebraic extension L/k, we write $\mathrm{Hom}_L(E_1, E_2)$ for the homomorphisms from E_1 to E_2 that are defined over L.[1]

The set $\mathrm{Hom}(E_1, E_2)$ forms an abelian group: for $\alpha, \beta \in \mathrm{Hom}(E_1, E_2)$ the sum $\alpha + \beta$ is defined pointwise via

$$(\alpha + \beta)(P) := \alpha(P) + \beta(P),$$

and the zero morphism from E_1 to E_2 is the identity element of $\mathrm{Hom}(E_1, E_2)$. Because addition is defined pointwise, if $\alpha(P) = \beta(P)$ for all $P \in E_1(\bar{k})$ then $\alpha = \beta$ because $\alpha - \beta$ is the zero morphism; we can thus test equality in $\mathrm{Hom}(E_1, E_2)$ pointwise.

Proposition 6.5. *Let E_1, E_2 be elliptic curves over a field k. For all $n \in \mathbb{Z}$ and all $\alpha \in \mathrm{Hom}(E_1, E_2)$ we have*

$$[n] \circ \alpha = n\alpha = \alpha \circ [n],$$

where the map $[n]$ on the LHS is multiplication-by-n on E_2 and the map $[n]$ on the RHS is multiplication-by-n on E_1.

Proof. For any $P \in E_1(\bar{k})$ and $\alpha \in \mathrm{Hom}(E_1, E_2)$ we have

$$([-1] \circ \alpha)(P) = -\alpha(P) = \alpha(-P) = (\alpha \circ [-1])(P),$$

since α is a group homomorphism, thus the proposition holds for $n = -1$ (as noted above, we can check equality of morphisms pointwise). All sides of the equalities are multiplicative in n, so it suffices to consider the case $n \geq 0$, where we have

$$([n] \circ \alpha)(P) = n\alpha(P) = \alpha(P) + \cdots + \alpha(P) = \alpha(P + \cdots P) = \alpha(nP) = (\alpha \circ [n])(P),$$

since α is a group homomorphism. The proposition follows. □

[1] Technically speaking, these homomorphisms are defined on the base changes E_{1_L} and E_{2_L} of E_1 and E_2 to L, so $\mathrm{Hom}_L(E_1, E_2)$ is really shorthand for $\mathrm{Hom}(E_{1_L}, E_{2_L})$.

Provided α and n are nonzero, both $[n]$ and α are surjective, as is $n\alpha$, thus $n\alpha \neq 0$; recall that by Theorem 4.17, every morphism of projective curves is either surjective or constant, and for elliptic curves (whose morphisms must preserve the distinguished point) the only constant morphism is the zero map. It follows that $\mathrm{Hom}(E_1, E_2)$ is a torsion free abelian group (but $\mathrm{Hom}(E_1, E_2) = \{0\}$ is possible).

Composition of homomorphisms distributes with addition: for any $\delta \in \mathrm{Hom}(E_0, E_1)$, $\alpha, \beta \in \mathrm{Hom}(E_1, E_2)$ and $\gamma \in \mathrm{Hom}(E_2, E_3)$ we have

$$(\alpha + \beta) \circ \gamma = \alpha \circ \gamma + \beta \circ \gamma \qquad \text{and} \qquad \delta \circ (\alpha + \beta) = \delta \circ \alpha + \delta \circ \beta,$$

since these identities hold pointwise (because $\alpha, \beta, \gamma, \delta$ are group homomorphisms).

Lemma 6.6. *Let $\delta\colon E_0 \to E_1$, $\alpha, \beta\colon E_1 \to E_2$, and $\gamma\colon E_2 \to E_3$ be isogenies. Then*

$$\begin{aligned} \alpha \circ \delta = \beta \circ \delta \quad &\implies \quad \alpha = \beta \\ \gamma \circ \alpha = \gamma \circ \beta \quad &\implies \quad \alpha = \beta. \end{aligned}$$

Proof. Isogenies are surjective, so in particular, γ, δ are not zero morphisms. We have

$$\begin{aligned} \alpha \circ \delta = \beta \circ \delta \Rightarrow \alpha \circ \delta - \beta\delta = 0 \Rightarrow (\alpha - \beta) \circ \delta = 0 \Rightarrow \alpha - \beta = 0 \Rightarrow \alpha = \beta \\ \gamma \circ \alpha = \gamma \circ \beta \Rightarrow \gamma \circ \alpha - \gamma \circ \beta = 0 \Rightarrow \gamma \circ (\alpha - \beta) = 0 \Rightarrow \alpha - \beta = 0 \Rightarrow \alpha = \beta. \end{aligned}$$

where the third arrow in both lines follows from the fact that a composition of morphisms is zero if and only if one of the morphisms in the composition is zero (because nonzero morphisms are surjective, as is their composition). □

6.3 The dual isogeny

To further develop our understanding of endomorphism rings (and isogenies in general) we now introduce the *dual isogeny*, whose existence is given by the following theorem. In the proof of the theorem we will appeal repeatedly to Theorem 5.11, which guarantees the existence of a separable isogeny with any given finite kernel, which is unique up to isomorphism. This implies that if $\alpha\colon E_1 \to E_2$ and $\alpha'\colon E_1 \to E_3$ are separable isogenies with the same kernel then there is an isomorphism $\iota\colon E_2 \to E_3$ such that $\alpha' = \iota \circ \alpha$. We will also make use of the fact that the kernel of an isogeny $\alpha\colon E_1 \to E_2$ of degree n is necessarily a subgroup of $E_1[n]$: by Theorem 5.8, $\#\ker\alpha = \deg_s \alpha$ is a divisor of $n = \deg\alpha$, so every $P \in \ker\alpha$ has order dividing n and is therefore an n-torsion point (satisfies $nP = 0$).

Theorem 6.7. *For any isogeny $\alpha\colon E_1 \to E_2$ of elliptic curves over a field k there exists a unique isogeny $\hat{\alpha}\colon E_2 \to E_1$ for which $\hat{\alpha} \circ \alpha = [n]$, where $n = \deg\alpha$.*

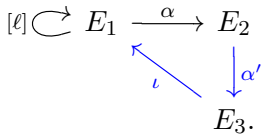
Proof. Uniqueness is immediate: if $\alpha_1 \circ \alpha = \alpha_2 \circ \alpha$ then $\alpha_1 = \alpha_2$ (by the cancellation law for composition of isogenies), so the equation $\hat{\alpha} \circ \alpha = [n]$ uniquely determines $\hat{\alpha}$.

To prove existence we proceed by induction on the number of prime factors of n, counted with multiplicity (recall from Corollary 5.12 that any isogeny can be written as a composition of isogenies of prime degree). Let p be the characteristic of the field k over which the elliptic curves E_1 and E_2 are defined.

If $n = 1$ has no prime factors then α is separable (otherwise we would have $p|\deg\alpha$) and has trivial kernel, and the same is true of the identity map $[1]$. It follows from Theorem 5.11 that there is an isomorphism $\iota\colon E_2 \to E_1$ such that $\iota \circ \alpha = [1]$, and we can take $\hat{\alpha} = \iota$.

We now suppose $n = \ell$ is prime. There are three cases to consider:

Case 1 ($\ell \neq p$): In this case α and $[\ell]$ are both separable and $\alpha(E_1[\ell])$ is a subgroup of $E_2(\bar{k})$ of cardinality $\deg[\ell]/\deg\alpha = \ell^2/\ell = \ell$. Let $\alpha'\colon E_2 \to E_3$ be the separable isogeny with $\alpha(E_1[\ell])$ as its kernel. The isogenies $\alpha' \circ \alpha$ and $[\ell]$ both have kernel $E_1[\ell]$, so there is an isomorphism $\iota\colon E_3 \to E_1$ for which $\iota \circ \alpha' \circ \alpha = [\ell]$, by Theorem 5.11, as shown below.

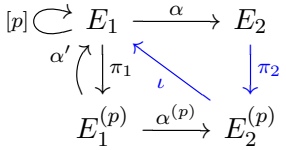


We now put $\hat{\alpha} := \iota \circ \alpha'$ to obtain $\hat{\alpha} \circ \alpha = [\ell]$ as desired.

Case 2 ($\ell = p$ and α separable): If α is separable then its kernel has order $\deg\alpha = p$ and we must have $\ker\alpha = E_1[p] \simeq \mathbb{Z}/p\mathbb{Z}$, by Theorem 6.1, and $\deg_s[p] = p$. Now $\deg[p] = p^2$, so by Corollary 5.4 we have $[p] = \alpha' \circ \pi_1$ for some separable isogeny $\alpha'\colon E_1^{(p)} \to E_1$ of degree p, where $\pi_1\colon E_1 \to E_1^{(p)}$ is the p-power Frobenius morphism.[2] We have $\pi_2 \circ \alpha = \alpha^{(p)} \circ \pi_1$, where $\alpha^{(p)}\colon E_1^{(p)} \to E_2^{(p)}$ is obtained by replacing each coefficient of α by its pth power, and

$$\ker(\alpha^{(p)} \circ \pi_1) = \ker(\pi_2 \circ \alpha) = \ker\alpha = \ker[p] = \ker(\alpha' \circ \pi_1),$$

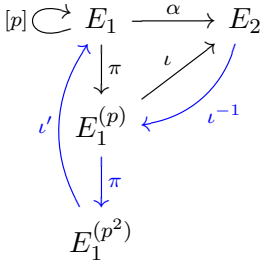
since the Frobenius morphisms π_1 and π_2 have trivial kernel, and it follows that $\alpha^{(p)}$ and α' are separable isogenies with the same kernel. There is thus an isomorphism $\iota\colon E_2^{(p)} \to E_1$ such that $\alpha' = \iota \circ \alpha^{(p)}$ (again by Theorem 5.11), as shown in the diagram below:

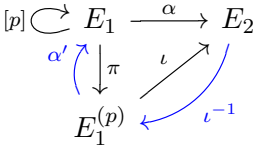


If we now put $\hat{\alpha} = \iota \circ \pi_2$ then

$$\hat{\alpha} \circ \alpha = \iota \circ \pi_2 \circ \alpha = \iota \circ \alpha^{(p)} \circ \pi_1 = \alpha' \circ \pi_1 = [p].$$

Case 3 ($\ell = p$ and α inseparable): In this case α must be purely inseparable, since its degree is prime, so $\alpha = \iota \circ \pi$ for some separable isogeny ι of degree $\deg_s\alpha = 1$, which must be an isomorphism. If $E[p] = \{0\}$ then $[p]$ is purely inseparable of degree p^2, so $[p] = \iota' \circ \pi^2$ for some isomorphism ι', and we may take $\hat{\alpha} = \iota' \circ \pi \circ \iota^{-1}$. If $E[p] \simeq \mathbb{Z}/p\mathbb{Z}$ then $[p] = \alpha' \circ \pi$ for some separable isogeny α' of degree p and we may take $\hat{\alpha} = \alpha' \circ \iota^{-1}$. The two cases are shown in the diagrams below.





[2] If $E_1\colon y^2 = x^3 + A_1x + B_1$ then $E_1^{(p)}$ denotes the elliptic curve $E_1^{(p)}\colon y^2 = x^3 + A_1^px + B_1^p$.

This completes the base case of our induction. If n is composite then we may decompose α into a sequence of isogenies of prime degree via Corollary 5.12. It follows that we can write $\alpha = \alpha_1 \circ \alpha_2$, where α_1, α_2 have degrees $n_1, n_2 < n$ with $n_1 n_2 = n$. Let $\hat{\alpha} = \hat{\alpha}_2 \circ \hat{\alpha}_1$, where the existence of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ is given by the inductive hypothesis. Then

$$\hat{\alpha} \circ \alpha = (\hat{\alpha}_2 \circ \hat{\alpha}_1) \circ \alpha = \hat{\alpha}_2 \circ \hat{\alpha}_1 \circ \alpha_1 \circ \alpha_2 = \hat{\alpha}_2 \circ [n_1] \circ \alpha_2 = \hat{\alpha}_2 \circ \alpha_2 \circ [n_1] = [n_2] \circ [n_1] = [n],$$

where $[n_1] \circ \alpha_2 = \alpha_2 \circ [n_1]$ by Proposition 6.5. □

Definition 6.8. The isogeny $\hat{\alpha}$ given by Theorem 6.7 is the *dual isogeny* of α.

Remark 6.9. One can define the dual isogeny for abelian varieties of any dimension, but in general if we have an isogeny of abelian varieties $\alpha\colon A_1 \to A_2$ then the dual isogeny

$$\hat{\alpha}\colon \hat{A}_2 \to \hat{A}_1,$$

is actually an isogeny between the *dual abelian varieties* $\hat{A}_2$ and $\hat{A}_1$. We won't give a definition of the dual abelian variety here, but the key point is that, in general, abelian varieties are not isomorphic to their duals. But abelian varieties of dimension one (elliptic curves) always are. This is yet another remarkable feature of elliptic curves.

As a matter of convenience we extend the notion of a dual isogeny to $\mathrm{Hom}(E_1, E_2)$ and $\mathrm{End}(E)$ by defining $\hat{0} = 0$, and we define $\deg 0 = 0$ so that $\hat{0} \circ 0 = [0]$ as in Theorem 6.7.

Lemma 6.10. *For an isogeny α of degree n we have $\alpha \circ \hat{\alpha} = [n]$, meaning that $\hat{\hat{\alpha}} = \alpha$. For any $n \in \mathbb{Z}$ the endomorphism $[n]$ is self-dual, that is, $\widehat{[n]} = [n]$.*

Proof. We have

$$(\alpha \circ \hat{\alpha}) \circ \alpha = \alpha \circ (\hat{\alpha} \circ \alpha) = \alpha \circ [n] = [n] \circ \alpha,$$

Isogenies are nonzero, so we may cancel α on the right to obtain $\alpha \circ \hat{\alpha} = [n]$. The last statement follows from the fact that $[n] \circ [n] = [n^2] = [\deg[n]]$. □

Lemma 6.11. *For any $\alpha, \beta \in \mathrm{Hom}(E_1, E_2)$ we have $\widehat{\alpha + \beta} = \hat{\alpha} + \hat{\beta}$.*

Proof. We will defer the proof of this lemma — the nicest proof uses the Weil pairing, which we will see later in the course. □

Lemma 6.12. *For any $\alpha \in \mathrm{Hom}(E_2, E_3)$ and $\beta \in \mathrm{Hom}(E_1, E_2)$ we have $\widehat{\alpha \circ \beta} = \hat{\beta} \circ \hat{\alpha}$.*

Proof. Let $m := \deg \alpha$ and $n := \deg \beta$. Then $\deg(\alpha \circ \beta) = mn$, by Corollary 5.10, and

$$(\hat{\beta} \circ \hat{\alpha}) \circ (\alpha \circ \beta) = \hat{\beta} \circ [m] \circ \beta = [m] \circ \hat{\beta} \circ \beta = [m] \circ [n] = [mn] = [\deg(\alpha \circ \beta)].$$

The lemma then follows from the definition of $\widehat{\alpha \circ \beta}$. □

6.4 Endomorphism rings

Definition 6.13. Let E/k be an elliptic curve. The endomorphism ring of E is the additive group $\mathrm{End}(E) := \mathrm{Hom}(E, E)$ with multiplication given by composition: $\alpha\beta := \alpha \circ \beta$.

Warning 6.14. Many authors use $\mathrm{End}(E)$ to mean $\mathrm{End}(E_{\bar{k}})$.

To verify that $\mathrm{End}(E)$ is in fact a ring, note that it has a multiplicative identity $1 = [1]$ (the identity morphism), and for all $\alpha, \beta, \gamma \in \mathrm{End}(E)$ and $P \in E(\bar{k})$ we have

$$((\alpha+\beta)\gamma)(P) = (\alpha+\beta)(\gamma(P)) = \alpha(\gamma(P)) + \beta(\gamma(P)) = (\alpha\gamma + \beta\gamma)(P)$$
$$(\gamma(\alpha+\beta))(P) = \gamma(\alpha(P) + \beta(P)) = \gamma(\alpha(P)) + \gamma(\beta(P)) = (\gamma\alpha + \gamma\beta)(P),$$

where we used the fact that γ is a group homomorphism to get the second identity.

For every integer n the multiplication-by-n map $[n]$ lies in $\mathrm{End}(E)$, and the map $n \mapsto [n]$ defines a ring homomorphism $\mathbb{Z} \to \mathrm{End}(E)$, since $[0] = 0, [1] = 1$, $[m] + [n] = [m+n]$ and $[m][n] = [mn]$. As noted above, $\mathrm{Hom}(E, E)$ is torsion free, so the homomorphism $n \mapsto [n]$ is injective and may regard $\mathbb{Z}$ as a subring of $\mathrm{End}(E)$; we will thus feel free to write n rather than $[n]$ when it is convenient to do so. Proposition 6.5 implies that $\mathbb{Z}$ lies in the center of $\mathrm{End}(E)$, since $n\alpha = \alpha n$ for all $\alpha \in \mathrm{End}(E)$. As we shall see, the ring $\mathrm{End}(E)$ need not be commutative, in general, which makes the elements that lie in its center of interest.

When $k = \mathbb{F}_q$ is a finite field, the q-power Frobenius endomorphism π_E also lies in the center of $\mathrm{End}(E)$. This follows from the fact that for any rational function $r \in \mathbb{F}_q(x_1, \ldots, x_n)$ we have $r(x_1, \ldots, x_n)^q = r(x_1^q, \ldots, x_n^q)$, and we can apply this to the rational maps defining any $\alpha \in \mathrm{End}(E)$. Thus the subring $\mathbb{Z}[\pi_E]$ generated by π_E lies in the center of $\mathrm{End}(E)$.

Remark 6.15. It can happen that $\mathbb{Z}[\pi_E] = \mathbb{Z}$. For example, when $E[p] = \{0\}$ and $q = p^2$ the multiplication-by-p map $[p]$ is purely inseparable and $[p]$ is necessarily the composition of $\pi^2 = \pi_E$ with an isomorphism. This isomorphism is typically $[\pm 1]$, in which case $\pi_E \in \mathbb{Z}$.

For any nonzero $\alpha, \beta \in \mathrm{End}(E)$, the product $\alpha\beta = \alpha \circ \beta$ is surjective, since α and β are both surjective; in particular, $\alpha\beta$ is not the zero morphism. It follows that $\mathrm{End}(E)$ has no zero divisors, so the cancellation law holds (on both the left and the right).

We now return to the setting of the endomorphism ring $\mathrm{End}(E)$ of an elliptic curve E/k.

Lemma 6.16. *For any endomorphism α we have $\alpha + \hat{\alpha} = 1 + \deg\alpha - \deg(1-\alpha)$.*

Note that in the statement of this lemma, $1 - \alpha$ denotes the endomorphism $[1] - \alpha$ and the integers $\deg\alpha$ and $\deg(1-\alpha)$ are viewed as elements of $\mathrm{End}(E)$ via the embedding $\mathbb{Z} \hookrightarrow \mathrm{End}(E)$ defined by $n \mapsto [n]$.

Proof. For any $\alpha \in \mathrm{End}(E)$ (including $\alpha = 0$) we have

$$\deg(1-\alpha) = \widehat{(1-\alpha)}(1-\alpha) = (\hat{1} - \hat{\alpha})(1-\alpha) = (1-\hat{\alpha})(1-\alpha) = 1 - (\alpha + \hat{\alpha}) + \deg(\alpha),$$

and therefore $\alpha + \hat{\alpha} = 1 + \deg\alpha - \deg(1-\alpha)$. □

A key consequence of the lemma is that $\alpha + \hat{\alpha}$ is always a multiplication-by-t map for some integer $t \in \mathbb{Z}$.

Definition 6.17. The *trace* of an endomorphism α is the integer $\mathrm{tr}\,\alpha := \alpha + \hat{\alpha}$.

Note that for any $\alpha \in \mathrm{End}(E)$ we have $\mathrm{tr}\,\hat{\alpha} = \mathrm{tr}\,\alpha$, and $\deg\hat{\alpha} = \deg\alpha$. This implies that α and $\hat{\alpha}$ have the same characteristic polynomial.

Theorem 6.18. *Let α be an endomorphism of an elliptic curve. Both α and its dual $\hat{\alpha}$ are solutions to*

$$\lambda^2 - (\mathrm{tr}\,\alpha)\lambda + \deg\alpha = 0.$$

Proof. $\alpha^2 - (\mathrm{tr}\,\alpha)\alpha + \deg\alpha = \alpha^2 - (\alpha+\hat{\alpha})\alpha + \hat{\alpha}\alpha = 0$, and similarly for $\hat{\alpha}$. □

6.5 Endomorphism restrictions to $E[n]$

Let E/k be an elliptic curve with $\mathrm{char}(k) = p$ (possibly $p = 0$). For any $\alpha \in \mathrm{End}(E)$, we may consider the restriction α_n of α to the n-torsion subgroup $E[n]$. Since α is a group homomorphism, it maps n-torsion points to n-torsion points, so α_n is an endomorphism of the abelian group $E[n]$.

Provided n is not divisible by p, we have $E[n] \simeq \mathbb{Z}/n\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$ with rank 2, and we can pick a basis $\langle P_1, P_2 \rangle$ for $E[n]$ as a $(\mathbb{Z}/n\mathbb{Z})$-module, so that every element of $E[n]$ can be written uniquely as a $(\mathbb{Z}/n\mathbb{Z})$-linear combination of P_1 and P_2 — it suffices to pick any $P_1, P_2 \in E[n]$ that generate $E[n]$ as an abelian group. Having fixed a basis for $E[n]$, we may represent α_n as a 2×2 matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, where $a, b, c, d \in \mathbb{Z}/n\mathbb{Z}$ are determined by

$$\alpha(P_1) = aP_1 + bP_2,$$
$$\alpha(P_2) = cP_1 + dP_2.$$

This matrix representation depends on our choice of basis but its conjugacy class does not; in particular the trace $\mathrm{tr}\,\alpha_n$ and determinant $\det \alpha_n$ are independent of our choice of basis.

A standard technique for proving that two endomorphisms α and β are equal is to prove that $\alpha_n = \beta_n$ for some sufficiently large n. If n^2 is larger than the degree of $\alpha - \beta$, then $\alpha_n = \beta_n$ implies $\ker(\alpha - \beta) > \deg(\alpha - \beta)$, which is impossible unless $\alpha - \beta = 0$, in which case $\alpha = \beta$. To handle situations where we don't know the degree of $\alpha - \beta$, or don't even know exactly what β is (maybe we just know β_n), we need a more refined result.

Lemma 6.19. *Let α and β be endomorphisms of an elliptic curve E/k and let m be the maximum of $\deg\alpha$ and $\deg\beta$. Let $n \geq 2\sqrt{m}+1$ be an integer prime to the characteristic of k, and also relatively prime to the integers $\deg\alpha$ and $\deg\beta$. If $\alpha_n = \beta_n$ then $\alpha = \beta$.*

Proof. We shall make use of the following fact. Let $r(x) = u(x)/v(x)$ be a rational function in $k(x)$ with $u \perp v$ and v monic. Suppose that we know the value of $r(x_i)$ for N distinct values $x_1, \ldots, x_N$ for which $v(x_i) \neq 0$. Provided that $N > 2\max\{\deg u, \deg v\} + 1$, the polynomials $u, v \in [x]$ can be uniquely determined using *Cauchy interpolation*; see [1, §5.8] for an efficient algorithm and a proof of its correctness. In particular, two rational functions with degrees bounded by N as above that agree on N distinct points must coincide.

Now let $\alpha(x, y) = \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$ be in standard form, with $u \perp v$, and v monic. If we know the value of $\alpha(P)$ at $2\deg\alpha + 2$ affine points $P \notin \ker\alpha$ with distinct x-coordinates, then we can uniquely determine u and v. For each $x_0 \in \bar{k}$ at most 2 points $P \in E(\bar{k})$ have x-coordinate x_0, so it suffices to know $\alpha(P)$ at $4\deg\alpha + 4$ affine points not in $\ker\alpha$.

For $n \geq 2\sqrt{m}+1$ we have $n^2 \geq 4m + 4\sqrt{m} + 1$, and $E[n]$ contains $n^2 - 1 \geq 4\deg\alpha + 4$ affine points, none of which lie in $\ker\alpha$, since $\#\ker\alpha$ divides $\deg\alpha$ which is coprime to n. Thus α_n uniquely determines the x-coordinate of $\alpha(P)$ for all $P \in E(\bar{k})$. The same argument applies to β_n and β, hence $\alpha(P) = \pm\beta(P)$ for all $P \in E(\bar{k})$. The kernel of at least one of $\alpha + \beta$ and $\alpha - \beta$ is therefore infinite, and it follows that $\alpha = \pm\beta$.

We have $n^2 > 4\deg\alpha \geq 4$, which implies that $\alpha(P)$ cannot lie in $E[2]$ for all $P \in E[n]$ (since $\#E[2] = 4$). Therefore $\alpha(P) \neq -\alpha(P)$ for some $P \in E[n]$, and for this P we have $\alpha(P) \neq -\alpha(P)$ and $\alpha_n(P) \neq -\alpha_n(P) = -\beta_n(P)$, so $\alpha \neq -\beta$ and we must have $\alpha = \beta$. □

The following theorem provides the key connection between endomorphisms and their restrictions to $E[n]$.

Theorem 6.20. *Let α be an endomorphism of an elliptic curve E/k and let n be a positive integer prime to the characteristic of k. Then*

$$\operatorname{tr}\alpha \equiv \operatorname{tr}\alpha_n \bmod n \qquad \textit{and} \qquad \deg\alpha \equiv \det\alpha_n \bmod n.$$

Proof. We will just prove the theorem for odd n prime to $\deg\alpha$ such that $n \geq 2\sqrt{\deg\alpha}+1$. The general proof relies on properties of the Weil pairing that we will see later in the course.

We note that the theorem holds for $\alpha = 0$, so we assume $\alpha \neq 0$. Let n be as above and let $t_n = \operatorname{tr}\alpha \bmod n$ and $d_n = \deg\alpha \bmod n$. Since α and $\hat{\alpha}$ both satisfy $\lambda^2 - (\operatorname{tr}\alpha)\lambda + \deg\alpha = 0$, both α_n and $\hat{\alpha}_n$ must satisfy $\lambda^2 - t_n\lambda + d_n = 0$. It follows that $\alpha_n + \hat{\alpha}_n$ and $\alpha_n\hat{\alpha}_n$ are the scalar matrices t_nI and d_nI, respectively. Let $\alpha_n = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, and let $\delta_n = \det\alpha_n$. The fact that $\hat{\alpha}_n\alpha_n = d_nI \neq 0$ with d_n prime to n implies that α_n is invertible, and we have

$$\hat{\alpha}_n = d_n\alpha_n^{-1} = \frac{d_n}{\det\alpha_n}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

If we put $\epsilon := d_n/\det\alpha_n$ and plug the expression for $\hat{\alpha}_n$ into $\alpha_n + \hat{\alpha}_n = t_nI$ we get

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \epsilon\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} t_n & 0 \\ 0 & t_n \end{bmatrix}.$$

Thus $a + \epsilon d = t_n$, $b - \epsilon b = 0$, $c - \epsilon c = 0$, and $d + \epsilon a = t_n$. Unless $a = d$ and $b = c = 0$, we must have $\epsilon = 1$, in which case $d_n = \det\alpha_n$ and $t_n = a + d = \operatorname{tr}\alpha_n$ as desired.

If $a = d$ and $b = c = 0$ then α_n is a scalar matrix. Let m be the unique integer with absolute value less than $n/2$ such that $\alpha_n = m_n$, where m_n is the restriction of the multiplication-by-m map to $E[n]$. We then have $\deg m = m^2$ and $n \geq 2\sqrt{\deg m}+1$. Since we also have $n \geq 2\sqrt{\deg\alpha}+1$ we must have $\alpha = m$, by Lemma 6.19. But then $\hat{\alpha} = \hat{m} = m = \alpha$, so $\operatorname{tr}\alpha = 2m \equiv \operatorname{tr} mI \equiv \operatorname{tr}\alpha_n \bmod n$ and $\deg\alpha = m^2 \equiv \det mI \equiv \det\alpha_n \bmod n$. □

References

[1] Joachim von zur Gathen and Jürgen Gerhard, *Modern Computer Algebra*, third edition, Cambridge University Press, 2013.

7 Point counting

7.1 Separable and inseparable endomorphisms

Recall that the Frobenius endomorphism π_E is inseparable. In order to prove Hasse's theorem we will need to use the fact that $\pi_E - 1$ is separable. This follows from a much more general result: adding a separable isogeny to an inseparable isogeny always yields a separable isogeny. Note that the sum of two separable isogenies need not be separable: in characteristic $p > 0$, if we have $a + b = p$ and both a and b prime to p, then $[a]$ and $[b]$ are both separable but $[a] + [b] = [a + b] = [p]$ is inseparable.

Lemma 7.1. *Let α and β be isogenies from E_1 to E_2, with α inseparable. Then $\alpha + \beta$ is inseparable if and only if β is inseparable.*

Proof. If β is inseparable then by Corollary 5.4 we can write $\alpha = \alpha_{\text{sep}} \circ \pi^m$ and $\beta = \beta_{\text{sep}} \circ \pi^n$, where π is the p-power Frobenius map and $m, n > 0$. We then have

$$\alpha + \beta = \alpha_{\text{sep}} \circ \pi^m + \beta_{\text{sep}} \circ \pi^n = (\alpha_{\text{sep}} \circ \pi^{m-1} + \beta_{\text{sep}} \circ \pi^{n-1}) \circ \pi,$$

which is inseparable (any composition involving an inseparable isogeny is inseparable because inseparable degrees multiply). If $\alpha+\beta$ is inseparable, then so is $-(\alpha+\beta)$, and $\alpha-(\alpha+\beta) = \beta$ is a sum of inseparable isogenies, which we have just shown is inseparable. □

Remark 7.2. Since the composition of an inseparable isogeny with any isogeny is always inseparable, Lemma 7.1 implies that the inseparable endomorphisms in $\text{End}(E)$ form an ideal (provided we view 0 as inseparable, which we do).

7.2 Hasse's Theorem

We are now ready to prove Hasse's theorem.

Theorem 7.3 (Hasse)**.** *Let $E/\mathbb{F}_q$ be an elliptic curve over a finite field. Then*

$$\#E(\mathbb{F}_q) = q + 1 - t,$$

where $t := \operatorname{tr} \pi_E$ is the trace of the Frobenius endomorphism π_E and $|t| \le 2\sqrt{q}$.

Proof. Recall that we defined $\mathbb{F}_q$ as the splitting field of $x^q - x$ over $\mathbb{F}_p$, where $p = \text{char}(\mathbb{F}_q)$, thus $\mathbb{F}_q = \{\alpha \in \overline{\mathbb{F}}_p : \alpha^q - \alpha = 0\} = \{\alpha \in \overline{\mathbb{F}}_q : \alpha^q - \alpha = 0\}$ is precisely the subfield of $\overline{\mathbb{F}}_q$ fixed by the q-power Frobenius automorphism $x \mapsto x^q$. The Frobenius endomorphism $\pi_E \colon E \to E$ is defined by $\pi_E(x : y : z) = (x^q : y^q : z^q)$, therefore

$$E(\mathbb{F}_q) = \{P \in E(\overline{\mathbb{F}}_q) : \pi_E(P) = P\} = \{P \in E(\overline{\mathbb{F}}_q) : \pi_E(P) - P = 0\} = \ker(\pi_E - 1),$$

where 1 denotes the multiplication-by-1 map $[1] \in \text{End}(E)$. The Frobenius endomorphism π_E is inseparable and -1 is separable, so by Lemma 7.1 the endomorphism $\pi_E - 1$ is separable, thus the cardinality of its kernel is equal to its degree (by Theorem 5.8). Therefore

$$\#E(\mathbb{F}_q) = \#\ker(\pi_E - 1) = \deg(\pi_E - 1) = \widehat{(\pi_E - 1)}(\pi_E - 1) = \hat{\pi}_E\pi_E + 1 - (\hat{\pi}_E + \pi_E) = q + 1 - t.$$

It remains only to show that $|t| \le 2\sqrt{q}$.

Consider the endomorphism $r\pi_E - s$ for $r, s \in \mathbb{Z}$ with $s \neq 0$. We have

$$\begin{aligned}\deg(r\pi_E - s) &= \widehat{(r\pi_E - s)}(r\pi_E - s) = (\hat{\pi}_E\hat{r} - \hat{s})(r\pi_E - s) = (\hat{\pi}_E r - s)(r\pi_E - s)\\ &= \hat{\pi}_E r^2\pi_E - \hat{\pi}_E rs - sr\pi_E + s^2 = r^2\hat{\pi}_E\pi_E - rs(\hat{\pi}_E + \pi_E) + s^2\\ &= r^2 \deg \pi_E - rs \operatorname{tr} \pi_E + s^2\\ &= r^2 q - rst + s^2,\end{aligned}$$

where we have used Lemmas 6.11 and 6.12, and the fact that $\mathbb{Z}$ is in the center of $\mathrm{End}(E)$. Dividing by s^2 and noting that $\deg(r\pi_E - s) \geq 0$ yields the inequality

$$q\,(r/s)^2 - t\,(r/s) + 1 \geq 0,$$

valid for all rational numbers r/s. Now $\mathbb{Q}$ is dense in $\mathbb{R}$, so we must have $qx^2 - tx + 1 \geq 0$ for all real numbers x. It follows that the discriminant $t^2 - 4q$ cannot be positive, which yields the desired bound $|t| \leq 2\sqrt{q}$. □

Recall that for an odd prime p the Legendre symbol $\left(\frac{a}{p}\right)$ is defined by

$$\left(\frac{a}{p}\right) = \left\{\begin{array}{ll} 1 & \text{if } y^2 = a \text{ has two solutions mod } p\\ 0 & \text{if } y^2 = a \text{ has one solution mod } p\\ -1 & \text{if } y^2 = a \text{ has no solutions mod } p\end{array}\right\} = \#\{\alpha \in \mathbb{F}_p : \alpha^2 = a\} - 1.$$

We extend the Legendre symbol to all finite fields $\mathbb{F}_q$ of odd characteristic by defining

$$\left(\frac{a}{\mathbb{F}_q}\right) = \#\{\alpha \in \mathbb{F}_q : \alpha^2 = a\} - 1 \in \{-1, 0, 1\}.$$

Thus $1 + \left(\frac{a}{\mathbb{F}_q}\right)$ counts the solutions to $y^2 = a$ in $\mathbb{F}_q$. It follows that if $E/\mathbb{F}_q$ is given by the Weierstrass equation $y^2 = x^3 + Ax + B$, then

$$\begin{aligned}\#E(\mathbb{F}_q) &= 1 + \sum_{x_0 \in \mathbb{F}_q}\left(1 + \left(\frac{x_0^3 + Ax_0 + B}{\mathbb{F}_q}\right)\right)\\ &= q + 1 + \sum_{x_0 \in \mathbb{F}_q}\left(\frac{x_0^3 + Ax_0 + B}{\mathbb{F}_q}\right). \qquad (1)\end{aligned}$$

Hasse's Theorem is equivalent to the statement that the sum in (1) has absolute value at most $2\sqrt{q}$. This is remarkable for a sum with q terms, almost all of which are ± 1. From a probabilistic point of view, one might expect that *on average* an $O(\sqrt{q})$ bound should hold, but Hasse's theorem guarantees that it *always* holds.

The bound in Hasse's theorem is the best possible. Later in the course we will see how to explicitly construct elliptic curves $E/\mathbb{F}_q$ with cardinalities matching every integer value in the *Hasse interval*

$$\mathcal{H}(q) := [q + 1 - 2\sqrt{q},\ q + 1 + 2\sqrt{q}] = [(\sqrt{q} - 1)^2, (\sqrt{q} + 1)^2]$$

when q is prime, and all but at most two integers when q is not prime.

7.3 Point counting

We now consider the problem of computing the cardinality of $E(\mathbb{F}_q)$, which is crucial to cryptographic applications; as we shall see, it is quite important to know the cardinality of the group one is working in. The most naïve approach one might take would be to evaluate the curve equation $y^2 = x^3 + Ax + B$ for E at every pair $(x_0, y_0) \in \mathbb{F}_q^2$, count the number of solutions, and add 1 for the point at infinity. This takes $O(q^2\,\mathsf{M}(\log q))$ time. Note that the input to this problem is the pair of coefficients $A, B \in \mathbb{F}_q$, which each have $O(n)$ bits, where $n = \log q$. Thus in terms of the size of its input, this algorithm takes

$$O\big(\exp(2n)\,\mathsf{M}(n)\big)$$

time, which is obviously exponential in n.

A slightly less naïve approach is to precompute a table of quadratic residues in $\mathbb{F}_q$ so that we can very quickly compute the extended Legendre symbol $\left(\frac{\cdot}{\mathbb{F}_q}\right)$. We can construct such a table in $O(q\,\mathsf{M}(\log q))$ time, and then compute

$$\#E(\mathbb{F}_q) = q + 1 + \sum_{x \in \mathbb{F}_q} \left(\frac{x^3 + Ax + B}{\mathbb{F}_q}\right)$$

in $O(q\,\mathsf{M}(\log q))$ time, yielding a total running time of

$$O(\exp(n)\,\mathsf{M}(n)).$$

So far we have not taken advantage of Hasse's theorem which gives us an interval $\mathcal{H}(q)$ of width $4\sqrt{q}$ which we know must contain the integer $\#E(\mathbb{F}_q)$ we wish to determine.

7.4 Computing the order of a point

Before giving an algorithm to compute $\#E(\mathbb{F}_q)$ using Hasse's theorem, let us first consider an easier problem: computing the order $|P|$ of a single point $P \in E(\mathbb{F}_q)$. Since the order of the group $E(\mathbb{F}_q)$ lies in $\mathcal{H}(q)$, we know that $\mathcal{H}(q)$ contains at least one integer M for which $MP = 0$, namely $M = \#E(\mathbb{F}_q)$, and any such M is a multiple of $|P|$. To find such an M, we set $M_0 = \lceil(\sqrt{q} - 1)^2\rceil$, compute M_0P using double-and-add scalar multiplication, and then generate the sequence of points

$$M_0P,\ (M_0 + 1)P,\ (M_0 + 2)P,\ \ldots,\ MP = 0,$$

by adding P repeatedly. Note that M is bounded by $M_0 + 4\sqrt{q}$, so $4\sqrt{q}$ additions suffice.

We then compute the prime factorization $M = p_1^{e_1} \cdots p_w^{e_w}$ (easy, compared to the time to find M, we could even use trial division). To compute the exact order of the point P we use the following generic algorithm.

Algorithm 7.4. Given an element P of an additive group and the prime factorization $M = p_1^{e_1} \cdots p_r^{e_r}$ of an integer M for which $MP = 0$, compute the order of P as follows:

1. Let $m = M = p_1^{e_1} \cdots p_r^{e_r}$.
2. For each prime p_i, while $p_i | m$ and $(m/p_i)P = 0$, replace m by m/p_i.
3. Output m.

When this procedure is complete we know that $mP = 0$ and $(m/p)P \neq 0$ for every prime p dividing m; this implies that $m = |P|$. You will analyze the efficiency of this algorithm and develop several improvements to it in Problem Set 4, but the number of group operations is clearly polynomial in $\log M$, which is all we need for the moment.

The time to compute $|P|$ is thus dominated by the time to find a multiple of $|P|$ in $\mathcal{H}(q)$. This involves $O(\sqrt{q})$ operations in $E(\mathbb{F}_q)$, yielding a bit complexity of $O(\sqrt{q}\ \mathsf{M}(\log q))$ or

$$O\big(\exp(n/2)\,\mathsf{M}(n)\big),$$

assuming that we use projective coordinates to avoid field inversions when adding points.

We will shortly see how this can be further improved, but first let us consider how to use our algorithm for computing $|P|$ to compute $\#E(\mathbb{F}_q)$. If we are lucky (and if q is large we almost always will be), the first multiple M of $|P|$ that we find in $\mathcal{H}(q)$ will actually be the only multiple of $|P|$ in $\mathcal{H}(q)$. If this happens, then we must have $M = \#E(\mathbb{F}_q)$. Otherwise, we might try our luck with a different point P. If we can find a combination of points for which the least common multiple of their orders has a unique multiple in $\mathcal{H}(q)$, then we can determine the group order. Unfortunately this will not always be possible, but before addressing that issue, let us consider the question of how long it might take to compute the least common multiple of the orders of *all* the points in $E(\mathbb{F}_q)$, which is a lot less than one might expect.

7.5 The group exponent

Definition 7.5. For a finite group G, the *exponent* of G, denoted $\lambda(G)$, is defined by

$$\lambda(G) = \operatorname{lcm}\{|\alpha| : \alpha \in G\}.$$

Note that $\lambda(G)$ is a divisor of $\#G$ and is divisible by the order of every element of G. Thus $\lambda(G)$ is the maximal possible order of an element of G, and when G is abelian this maximum is achieved: there exists an element $\alpha \in G$ with order $|\alpha| = \lambda(G)$. To see this, note that the structure theorem for finite abelian groups allows us to decompose G as

$$G \simeq \mathbb{Z}/n_1\mathbb{Z} \oplus \mathbb{Z}/n_2\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/n_r\mathbb{Z},$$

with $n_i | n_{i+1}$ for $1 \leq i < r$. Thus $\lambda(G) = n_r$, and any $\alpha = (\alpha_1, \ldots, \alpha_r) \in G$ for which α_r is a generator for $\mathbb{Z}/n_r\mathbb{Z}$ will necessarily satisfy $|\alpha| = \lambda(G)$.

Rather than searching for a single α with maximal order, it is enough to find any set of elements $S \subseteq G$ for which $\operatorname{lcm}\{|\alpha| : \alpha \in S\} = \lambda(G)$. If we choose S randomly, how large does it need to be to have a good chance of determining $\lambda(G)$? The answer is surprisingly small: for $|S| = 2$ we already have a better than $50/50$ chance.

Theorem 7.6. *Let G be a finite abelian group with exponent $\lambda(G)$. Let α and β be uniformly distributed random elements of G. Then*

$$\Pr[\operatorname{lcm}(|\alpha|, |\beta|) = \lambda(G)] > \frac{6}{\pi^2}.$$

Proof. We first reduce to the case that G is cyclic. As noted above, $G \simeq \mathbb{Z}/n_1\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}/n_r\mathbb{Z}$ with $n_i | n_{i+1}$ and $\lambda(G) = n_r$. Let α_r and β_r be the projections of α and β to $\mathbb{Z}/n_r\mathbb{Z}$. Then $\operatorname{lcm}(|\alpha_r|, |\beta_r|) = \lambda(G)$ certainly implies $\operatorname{lcm}(|\alpha|, |\beta|) = \lambda(G)$, thus

$$\Pr[\operatorname{lcm}(|\alpha|, |\beta|) = \lambda(G)] \geq \Pr[\operatorname{lcm}(|\alpha_r|, |\beta_r|) = \lambda(G)],$$

and in the worst case G is cyclic and this inequality is an equality, which we now assume.

So let $G = \langle\gamma\rangle$ and let $p_1^{e_1} \cdots p_k^{e_k}$ be the prime factorization of $|\gamma| = \lambda(G) = \#G$. Then $\alpha = a\gamma$, with $0 \le a < |\gamma|$, and unless a is divisible by p_i, which occurs with probability $1/p_i$, the order of α will be divisible by $p_i^{e_i}$ (and similarly for β). The two probabilities for α and β are independent, thus with probability $1 - 1/p_i^2$ at least one of α and β has order divisible by $p_i^{e_i}$. Call this event E_i. The events $E_1, \ldots, E_k$ are independent, since we may write G as a direct sum of cyclic groups of prime-power orders $p_1^{e_1}, \ldots, p_k^{e_k}$, and the projections of α and β to each of these cyclic groups are uniformly and independently distributed. Thus

$$\begin{aligned}\Pr[\operatorname{lcm}(|\alpha|, |\beta|) = \lambda(G)] &= \Pr[E_1 \cap \cdots \cap E_k] \\ &= \prod_{p|\lambda(G)} (1 - p^{-2}) > \prod_p (1 - p^{-2}) = \left(\sum_{n=1}^{\infty} \frac{1}{n^2}\right)^{-1} = \frac{1}{\zeta(2)} = \frac{6}{\pi^2},\end{aligned}$$

where $\zeta(s) = \sum n^{-s}$ is the Riemann zeta function. □

Theorem 7.6 implies that if we generate random points $P \in E(\mathbb{F}_q)$ and accumulate the least common multiple N of their orders, we should expect to obtain $\lambda(E(\mathbb{F}_q))$ within $O(1)$ iterations. Regardless of when we obtain $\lambda(E(\mathbb{F}_q))$, at every stage we know that N divides $\#E(\mathbb{F}_q)$, and if we ever find that N has a unique multiple M in the Hasse interval $\mathcal{H}(q)$, then we know that $\#E(\mathbb{F}_q) = M$.

Unfortunately this might not ever happen; it can happen that $\lambda(E(\mathbb{F}_q)) \le 4\sqrt{q}$, in which case it is possible for $\lambda(E(\mathbb{F}_q))$ to have more than one multiple in $\mathcal{H}(q)$. To deal with this problem we need to consider the *quadratic twist* of E, which you saw on Problem Set 1.

7.6 The quadratic twist of an elliptic curve

Suppose s is an element of $\mathbb{F}_q$ that is *not* a square, meaning that $\left(\frac{s}{\mathbb{F}_q}\right) = -1$. If we consider the elliptic curve $\widetilde{E}$ defined by $sy^2 = x^3 + Ax + B$, then the affine point (x, y) will lie on the curve if and only if $x^3 + Ax + B$ is *not* a square. Thus

$$\#\widetilde{E}(\mathbb{F}_q) = q + 1 - \sum_{x \in \mathbb{F}_q} \left(\frac{x^3 + Ax + B}{\mathbb{F}_q}\right),$$

and it follows that if $\#E(\mathbb{F}_q) = q+1-t$, then $\#\widetilde{E}(\mathbb{F}_q) = q+1+t$. The curve $\widetilde{E}$ is called the *quadratic twist* of E (by s). We can put the curve equation for $\widetilde{E}$ in standard Weierstrass form by substituting x/s for x and y/s^2 for y and then clearing denominators, yielding

$$y^2 = x^3 + s^2Ax + s^3B.$$

Notice that it does not matter which non-residue s we choose. As you showed in Problem Set 1, if s and s' are any two non-squares in $\mathbb{F}_q$, then the corresponding curves $\widetilde{E}$ and $\widetilde{E'}$ are isomorphic over $\mathbb{F}_q$; thus we refer to $\widetilde{E}$ as "the" quadratic twist of E.[1]

Our interest in the quadratic twist of E lies in the fact that

$$\#E(\mathbb{F}_q) + \#\widetilde{E}(\mathbb{F}_q) = 2q + 2.$$

Thus if we can compute either $\#E(\mathbb{F}_q)$ or $\#\widetilde{E}(\mathbb{F}_q)$, we can easily determine both values.

[1]This situation is specific to finite fields. Over $\mathbb{Q}$, for example, every elliptic curve has infinitely many quadratic twists that are not isomorphic over $\mathbb{Q}$ (of course they are all isomorphic over $\overline{\mathbb{Q}}$).

7.7 Mestre's Theorem

As noted above, it is not necessarily the case that the exponent of $E(\mathbb{F}_p)$ has a unique multiple in the Hasse interval. But if we also consider the quadratic twist $\widetilde{E}(\mathbb{F}_p)$, then a theorem of Mestre (published by Schoof in [4]) ensures that for all primes $p > 229$, either $\lambda(E(\mathbb{F}_p))$ or $\lambda(\widetilde{E}(\mathbb{F}_p))$ has a unique multiple in the Hasse interval $\mathcal{H}(p)$. A generalization of this theorem that works for arbitrary prime powers q can be found in [2], but we will restrict ourselves to the case of primes $p > 229$ for the sake of simplicity.

Theorem 7.7 (Mestre)**.** *Let $p > 229$ be prime, and let $E/\mathbb{F}_p$ be an elliptic curve with quadratic twist $\widetilde{E}/\mathbb{F}_p$. At least one of the integers $\lambda(E(\mathbb{F}_p))$ and $\lambda(\widetilde{E}(\mathbb{F}_p))$ has a unique multiple in the Hasse interval $\mathcal{H}(p) = [(\sqrt{p}-1)^2, (\sqrt{p}+1)^2]$.*

Proof. Let $E(\mathbb{F}_p) \simeq \mathbb{Z}/n\mathbb{Z} \oplus \mathbb{Z}/N\mathbb{Z}$ and $\widetilde{E}(\mathbb{F}_p) \simeq \mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/M\mathbb{Z}$, where $n|N$ and $m|M$. Let t be the trace of the Frobenius endomorphism π of E. We have $E[n] = E(\mathbb{F}_p)[n]$, so π fixes $E[n]$ and the matrix π_n corresponding to the restriction of π to $E[n]$ is the identity matrix. The matrix π_{n^2} then has the form

$$\pi_{n^2} = \begin{bmatrix} 1+an & bn \\ cn & 1+dn \end{bmatrix},$$

for some $a, b, c, d \in \mathbb{Z}/n\mathbb{Z}$, and we have

$$p \equiv \det \pi_{n^2} \equiv 1 + (a+d)n \bmod n^2,$$
$$t \equiv \operatorname{tr} \pi_{n^2} \equiv 2 + (a+d)n \bmod n^2.$$

It follows that $4p - t^2 \equiv 0 \bmod n^2$. The trace of Frobenius for $\widetilde{E}$ is $-t$, and we similarly obtain $4p - t^2 \equiv 0 \bmod m^2$. Thus $\operatorname{lcm}(m^2, n^2)$ divides $4p - t^2$. We also have $t = un + 2$ and $t = vm - 2$, for some integers u and v, and subtracting these equations yields $vm - un = 4$. This implies $\gcd(m, n) \le 4$, and therefore $\gcd(m^2, n^2) \le 16$. Thus

$$\frac{m^2 n^2}{16} \le \operatorname{lcm}(m^2, n^2) \le 4p - t^2 \le 4p. \tag{2}$$

Suppose for the sake of contradiction that $N = \lambda(E(\mathbb{F}_p))$ and $M = \lambda(\widetilde{E}(\mathbb{F}_p))$ both have more than one multiple in $\mathcal{H}(p)$. Then M and N are both less than $4\sqrt{p}$ and $MN < 16p$. Since mM and nN lie in $\mathcal{H}(p)$, both are greater than $(\sqrt{p}-1)^2$, and $mnMN > (\sqrt{p}-1)^4$. It follows that $mn > (\sqrt{p}-1)^4/(16p)$. Dividing by 4 and squaring both sides yields

$$\frac{m^2 n^2}{16} > \frac{(\sqrt{p}-1)^8}{4096p^2}. \tag{3}$$

Combining (2) and (3), we have

$$16384p^3 > (\sqrt{p}-1)^8. \tag{4}$$

This implies that if neither M nor N has a unique multiple in $\mathcal{H}(p)$, then $p < 17413$. An exhaustive computer search for $p < 17413$ finds that in fact we must have $p \le 229$. □

7.8 Computing the group order with Mestre's Theorem

We now give a complete algorithm to compute $\#E(\mathbb{F}_p)$ using Mestre's theorem, assuming that p is a prime greater than 229 (if p is smaller than this we can easily count points using one of our naïve algorithms); see [2] for an analogous algorithm that works for all prime powers $q > 49$. As usual, $\mathcal{H}(p) := [(\sqrt{p}-1)^2, (\sqrt{p}+1)^2]$ denotes the Hasse interval.

Algorithm 7.8. Given $E/\mathbb{F}_p$ with $p > 229$ prime, compute $\#E(\mathbb{F}_p)$ as follows:

1. Compute a quadratic twist $\widetilde{E}$ of E using a randomly chosen non-square $s \in \mathbb{F}_p$.
2. Let $E_0 = E$ and $E_1 = \widetilde{E}$, let $N_0 = N_1 = 1$, and let $i = 0$.
3. While neither N_0 nor N_1 has a unique multiple in $\mathcal{H}(p)$:
   a. Generate a random point $P \in E_i(\mathbb{F}_p)$.
   b. Find an integer $M \in \mathcal{H}(p)$ such that $MP = 0$.
   c. Factor M and compute $|P|$ via Algorithm 7.4.
   d. Replace N_i by $\operatorname{lcm}(N_i, |P|)$ and replace i by $1 - i$.
4. If N_0 has a unique multiple M_0 in $\mathcal{H}(p)$ return M_0, otherwise return $2p + 2 - M_1$, where M_1 is the unique multiple of N_1 in $\mathcal{H}(p)$ guaranteed by Mestre's theorem.

It is clear that the output of the algorithm is correct, and it follows from Theorems 7.6 and 7.7 that the expected number of iterations of step 3 is $O(1)$. We thus have a *Las Vegas* algorithm to compute $\#E(\mathbb{F}_p)$. Its running time is dominated by the time to find M in step 3b, and we obtain a total expected running time of $O(\sqrt{p}\ \mathsf{M}(\log p))$, or

$$O\big(\exp(n/2)\,\mathsf{M}(n)\big).$$

We now show how this complexity can be improved using the *baby-steps giant-steps* method to find a suitable M in step 3b.

7.9 Baby-steps giant-steps

The baby-steps giant-steps method is a generic group algorithm that was first introduced by Daniel Shanks in [5] and subsequently generalized by many authors. In the context of searching $\mathcal{H}(q)$ for an integer M such that $MP = 0$, it works as follows.

Algorithm 7.9. Given $P \in E(\mathbb{F}_q)$ compute $M \in \mathcal{H}(q)$ such that $MP = 0$:

1. Pick integers r and s such that $rs \geq 4\sqrt{q}$ and let $a := \lceil (\sqrt{q}-1)^2 \rceil = \min(\mathcal{H}(q) \cap \mathbb{Z})$.
2. Compute the set $S_{\text{baby}} := \{0, P, 2P, \ldots, (r-1)P\}$ of *baby steps*.
3. Compute the set $S_{\text{giant}} := \{aP, (a+r)P, (a+2r)P, \ldots, (a+(s-1)r)P\}$ of *giant steps*.
4. For each giant step $P_{\text{giant}} = (a+ir)P \in S_{\text{giant}}$, check whether $P_{\text{giant}} + P_{\text{baby}} = 0$ for some baby step $P_{\text{baby}} = jP \in S_{\text{baby}}$. If so, output $M = a + ri + j$.

Note that *every* integer in $\mathcal{H}(q)$ can be written in the form $a + ir + j$ with $0 \leq i < s$ and $0 \leq j < r$, and for at least one such M we must have

$$MP = (a+ri)P + jP = P_{\text{giant}} + P_{\text{baby}} = 0$$

for some $P_{\text{giant}} \in S_{\text{giant}}$ and $P_{\text{baby}} \in S_{\text{baby}}$; this shows that the algorithm is correct.

To implement this algorithm efficiently, we typically store the baby steps S_{baby} in a lookup table (such as a hash table or binary tree) and as each giant step P_{giant} is computed, we look up $-P_{\text{giant}}$ in this table. Alternatively, one may compute the sets S_{baby} and S_{giant} in their entirety, sort both sets, and then efficiently search for a match. In both cases, we need the points in S_{baby} and S_{giant} to be uniquely represented.

If we are using projective coordinates this means we must convert each point to affine form: the point $(x : y : z)$ is put in the form $(x/z : y/z : 1)$ by computing the inverse of z in $\mathbb{F}_q$. Done naïvely, this requires $r + s$ field inversions, which costs $O((r + s)\,\mathsf{M}(n) \log n)$, but by using the method described in the next section, it is possible to perform $r + s$ field inversions in $O((r + s)\,\mathsf{M}(n))$ time. Assuming this is done, if we choose $r \approx s \approx 2q^{1/4}$, then the running time of the algorithm above is $O(q^{1/4}\,\mathsf{M}(\log q))$.

Using the baby-steps giant-steps method to implement step 3b of Algorithm 7.8 thus allows us to compute $\#E(\mathbb{F}_q)$ in expected time

$$O\big(\exp(n/4)\,\mathsf{M}(n)\big).$$

7.10 Batching field inversions

Suppose we are given a list of elements $\alpha_1, \ldots, \alpha_m \in \mathbb{F}_q$ whose inverses we wish to compute. The following algorithm accomplishes this using just one field inversion.

Algorithm 7.10. Given $\alpha_1, \ldots, \alpha_m \in \mathbb{F}_q$ compute $\alpha_1^{-1}, \ldots, \alpha_m^{-1}$ as follows:

1. Set $\beta_0 = 1$ and $\beta_i = \beta_{i-1}\alpha_i$ for i from 1 to m. $[\beta_i = (\alpha_1 \cdots \alpha_i)]$
2. Compute $\gamma_m = \beta_m^{-1}$. $[\gamma_m = (\alpha_1 \cdots \alpha_m)^{-1}]$
3. For i from m down to 1:
   a. Compute $\alpha_i^{-1} = \beta_{i-1}\gamma_i$. $[\alpha_i^{-1} = (\alpha_1 \cdots \alpha_{i-1})(\alpha_1 \cdots \alpha_i)^{-1}]$
   b. Compute $\gamma_{i-1} = \gamma_i\alpha_i$. $[\gamma_{i-1} = (\alpha_1 \cdots \alpha_{i-1})^{-1}]$

The algorithm performs less than $3m$ multiplications in $\mathbb{F}_q$ and just one inversion in $\mathbb{F}_q$. Provided that $m = \Omega(\log n)$, its running time is $O(m\,\mathsf{M}(n))$.

In the context of Algorithm 7.9, if we are using a table of baby steps, we can compute all of the baby steps using projective coordinates, convert them to affine form using just one field inversion, and then construct the lookup table. For the giant steps we work in batches of size $m > \log n$, converting an entire batch to affine form using one field inversion and then performing table lookups.

7.11 Optimizations

There are a wide range of optimizations to the baby-steps giant-steps method that have been developed over the years. Here we mention just a few.

1. **Optimize expected time**: If we suppose that M is uniformly distributed over an interval of width N, then we should use $r \approx \sqrt{N/2}$ baby steps so that the average number of giant steps is $s/2 \approx \sqrt{2N}/2 = \sqrt{N/2}$.

2. **Optimize for known distribution**: In the case of elliptic curves we know that M is *not* uniformly distributed – it has a semi-circular distribution.[2] This means we should search from the middle outwards by taking our first giant step in the middle of the interval (at $q+1$), and then alternating steps on either side. We should also choose the number of baby steps to optimize the expected time, using the fact that the expected distance between M and the middle of the interval is $\frac{8}{3\pi}\sqrt{q}$.
3. **Fast inverses**: In groups such as $E(\mathbb{F}_q)$ where we can compute inverses very quickly (the inverse of the point (x,y) is just $(x,-y)$), it makes sense to compute $-P_{\text{giant}}$ at the same time we compute P_{giant} and see whether either matches a baby step; equivalently, whether $P_{\text{giant}} \pm P_{\text{baby}} = 0$ holds. This allows us to double the width of the giant steps and use half as many, or (better), reduce both the number of baby steps and giant steps by a factor of $\sqrt{2}$.
4. **Parity**: We can easily determine the parity of $\#E(\mathbb{F}_q)$ by checking whether it has a point of order 2. If the curve equation is $y^2 = f(x) = x^3 + Ax + B$, then $\#E(\mathbb{F}_q)$ has even parity if and only if $f(x)$ has a root in $\mathbb{F}_q$ (recall that points of order 2 have y-coordinate 0), which we can determine using a root-finding algorithm.[3] Once we know the parity of M we can modify Algorithm 8.1 to only use baby steps that correspond to multiples of P with the same parity (so if M is odd we compute baby steps $P, 3P, 5P, \ldots$, adding $2P$ to each previous step), and use giant steps with even parity. We should then reduce the number of baby steps by a factor of $\sqrt{2}$.

References

[1] William D. Banks and Igor E. Shparlinski, *Sato-Tate, cyclicity, and divisibility statistics on average for elliptic curves of small height*, Israel J. Math. **173** (2009), 253–277.

[2] John E. Cremona and Andrew V. Sutherland, *On a theorem of Mestre and Schoof*, Journal de Théorie des Nombres de Bordeaux **22** (2010), 353–358.

[3] Joachim von zur Gathen and Jürgen Gerhard, *Modern computer algebra*, third edition, Cambridge University Press, 2013.

[4] René Schoof, *Counting points on elliptic curves over finite fields*, Journal de Théorie des Nombres de Bordeaux **7** (1995), 219–254.

[5] Daniel Shanks, *Class number, a theory of factorization and genera*, in 1969 Number Theory Institute (Proc. Symp. Pure Math., Vol. XX), Amer. Math. Soc., 1971, 415–440.

[6] Joseph H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics **106**, second edition, Springer 2009.

[7] Lawrence C. Washington, *Elliptic Curves: Number theory and cryptography*, second edition, Chapman and Hall/CRC, 2008.

[2] This follows from results showing that the Sato-Tate conjecture holds "on average"; see [1].
[3] In fact we only need to check whether $\deg \gcd(x^q - x, f(x)) > 0$, so we can do this deterministically.

8 Schoof's algorithm

In the early 1980s, René Schoof [3, 4] introduced the first polynomial-time algorithm to compute $\#E(\mathbb{F}_q)$. Extensions of Schoof's algorithm remain the point-counting method of choice when the characteristic of $\mathbb{F}_q$ is large (e.g., when q is a cryptographic-size prime).[1]

Schoof's basic strategy is simple: compute the trace of Frobenius t modulo many small primes ℓ and use the Chinese remainder theorem to uniquely determine t, which then determines $\#E(\mathbb{F}_q) = q + 1 - t$. Here is a high-level version of the algorithm.

Algorithm 8.1. Given an elliptic curve E over a finite field $\mathbb{F}_q$ compute $\#E(\mathbb{F}_q)$ as follows:

1. Initialize $M \leftarrow 1$ and $t \leftarrow 0$.
2. While $M \le 4\sqrt{q}$, for increasing primes $\ell = 2, 3, 5, \dots$ that do not divide q:
   a. Compute $t_\ell = \operatorname{tr}\pi \bmod \ell$.
   b. Set $t \leftarrow \Big(M(M^{-1} \bmod \ell)t_\ell + \ell(\ell^{-1} \bmod M)t\Big) \bmod \ell M$ and then $M \leftarrow \ell M$.
3. If $t > M/2$ then set $t \leftarrow t - M$.
4. Output $q + 1 - t$.

Step 2b uses an iterative version of the Chinese remainder theorem to ensure that

$$t \equiv \operatorname{tr}\pi_E \bmod M$$

holds throughout.[2] This invariant holds trivially after step 1, modulo $M = 1$, and is maintained in step 2b: note that the integer $M(M^{-1} \bmod \ell)$ is congruent to $1 \bmod \ell$ and $0 \bmod M$, while the integer $\ell(\ell^{-1} \bmod M)$ is congruent to $0 \bmod \ell$ and $1 \bmod M$.

Once M exceeds $4\sqrt{q}$, the value of $t \in \mathbb{Z}/M\mathbb{Z}$ uniquely determines $\operatorname{tr}\pi_E \in \mathbb{Z}$: by Hasse's theorem, $|\operatorname{tr}\pi_E| \le 2\sqrt{q} < M/2$, and this allows us to determine the sign of $\operatorname{tr}\pi_E$ in step 3. The key to the algorithm is the implementation of step 2a, which is described in the next section, but let us first consider the primes ℓ that the algorithm uses. Let $\ell_{\max}$ be the largest prime ℓ for which the algorithm computes t_ℓ. The Prime Number Theorem implies[3]

$$\sum_{\text{primes } \ell \le x} \log \ell \sim x,$$

so $\ell_{\max} \approx \log(4\sqrt{q}) \approx \frac{1}{2}n = O(n)$, and we need $O\big(\frac{n}{\log n}\big)$ primes ℓ (as usual, $n = \log q$). The cost of updating t and M is bounded by $O(\mathsf{M}(n)\log n)$, thus if we can compute t_ℓ in time bounded by a polynomial in n and ℓ, then the whole algorithm will run in polynomial time.

8.1 Computing the trace of Frobenius modulo 2.

We first consider the case $\ell = 2$. Assuming q is odd (which we do), $t = q + 1 - \#E(\mathbb{F}_q)$ is divisible by 2 if and only if $\#E(\mathbb{F}_q)$ is divisible by 2, equivalently, if and only if $E(\mathbb{F}_q)$ contains a point of order 2. If E has Weierstrass equation $y^2 = f(x)$, then the points of

[1] There are deterministic p-adic algorithms for computing $\#E(\mathbb{F}_q)$ that are faster than Schoof's algorithm when the characteristic p of $\mathbb{F}_q$ is very small; see [2]. But their running times are exponential in $\log p$.

[2] There are faster ways to apply the Chinese remainder theorem; see [1, §10.3]. They are not relevant here because the complexity is overwhelmingly dominated by step 2a.

[3] In fact we only need Chebyshev's Theorem to get this.

order 2 in $E(\mathbb{F}_q)$ are precisely those of the form $(x_0, 0)$, where $x_0 \in \mathbb{F}_q$ is a root of $f(x)$. Recall from Lecture 4 that the distinct roots of f in $\mathbb{F}_q$ are precisely the roots of $\gcd(x^q - x, f(x))$. We can thus compute $t_2 \coloneqq \operatorname{tr} \pi_E \bmod 2$ as

$$t_2 = \begin{cases} 0 & \text{if } \deg\Big(\gcd(f(x), x^q - x)\Big) > 0; \\ 1 & \text{otherwise.} \end{cases}$$

Note that this is a deterministic computation (we need randomness to efficiently *find* the roots of f, but not to *count* them), and it takes $O(n\,\mathsf{M}(n))$ time.

Having addressed the case $\ell = 2$ we henceforth assume that ℓ is odd.

8.2 The characteristic equation of Frobenius modulo ℓ

Recall that for $E/\mathbb{F}_q$, the Frobenius endomorphism $\pi_E \in \operatorname{End}(E)$ is defined by the rational map $(x : y : z) \mapsto (x^q : y^q : z^q)$. By Theorem 6.18, it satisfies the characteristic equation

$$\pi_E^2 - t\pi_E + q = 0,$$

with $t = \operatorname{tr} \pi_E$ and $q = \deg \pi_E$. Restricting to the ℓ-torsion subgroup $E[\ell]$ yields

$$\pi_\ell^2 - t_\ell \pi_\ell + q_\ell = 0, \tag{1}$$

which we view as an identity in $\operatorname{End}(E[\ell])$. Here $t_\ell \equiv t \bmod \ell$ and $q_\ell \equiv q \bmod \ell$ can be viewed either as restrictions of the scalar multiplication maps $[t]$ and $[q]$, or simply as scalars in $\mathbb{Z}/\ell\mathbb{Z}$ multiplied by $[1]_\ell$, the restriction of $[1] \in \operatorname{End}(E)$ to $E[\ell]$ (equivalently the multiplicative identity in the ring $\operatorname{End}(E[\ell])$). We shall take the latter view, regarding

$$q_\ell = q_\ell \cdot [1]_\ell = [1]_\ell + \cdots + [1]_\ell$$

as the sum of q_ℓ copies of $[1]_\ell$, and similarly for t_ℓ. We can efficiently compute q_ℓ using our usual double-and-add method to perform scalar multiplication by q_ℓ, provided that we know how to explicitly represent and perform ring operations on elements of $\operatorname{End}(E[\ell])$; this is the topic of the next section.

Our strategy for determining t_ℓ is simple: for $c = 0, 1, \ldots, \ell - 1$ compute $\pi_\ell^2 - c\pi_\ell + q_\ell$ and check whether it is equal to 0.

The following lemma shows that whenever this occurs (which it must, since (1) guarantees this for $c = t_\ell$) we must have $c = t_\ell \in \mathbb{Z}/\ell\mathbb{Z}$. In fact we will prove something stronger.

Lemma 8.2. *Let $E/\mathbb{F}_q$ be an elliptic curve with Frobenius endomorphism π, let ℓ be a prime not dividing q, and let $P \in E[\ell]$ be nonzero. Suppose that for some integer c the equation*

$$\pi_\ell^2(P) - c\pi_\ell(P) + q_\ell(P) = 0$$

holds. Then $c \equiv t_\ell = \operatorname{tr} \pi \bmod \ell$.

Proof. From equation (1) we have

$$\pi_\ell^2(P) - t_\ell \pi_\ell(P) + q_\ell P = 0,$$

and we are assuming that

$$\pi_\ell^2(P) - c\pi_\ell(P) + q_\ell P = 0.$$

Subtracting these equations yields $(c - t_\ell)\pi_\ell(P) = 0$. Since $\pi_\ell P$ is a nonzero element of $E[\ell]$ and ℓ is prime, the point $\pi_\ell(P)$ has order ℓ, which must divide $c - t_\ell$. So $c \equiv t_\ell \bmod \ell$. □

8.3 Arithmetic in End($E[\ell]$)

Let $h = \psi_\ell(x, y)$ be the ℓth division polynomial of E. We have assumed that ℓ is odd, so by Lemma 5.20, we in fact have $h \in \mathbb{F}_q[x]$ (no dependence on y). As we proved in Lecture 5, a nonzero point $P = (x_0, y_0) \in E(\overline{\mathbb{F}}_q)$ lies in $E[\ell]$ if and only if $h(x_0) = 0$; this follows from Corollary 4.28 and Theorem 5.21. To represent elements of $\mathrm{End}(E[\ell])$ as rational maps, we can thus treat the polynomials appearing in these maps as elements of the ring

$$\mathbb{F}_q[x, y]/(h(x), y^2 - f(x)),$$

where $y^2 = f(x) = x^3 + Ax + B$ is the Weierstrass equation for E.

In the case of the Frobenius endomorphism, we have

$$\begin{aligned} \pi_\ell &= \big(x^q \bmod h(x),\ y^q \bmod (h(x), y^2 - f(x))\big) \\ &= \Big(x^q \bmod h(x),\ \Big(f(x)^{(q-1)/2} \bmod h(x)\Big)\, y\Big), \end{aligned} \tag{2}$$

and we also note that

$$[1]_\ell = (x \bmod h(x),\ \big(1 \bmod h(x)\big)\, y).$$

We can thus represent all of the nonzero endomorphisms that appear in equation (1) in the form $(a(x), b(x)\, y)$, where a and b are elements of the polynomial ring $R = \mathbb{F}_q[x]/(h(x))$ that we may uniquely represent as polynomials in $\mathbb{F}_q[x]$ of degree less than $\deg h = (\ell^2 - 1)/2$ by taking their remainders modulo h.

8.3.1 Multiplication in End($E[\ell]$).

If $\alpha_1 = \big(a_1(x), b_1(x)y\big)$ and $\alpha_2 = \big(a_2(x), b_2(x)y\big)$ are two elements of $\mathrm{End}(E[\ell])$, the product $\alpha_1\alpha_2$ in $\mathrm{End}(E[\ell])$ is defined by the composition

$$\alpha_1 \circ \alpha_2 = (a_1(a_2(x)),\ b_1(a_2(x))b_2(x)\, y),$$

where we may reduce $a_3(x) = a_1(a_2(x))$ and $b_3(x) = b_1(a_2(x))b_2(x)$ modulo $h(x)$.

8.3.2 Addition in End($E[\ell]$).

Addition of endomorphisms is defined pointwise in terms of addition on the elliptic curve. Given $\alpha_1 = \big(a_1(x), b_1(x)y\big)$ and $\alpha_2 = \big(a_2(x), b_2(x)y\big)$, to compute $\alpha_3 = \alpha_1 + \alpha_2$, we simply apply the formulas for point addition to the coordinate functions of α_1 and α_2. Recall that the general formula for addition of non-opposite affine points $(x_3, y_3) = (x_1, y_1) + (x_2, y_2)$ on the elliptic curve $E\colon y^2 = x^3 + Ax + B$ is given by the formulas

$$x_3 = m^2 - x_1 - x_2, \qquad y_3 = m(x_1 - x_3) - y_1,$$

where

$$m = \frac{y_1 - y_2}{x_1 - x_2} \quad (\text{if } x_1 \neq x_2), \qquad m = \frac{3x_1^2 + A}{2y_1} \quad (\text{if } x_1 = x_2).$$

Using the coordinate functions $x_1 = a_1(x),\ x_2 = a_2(x),\ y_1 = b_1(x)y,\ y_2 = b_2(x)y$, in the case $x_1 \neq x_2$ we have

$$m(x, y) = \frac{b_1(x) - b_2(x)}{a_1(x) - a_2(x)} y = r(x)y,$$

where $r = (b_1 - b_2)/(a_1 - a_2)$, and when $x_1 = x_2$ we have

$$m(x,y) = \frac{3a_1(x)^2 + A}{2b_1(x)y} = \frac{3a_1(x)^2 + A}{2b_1(x)f(x)}y = r(x)y,$$

where now $r = (3a_1^2 + A)/(2b_1 f)$. Noting that $m(x,y)^2 = (r(x)y)^2 = r(x)^2 f(x)$, the sum $\alpha_1 + \alpha_2 = \alpha_3 = (a_3(x),\ b_3(x)y)$ is defined by

$$\begin{aligned} a_3 &= r^2 f - a_1 - a_2, \\ b_3 &= r(a_1 - a_3) - b_1. \end{aligned}$$

In both cases, provided that the polynomial v in the denominator of the rational function $r = u/v$ is invertible in the ring $\mathbb{F}_q[x]/(h(x))$, we can express r as a polynomial $uv^{-1} \bmod h$ and write $\alpha_3 = \big(a_3(x), b_3(x)y\big)$ in our desired form, with $a_3, b_3 \in \mathbb{F}_q[x]/(h(x))$ uniquely represented by polynomials in $\mathbb{F}_q[x]$ of degree less than the degree of h.

But this may not always be possible, because the ℓ-division polynomial $h(x)$ need not be irreducible. Indeed, if ℓ divides $\#E(\mathbb{F}_q)$ it certainly will not be irreducible, since $h(x)$ will then have rational roots corresponding to the x-coordinates of rational points of order ℓ, and even when $\ell \nmid \#E(\mathbb{F}_q)$, if E admits a rational isogeny α of degree ℓ then $h(x)$ will be divisible by the polynomial of degree $(\ell - 1)/2$ whose roots are the x-coordinates of the nonzero points in the kernel of α. Thus the ring $\mathbb{F}_q[x]/(h(x))$ is not necessarily a field; it may contain zero divisors, and these elements are not invertible.

At first glance this might appear to be a problem, but in fact it can only help us. If we encounter a rational function $r = u/v$ whose denominator v is not invertible in $\mathbb{F}_q[x]/(h(x))$ then we can obtain a non-trivial factor of h by computing $\gcd(v,h)$: if $v = a_1 - a_2$ then v is nonzero and has degree less than $\deg h$, since in this case $a_1 \neq a_2$ and $\deg(a_1 - a_2) < \deg(h)$, and if $v = 2b_1 f$ then $\gcd(v,h)$ must divide b_1, because h and f cannot share a common factor (the roots of $f(x)$ in $\overline{\mathbb{F}}_q$ are x-coordinates of 2-torsion points, the roots of $h(x)$ in $\overline{\mathbb{F}}_q$ are x-coordinates of ℓ-torsion points, and $\ell \neq 2$), and $b_1 \neq 0$ has degree less than $\deg h$.

Our strategy in this situation is to simply replace h by $g = \gcd(v,h)$ and compute t_ℓ by working in the smaller quotient ring $\mathbb{F}_q[x]/(g(x))$, which will be faster because $\deg g < \deg h$; in fact in this situation we will always have $\deg g \leq (\ell - 1)/2$, which is much smaller than $\deg h = (\ell^2 - 1)/2$. Lemma 8.2 implies that we can restrict our attention to the action of π_ℓ on points $P \in E[\ell]$ whose x-coordinates are roots of $g(x)$, even if $\deg g = 1$.

8.4 Algorithm to compute the trace of Frobenius modulo ℓ

We now give an algorithm to compute t_ℓ, the trace of Frobenius modulo ℓ.

Algorithm 8.3. Given $E : y^2 = f(x)$ over $\mathbb{F}_q$ and an odd prime ℓ, compute t_ℓ as follows:

1. Compute the ℓth division polynomial $h = \psi_\ell \in \mathbb{F}_q[x]$ for E.
2. Compute $\pi_\ell = (x^q \bmod h,\ (f^{(q-1)/2} \bmod h)y)$ and $\pi_\ell^2 = \pi_\ell \circ \pi_\ell$.
3. Use scalar multiplication to compute $q_\ell = q_\ell[1]_\ell$, and then compute $\pi_\ell^2 + q_\ell$.
   (If a non-invertible denominator arises, update h and return to step 2).
4. Compute $0, \pi_\ell, 2\pi_\ell, 3\pi_\ell, \ldots, c\pi_\ell$, until $c\pi_\ell = \pi_\ell^2 + q_\ell$.
   (If a non-invertible denominator arises, update h and return to step 2).
5. Output $t_\ell = c$.

Throughout the algorithm, elements of $\mathrm{End}(E[\ell])$ are represented in the form $(a(x), b(x)y)$, with $a, b \in R = \mathbb{F}_q[x]/(h(x))$, and all polynomial operations take place in the ring R. If a non-invertible denominator v is found in either steps 3 or 4 we replace h with whichever of $\gcd(h, v)$ and $h/\gcd(h, v)$ has lower degree; this guarantees that the degree of h is reduced by at least a factor of 2 (but see the next section for a further discussion).

The correctness of the algorithm follows from equation (1) and Lemma 8.2. The algorithm is guaranteed to find some $c\pi_l = \pi_l^2 + q_l$ in step 4 with $c < \ell$, since we know that $c = t_\ell$ works. Although we may be working modulo a proper factor g of h, every root x_0 of g is a root of h and therefore corresponds to a pair of nonzero points $P = (x_0, \pm y_0) \in E[\ell]$ for which $\pi_\ell^2(P) - c\pi_\ell(P) + q_\ell P = 0$ holds (there is at least one such root, since $\deg g > 0$), and Lemma 8.2 implies that we must have $c = t_\ell$.

The computation of the division polynomial in step 1 of the algorithm can be efficiently accomplished using the double-and-add approach described in Problem Set 3. You will have the opportunity to do a careful complexity analysis of Algorithm 8.3 in the next problem set, but it is easy to see that its running time is polynomial in $n = \log q$ and ℓ: every operation involves polynomials over $\mathbb{F}_q$ of degree less than ℓ^2, in step 4 we can have at most ℓ iterations, and we can return to step 2 at most $2\log \ell$ times (in fact this can happen only once). A simple implementation of the algorithm can be found in this Sage notebook.

8.5 Factors of the division polynomial

As we saw when running our implementation of Schoof's algorithm in Sage, we do occasionally encounter non-invertible denominators and thereby obtain a proper factor g of the ℓ-division polynomial $h = \psi_\ell$. This is not too surprising, since there is no reason why h should necessarily be irreducible, but it is worth noting that whenever this occurs the degree of g is always exactly $(\ell - 1)/2$. Why is this the case?

Any point $P = (x_0, y_0) \in E(\overline{\mathbb{F}}_q)$ for which $g(x_0) = 0$ lies both in $E[\ell]$ and in the kernel of an endomorphism α (since x_0 is a root of the denominator of a rational function defining α). The point P is nonzero, so it generates a cyclic group C of order ℓ which must be a subgroup of $\ker \alpha$. It follows that over $\overline{\mathbb{F}}_q$ the polynomial g has at least $(\ell - 1)/2$ roots, one for each pair of nonzero points $(x_i, \pm y_i)$ in C (note that ℓ is odd). If g has any other roots, then there is a point Q that lies in the intersection of $E[\ell] \cap \ker \alpha$ but not in C, in which case we must have $\ker \alpha = E[\ell]$, since $E[\ell]$ has ℓ-rank 2; but this is impossible because g is a proper factor of the ℓ-division polynomial h (whose roots are distinct because $\ell \nmid q$). So g must have exactly $(\ell-1)/2$ roots in $\overline{\mathbb{F}}_q$. Reducing the polynomials that define our endomorphism modulo g corresponds to working in the subring $\mathrm{End}(C)$ of $\mathrm{End}(E[\ell])$.

If we are lucky enough to find such a proper factor g of h, our algorithm then speeds up by at least a factor of ℓ, since we are working modulo a polynomial of degree $(\ell - 1)/2$ rather than $(\ell^2 - 1)/2$. While we are fairly unlikely to stumble across such a g by chance, it turns out that in fact such a g exists for half of the primes ℓ (asymptotically speaking). Not long after Schoof published his result, Noam Elkies found a way to directly compute these polynomials, whose roots are the x-coordinates of points $P = (x_0, y_0)$ that lie in the kernel of a rational isogeny of degree ℓ. We will learn about Elkies' technique later in the course when we discuss modular polynomials. There is another optimization due to A.O.L. Atkin that applies to primes ℓ for which Elkies' optimization does not; together these yield what is known as the Schoof-Elkies-Atkin (SEA) algorithm.

8.6 Some historical remarks

When Schoof originally developed this algorithm, it was not clear to him that it had any practical use. This is in part because he (and others) were unduly pessimistic about its practical efficiency, in part because robust implementations of fast integer and polynomial arithmetic were not as widely available then as they are now. Even the simple Sage implementation given in the worksheet is already noticeably faster than the baby-steps giant-steps algorithm for $q \approx 2^{80}$ and can readily handle computations over fields of cryptographic size (it might take a day or two for $q \approx 2^{256}$, but this could be improved by at least an order of magnitude using a lower-level implementation in C or C++).

To better motivate his algorithm, Schoof gave an application that is of purely theoretical interest: he showed that it could be used to deterministically compute the square root of an integer a modulo a prime p in time that grows polynomially in $\log p$ when a is held fixed; we will see exactly how this works when we cover the theory of complex multiplication. Previously, no deterministic polynomial-time algorithm was known for this problem, unless one assumes the extended Riemann hypothesis. But Schoof's square-root application is really of no practical use; as we have seen, there are fast probabilistic algorithms to compute square roots modulo a prime, and unless the extended Riemann hypothesis is false, there are even deterministic algorithms that are much faster than Schoof's approach.

By contrast, in showing how to compute $\#E(\mathbb{F}_q)$ in polynomial time, Schoof solved a practically important problem for which the best previously known algorithms were fully exponential (including randomized algorithms), despite the efforts of many experts working in the field. While perhaps not fully appreciated at the time, this has to be regarded as a major breakthrough, both from a theoretical and practical perspective. Improved versions of Schoof's algorithm (the SEA algorithm) are now the method of choice for computing $\#E(\mathbb{F}_q)$ in fields of large characteristic. In particular, the PARI/GP library that is used by Sage includes an implementation of the SEA algorithm, and over 256-bit fields it takes only a few seconds to compute $\#E(\mathbb{F}_q)$. Today it is feasible to compute $\#E(\mathbb{F}_q)$ even when q is a prime with 5,000 decimal digits (over 16,000 bits), which represents the current record [5].

References

[1] Joachim von zur Gathen and Jürgen Gerhard, *Modern computer algebra*, third edition, Cambridge University Press, 2013.

[2] Takakazu Satoh, *On p-adic point counting algorithms for elliptic curves over finite fields*, ANTS V, LNCS **2369** (2002), 43–66.

[3] René Schoof, *Elliptic curves over finite fields and the computation of square roots mod p*. Mathematics of Computation **44** (1985), 483–495.

[4] René Schoof, *Counting points on elliptic curves over finite fields*, Journal de Théorie des Nombres de Bordeaux **7** (1995), 219–254.

[5] Andrew V. Sutherland, *On the evaluation of modular polynomials*, in Proceedings of the Tenth Algorithmic Number Theory Symposium (ANTS X), Open Book Series **1**, Mathematical Science Publishers, 2013, 531–555.

9 The discrete logarithm problem

In its most standard form, the *discrete logarithm problem* in a finite group G can be stated as follows:

> *Given $\alpha \in G$ and $\beta \in \langle\alpha\rangle$, find the least positive integer x such that $\alpha^x = \beta$.*

In additive notation (which we will often use), this means $x\alpha = \beta$. In either case, we call x the discrete logarithm of β with respect to the base α and denote it $\log_\alpha \beta$.[1] Note that in the form stated above, where x is required to be positive, the discrete logarithm problem includes the problem of computing the order of α as a special case: $|\alpha| = \log_\alpha 1_G$.

We can also formulate a slightly stronger version of the problem:

> *Given $\alpha, \beta \in G$, compute $\log_\alpha \beta$ if $\beta \in \langle\alpha\rangle$ and otherwise report that $\beta \notin \langle\alpha\rangle$.*

This can be a significantly harder problem. For example, if we are using a Las Vegas algorithm, when β lies in $\langle\alpha\rangle$ we are guaranteed to eventually find $\log_\alpha \beta$, but if not, we will never find it and it may be impossible to tell whether we are just very unlucky or $\beta \notin \langle\alpha\rangle$. On the other hand, with a deterministic algorithm such as the baby-steps giant-steps method, we can unequivocally determine whether β lies in $\langle\alpha\rangle$ or not.

There is also a generalization called the *extended discrete logarithm*:

> *Given $\alpha, \beta \in G$, determine the least positive integer y such that $\beta^y \in \langle\alpha\rangle$, and then output the pair (x, y), where $x = \log_\alpha \beta^y$.*

This yields positive integers x and y satisfying $\beta^y = \alpha^x$, where we minimize y first and x second. Note that there is always a solution: in the worst case $x = |\alpha|$ and $y = |\beta|$.

Finally, one can also consider a vector form of the discrete logarithm problem:

> *Given $\alpha_1, \dots, \alpha_r \in G$ and $n_1, \dots, n_r \in \mathbb{Z}_{>0}$ such that every $\beta \in G$ can be written uniquely as $\beta = \alpha_1^{e_1} \cdots \alpha_r^{e_r}$ with $e_i \in [1, n_i]$, compute the exponent vector $(e_1, \dots, e_r)$ associated to a given β.*

Note that the group G need not be abelian in order for the hypothesis to apply, it suffices for G to be *polycyclic* (this means it admits a subnormal series with cyclic quotients).

The extended discrete and vector forms of the discrete logarithm problem play an important role in algorithms to compute the structure of a finite abelian group, but in the lectures we will focus primarily on the standard form of the discrete logarithm problem (which we may abbreviate to DLP).

Example 9.1. Suppose $G = \mathbb{F}_{101}^\times$. Then $\log_3 37 = 24$, since $3^{24} \equiv 37 \bmod 101$.

Example 9.2. Suppose $G = \mathbb{F}_{101}^+$. Then $\log_3 37 = 46$, since $46 \cdot 3 \equiv 37 \bmod 101$.

Both of these examples involve groups where the discrete logarithm is easy to compute (and not just because 101 is a small number), but for very different reasons. In Example 9.1 we are working in a group of order $100 = 2^2 \cdot 5^2$. As we will see in the next lecture, when the group order is a product of small primes (i.e. *smooth*), it is easy to compute discrete logarithms. In Example 9.2 we are working in a group of order 101, which is prime, and in

[1] The multiplicative terminology stems from the fact that most of the early work on computing discrete logarithms focused on the case where G is the multiplicative group of a finite field.

terms of the group structure, this represents the hardest case. But in fact it is very easy to compute discrete logarithms in the additive group of a finite field! All we need to do is compute the multiplicative inverse of 3 modulo 101 (which is 34) and multiply by 37. This is a small example, but even if the field size is very large, we can use the extended Euclidean algorithm to compute multiplicative inverses in quasi-linear time.

So while the DLP is generally considered a "hard problem", its difficulty really depends not on the order of the group (or its structure), but on how the group is explicitly represented. Every group of prime order p is isomorphic to $\mathbb{Z}/p\mathbb{Z}$; computing the discrete logarithm amounts to computing this isomorphism. The reason it is easy to compute discrete logarithms in $\mathbb{Z}/p\mathbb{Z}$ has nothing to do with the structure of $\mathbb{Z}/p\mathbb{Z}$ as an additive group, rather it is the fact that $\mathbb{Z}/p\mathbb{Z}$ also has a ring structure; in particular, it is a Euclidean domain, which allows us to use the extended Euclidean algorithm to compute multiplicative inverses. This involves operations (multiplication) other than the standard group operation (addition), which is in some sense "cheating".

Even when working in the multiplicative group of a finite field, where the DLP is believed to be much harder, we can do substantially better than in a generic group. As we shall see, there are sub-exponential time algorithms for this problem, whereas in the generic setting defined below, only exponential time algorithms exist, as we will prove in the next lecture.

9.1 Generic group algorithms

In order to formalize the notion of "not cheating", we define a *generic group algorithm* (or just a *generic algorithm*) to be one that interacts with an abstract group G solely through a *black box* (sometimes called an *oracle*). All group elements are opaquely encoded as bit-strings via a map $\mathsf{id}\colon G \to \{0,1\}^m$ chosen by the black box. The black box supports the following operations.

1. `identity`: output $\mathsf{id}(1_G)$.
2. `inverse`: given input $\mathsf{id}(\alpha)$, output $\mathsf{id}(\alpha^{-1})$.
3. `compose`: given inputs $\mathsf{id}(\alpha)$ and $\mathsf{id}(\beta)$, output $\mathsf{id}(\alpha\beta)$.
4. `random`: output $\mathsf{id}(\alpha)$ for a uniformly distributed random element $\alpha \in G$.

In the description above we used multiplicative notation; in additive notation we would have outputs $\mathsf{id}(0_G)$, $\mathsf{id}(-\alpha)$, $\mathsf{id}(\alpha+\beta)$ for the operations `identity`, `inverse`, `compose`, respectively.

Some models for generic group algorithms also include a black box operation for testing equality of group elements, but we will instead assume that group elements are *uniquely identified*; this means that the identification map $\mathsf{id}\colon G \to \{0,1\}^m$ used by the black box is injective. With uniquely identified group elements we can test equality by simply comparing identifiers, without needing to consult the black box.[2]

The black box is allowed to use *any* injective identification map (e.g., a random one). A *generic algorithm* cannot depend on a particular choice of the identification map; this prevents it from taking advantage of how group elements are represented. We have already seen several examples of generic group algorithms, including various exponentiation algorithms, fast order algorithms, and the baby-steps giant-steps method.

[2]We can also sort bit-strings or index them with a hash table or other data structure; this is essential to an efficient implementation of the baby-steps giant-steps algorithm.

We measure the time complexity of a generic group algorithm by counting *group operations*, the number of interactions with the black box. This metric has the virtue of being independent of the actual software and hardware implementation, allowing one to make comparisons that remain valid even as technology improves. But if we want to get a complete measure of the complexity of solving a problem in a particular group, we need to multiply the group operation count by the bit-complexity of each group operation, which of course depends on the black box. To measure the space complexity, we count the total number of group identifiers stored at any one time (i.e. the maximum number of group identifiers the algorithm ever has to remember).

These complexity metrics do not account for any other work done by the algorithm. If the algorithm wants to compute a trillion digits of pi, or factor some huge integer, it can effectively do so "for free". The implicit assumption is that the cost of any useful auxiliary computations are at worst proportional to the number of group operations — this is true of all the algorithms we will consider.

9.2 Generic algorithms for the discrete logarithm problem

We now consider generic algorithms for the discrete logarithm problem in the standard setting of a cyclic group $\langle\alpha\rangle$. We shall assume throughout that $N := |\alpha|$ is known. This is a reasonable assumption for three reasons: (1) in cryptographic applications it is quite important to know N (or at least to know that N is prime), (2) the lower bound we shall prove applies even when the algorithm is given N, (3) for a generic algorithm, computing $|\alpha|$ is strictly easier than solving the discrete logarithm problem [12], and in most cases of practical interest (the group of rational points on an elliptic curve over a finite field, for example), there are (non-generic) polynomial time algorithms to compute N.

The cyclic group $\langle\alpha\rangle$ is isomorphic to the additive group $\mathbb{Z}/N\mathbb{Z}$. For generic group algorithms we may as well assume $\langle\alpha\rangle$ *is* $\mathbb{Z}/N\mathbb{Z}$, generated by $\alpha = 1$, since every cyclic group of order N looks the same when it is hidden inside a black box. Of course with the black box picking arbitrary group identifiers in $\{0,1\}^m$, we cannot actually tell which integer x in $\mathbb{Z}/N\mathbb{Z}$ corresponds to a particular group element β; indeed, x is precisely the discrete logarithm of β that we wish to compute! Thus computing discrete logarithms amounts to explicitly computing the isomorphism from $\langle\alpha\rangle$ to $\mathbb{Z}/N\mathbb{Z}$ that sends α to 1. Computing the isomorphism in the reverse direction is easy: this is just exponentiation. Thus we have (in multiplicative notation):

$$\begin{aligned}\langle\alpha\rangle &\simeq \mathbb{Z}/N\mathbb{Z}\\ \beta &\to \log_\alpha \beta\\ \alpha^x &\leftarrow x\end{aligned}$$

Cryptographic applications of the discrete logarithm problem rely on the fact that it is easy to compute $\beta = \alpha^x$ but hard (in general) to compute $x = \log_\alpha \beta$. In order to simplify our notation we will write the group operation additively, so that $\beta = x\alpha$.

9.3 Linear search

Starting from α, compute

$$\alpha, 2\alpha, 3\alpha, \ldots, x\alpha = \beta,$$

and then output x (or if we reach $N\alpha$ without finding β, report that $\beta \notin \langle\alpha\rangle$). This uses at most N group operations, and the average over all inputs is $N/2$ group operations.

We mention this algorithm only for the sake of comparison. Its time complexity is not attractive, but we note that its space complexity is $O(1)$ group elements.

9.4 Baby-steps giant-steps

Pick positive integers r and s such that $rs > N$, and then compute:

$$\begin{array}{ll} \text{baby steps:} & 0,\ \alpha,\ 2\alpha,\ 3\alpha,\ \ldots,\ (r-1)\alpha, \\ \text{giant steps:} & \beta,\ \beta - r\alpha,\ \beta - 2r\alpha,\ \ldots,\ \beta - (s-1)r\alpha, \end{array}$$

A collision occurs when we find a baby step that is equal to a giant step. We then have

$$i\alpha = \beta - jr\alpha,$$

for some nonnegative integers $i < r$ and $j < s$. If $i = j = 0$, then β is the identity and $\log_\alpha \beta = N$. Otherwise,

$$\log_\alpha \beta = i + jr.$$

Typically the baby steps are stored in a lookup table, allowing us to check for a collision as each giant step is computed, so we don't necessarily need to compute all the giant steps. We can easily detect $\beta \notin \langle\alpha\rangle$, since every integer in $[1, N]$ can be written in the form $i + jr$ with $0 \le i < r$ and $0 \le j < s$. If we do not find a collision, then $\beta \notin \langle\alpha\rangle$.

The baby-steps giant-steps algorithm uses $r + s$ group operations, which is $O(\sqrt{N})$ if we choose $r \approx s \approx \sqrt{N}$. It requires space for r group elements (the baby steps), which is also $O(\sqrt{N})$ but can be made smaller if we are willing to increase the running time by making s larger; there is thus a time-space trade-off we can make, but the product of the time and space complexity is always $\Omega(N)$.

The two algorithms above are insensitive to any special properties of N, their complexities depend only on its approximate size. In fact, if we assume that $\beta \in \langle\alpha\rangle$ then we do not even need to know N: this is clear for the linear search, and for the baby-steps giant-steps method we could simply start by assuming $N = 2$ and if/when that fails, keep doubling N and rerunning the algorithm until we succeed. This still yields an $O(\sqrt{N})$ complexity.[3]

For the next algorithm we consider, it is quite important to know N exactly, in fact we will assume that we know its prime factorization; factoring N does not require any group operations, so in our complexity model which counts group operations, a generic algorithm can factor any integer N "for free". In practical terms, there are algorithms to factor N that are much faster than the generic lower bound we will prove below; as we will see in the next lecture, the elliptic curve factorization method is one such algorithm.

9.5 The Pohlig-Hellman algorithm

We now introduce the Pohlig-Hellman[4] algorithm, a recursive method to reduce the discrete logarithm problem in cyclic groups of composite order to discrete logarithm problems in cyclic groups of prime order.

[3] There are more efficient ways to do an unbounded baby-steps giant-steps search, see [12, 14].

[4] The article by Pohlig and Hellman [6] notes that essentially equivalent versions of the algorithm were independently found by R. Silver, and by R. Schroeppel and H. Block, none of whom published the result.

We first reduce to the prime power case. Suppose $N = N_1N_2$ with $N_1 \perp N_2$. Then $\mathbb{Z}/N\mathbb{Z} \simeq \mathbb{Z}/N_1\mathbb{Z} \oplus \mathbb{Z}/N_2\mathbb{Z}$, by the Chinese remainder theorem, and we can make this isomorphism completely explicit via

$$\begin{array}{rcl} x & \to & (x \bmod N_1,\ x \bmod N_2), \\ (M_1x_1 + M_2x_2) \bmod N & \leftarrow & (x_1,\ x_2), \end{array}$$

where

$$M_1 = N_2(N_2^{-1} \bmod N_1) \equiv \begin{cases} 1 \bmod N_1, \\ 0 \bmod N_2, \end{cases} \tag{1}$$

$$M_2 = N_1(N_1^{-1} \bmod N_2) \equiv \begin{cases} 0 \bmod N_1, \\ 1 \bmod N_2. \end{cases} \tag{2}$$

Note that computing M_i and N_i does not involve group operations and is independent of β; with the fast Euclidean algorithm the time to compute M_1 and M_2 is $O(\mathsf{M}(n)\log n)$ bit operations, where $n = \log N$.

Let us now consider the computation of $x = \log_\alpha \beta$. Define

$$x_1 := x \bmod N_1 \qquad \text{and} \qquad x_2 := x \bmod N_2,$$

so that $x = M_1x_1 + M_2x_2$, and

$$\beta = (M_1x_1 + M_2x_2)\alpha.$$

Multiplying both sides by N_2 and distributing the scalar multiplication yields

$$N_2\beta = M_1x_1N_2\alpha + M_2x_2N_2\alpha. \tag{3}$$

As you proved in Problem Set 1, the order of $N_2\alpha$ is N_1 (since $N_1 \perp N_2$). From (1) and (2) we have $M_1 \equiv 1 \bmod N_1$ and $M_2 \equiv 0 \bmod N_1$, so the second term in (3) vanishes and the first term can be simplified, yielding

$$N_2\beta = x_1N_2\alpha.$$

We similarly find that $N_1\beta = x_2N_1\alpha$. Therefore

$$\begin{aligned} x_1 &= \log_{N_2\alpha} N_2\beta, \\ x_2 &= \log_{N_1\alpha} N_1\beta. \end{aligned}$$

If we know x_1 and x_2 then we can compute $x = (M_1x_1 + M_2x_2) \bmod N$. Thus the computation of $x = \log_\alpha \beta$ can be reduced to the computation of $x_1 = \log_{N_2\alpha} N_2\beta$ and $x_2 = \log_{N_1\alpha} N_1\beta$. If N is a prime power this doesn't help (either $N_1 = N$ or $N_2 = N$), but otherwise these two discrete logarithms involve groups of smaller order. In the best case $N_1 \approx N_2$, and we reduce our original problem to two subproblems of half the size, and this reduction involves only $O(n)$ groups operations (the time to compute $N_1\alpha, N_1\beta, N_2\alpha, N_2\beta$ using double-and-add scalar multiplication).

By applying the reduction above recursively, we can reduce to the case where N is a prime power p^e, which we now assume. Let $e_0 = \lceil e/2 \rceil$ and $e_1 = \lfloor e/2 \rfloor$. We may write $x = \log_\alpha \beta$ in the form $x = x_0 + p^{e_0}x_1$, with $0 \le x_0 < p^{e_0}$ and $0 \le x_1 < p^{e_1}$. We then have

$$\begin{aligned} \beta &= (x_0 + p^{e_0}x_1)\alpha, \\ p^{e_1}\beta &= x_0p^{e_1}\alpha + x_1p^e\alpha. \end{aligned}$$

The second term in the last equation is zero, since α has order $N = p^e$, so

$$x_0 = \log_{p^{e_1}\alpha} p^{e_1}\beta.$$

We also have $\beta - x_0\alpha = p^{e_0}x_1\alpha$, therefore

$$x_1 = \log_{p^{e_0}\alpha}(\beta - x_0\alpha).$$

If N is not prime, this again reduces the computation of $\log_\alpha \beta$ to the computation of two smaller discrete logarithms (of roughly equal size) using $O(n)$ group operations.

The Pohlig-Hellman method [6] recursively applies the two reductions above to reduce the problem to a set of discrete logarithm computations in groups of prime order.[5] For these computations we must revert to some other method, such as baby-steps giant-steps (or Pollard-rho, which we will see shortly). When N is a prime p, the complexity is then $O(\sqrt{p})$ group operations.

9.6 Complexity analysis

Let $N = p_1^{e_1} \cdots p_r^{e_r}$ be the prime factorization of N. Reducing to the prime-power case involves at most $\lg r = O(\log n)$ levels of recursion, where $n = \log N$ (in fact the prime number theorem implies $\lg r = O(\log n/\log\log n)$, but we won't use this). The exponents e_i are all bounded by $\lg N = O(n)$, thus reducing prime powers to the prime case involves at most an additional $O(\log n)$ levels of recursion, since the exponents are reduced by roughly a factor of 2 at each level.

The total depth of the recursion tree is thus $O(\log n)$. Note that we do not need to assume anything about the prime factorization of N in order to obtain this bound; in particular, even if the prime powers $p_i^{e_i}$ vary widely in size, this bound still holds.

The product of the orders of the bases used at any given layer of the recursion tree never exceeds N. The number of group operations required at each internal node of the recursion tree is linear in the bit-size of the order of the base, since only $O(1)$ scalar multiplications are used in each recursive step. Thus if we exclude the primes order cases at the leaves, every layer of the recursion tree uses $O(n)$ group operations. If we use the baby-steps giant-steps algorithm to handle the prime order cases, each leaf uses $O(\sqrt{p_i})$ group operations and the total running time is

$$O\left(n\log n + \sum e_i\sqrt{p_i}\right)$$

group operations, where the sum is over the distinct prime divisors p_i of N. We can also bound this by

$$O\left(n\log n + n\sqrt{p}\right),$$

where p is the largest prime dividing N. The space complexity is $O(\sqrt{p})$ group elements, assuming we use a baby-steps giant-steps search for the prime cases; this can be reduced to $O(1)$ using the Pollard-rho method (which is the next algorithm we will consider), but this results in a probabilistic (Las Vegas) algorithm, whereas the standard Pohlig-Hellman approach is deterministic.

[5] The original algorithm of Pohlig and Hellman actually used an iterative approach that is not as fast as the recursive approach suggested here. The recursive approach for the prime-power case that we use here appears in [11, §11.2.3]. When $N = p^e$ is a power of a prime $p = O(1)$, the complexity of the original Pohlig-Hellman algorithm is $O(n^2)$, versus the $O(n\log n)$ bound we obtain here (this can be further improved to $O(n\log n/\log\log n)$ via [13]).

The Pohlig-Hellman algorithm can be extremely efficient when N is composite; if N is sufficiently smooth its running time is quasi-linear in $n = \log N$, comparable to the cost of exponentiation. Thus it is quite important to use groups of prime (or near-prime) order in cryptographic applications of the discrete logarithm problem. This is one of the motivations for efficient point-counting algorithms for elliptic curves: we really need to know the exact group order before we can consider a group suitable for cryptographic use.

9.7 Randomized algorithms for the discrete logarithm problem

So far we have only considered deterministic algorithms for the discrete logarithm problem. We now consider a probabilistic approach. Randomization will not allow us to achieve a better time complexity (a fact we will prove shortly), but we can achieve a much better space complexity. This also makes it much easier to parallelize the algorithm, which is crucial for large-scale computations (one can construct a parallel version of the baby-steps giant-steps algorithm, but detecting collisions is more complicated and requires a lot of communication).

9.7.1 The birthday paradox

Recall what the so-called *birthday paradox* tells us about collision frequency: if we drop $\Omega(\sqrt{N})$ balls randomly into $O(N)$ bins then the probability that some bin contains more than one ball is bounded below by some nonzero constant that we can make arbitrarily close to 1 by increasing the number of balls by a constant factor. Given $\beta \in \langle\alpha\rangle$, the baby-steps giant-steps method for computing $\log_\alpha \beta$ can be viewed as dropping $\sqrt{2N}$ balls corresponding to linear combinations of α and β into N bins corresponding to the elements of $\langle\alpha\rangle$. Of course these balls are not dropped randomly, they are dropped in a pattern that guarantees a collision.

But if we instead computed $\sqrt{2N}$ random linear combinations of α and β, we would still have a good chance of finding a collision (better than 50/50, in fact). The main problem with this approach is that in order to find the collision we would need to keep a record of all the linear combinations we have computed, which takes space. In order to take advantage of the birthday paradox in a way that uses less space we need to be a bit more clever.

9.7.2 Random walks on a graph

We now want to view the group $G = \langle\alpha\rangle$ as the vertex set V of a connected graph Γ whose edges $e_{ij} = (\gamma_i, \gamma_j)$ are labeled with the group element $\delta_{ij} = \gamma_j - \gamma_i$ satisfying $\gamma_i + \delta_{ij} = \gamma_j$ (a Cayley graph, for example). If we know how to express each δ_{ij} as a linear combination of α and $\beta \in \langle\alpha\rangle$, then any cycle in Γ yields a linear relation involving α and β. Provided the coefficient of β is invertible modulo $N := |\alpha|$, we can use this relation to compute $\log_\alpha \beta$.

Suppose we use a random function $f\colon V \to V$ to construct a walk from a random starting point $v_0 \in V$ as follows:

$$\begin{aligned} v_1 &= f(v_0) \\ v_2 &= f(v_1) \\ v_3 &= f(v_2) \\ &\vdots \end{aligned}$$

Since f is a function, if we ever repeat a vertex, say $v_\rho = v_\lambda$ for some $\rho > \lambda$, we will be permanently stuck in a cycle, since we then have $f(v_{\rho+i}) = f(v_{\lambda+i})$ for all $i \geq 0$. Note that V is finite, so this must happen eventually.

Our random walk consists of two parts, a path from v_0 to the vertex v_λ, the first vertex that is visited more than once, and a cycle consisting of the vertices $v_\lambda, v_{\lambda+1}, \ldots, v_{\rho-1}$. This can be visualized as a path in the shape of the Greek letter ρ, which explains the name of the *ρ-method* we wish to consider.

In order to extract information from this cycle we need to augment the function f so that we can associate linear combinations $a\alpha + b\beta$ to each edge in the cycle. But let us first compute the expected number of steps a random walk takes to reach its first collision.

Theorem 9.3. *Let V be a finite set. For any $v_0 \in V$, the expected value of ρ for a walk from v_0 defined by a random function $f\colon V \to V$ is*

$$\mathrm{E}[\rho] \sim \sqrt{\pi N/2},$$

as the cardinality N of V tends to infinity.

This theorem was stated in lecture without proof; here we give an elementary proof.

Proof. Let $P_n = \Pr[\rho > n]$. We have $P_0 = 1$ and $P_1 = (1 - 1/N)$, and in general

$$P_n = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{n}{N}\right) = \prod_{i=1}^{n}\left(1 - \frac{i}{N}\right)$$

for any $n < N$ (and $P_n = 0$ for $n \geq N$). We compute the expectation of ρ as

$$\begin{aligned}
\mathrm{E}[\rho] &= \sum_{n=1}^{N-1} n \cdot \Pr[\rho = n] \\
&= \sum_{n=1}^{N-1} n \cdot (P_{n-1} - P_n), \\
&= 1(P_0 - P_1) + 2(P_1 - P_2) + \ldots + n(P_{n-1} - P_n) \\
&= \sum_{n=0}^{N-1} P_n - nP_n. \qquad (4)
\end{aligned}$$

In order to determine the asymptotic behavior of $\mathrm{E}[\rho]$ we need tight bounds on P_n. Using the fact that $\log(1-x) < -x$ for $0 < x < 1$, we obtain an upper bound on P_n:

$$\begin{aligned}
P_n &= \exp\left(\sum_{i=1}^{n} \log\left(1 - \frac{i}{N}\right)\right) \\
&< \exp\left(-\frac{1}{N}\sum_{i=1}^{n} i\right) \\
&< \exp\left(\frac{-n^2}{2N}\right).
\end{aligned}$$

To establish a lower bound, we use the fact that $\log(1-x) > -x - x^2$ for $0 < x < \frac{1}{2}$, which can be verified using the Taylor series expansion for $\log(1-x)$.

$$\begin{aligned} P_n &= \exp\left(\sum_{i=1}^{n} \log\left(1 - \frac{i}{N}\right)\right) \\ &> \exp\left(-\sum_{i=1}^{n}\left(\frac{i}{N} + \frac{i^2}{N^2}\right)\right). \end{aligned}$$

We now let $M = N^{3/5}$ and assume $n < M$. In this range we have

$$\begin{aligned} \sum_{i=1}^{n}\left(\frac{i}{N} + \frac{i^2}{N^2}\right) &< \sum_{i=1}^{n}\left(\frac{i}{N} + N^{-\frac{4}{5}}\right) \\ &< \frac{n^2+n}{2N} + N^{-\frac{1}{5}} \\ &< \frac{n^2}{2N} + \frac{1}{2}N^{-\frac{2}{5}} + N^{-\frac{1}{5}} \\ &< \frac{n^2}{2N} + 2N^{-\frac{1}{5}}, \end{aligned}$$

which implies

$$\begin{aligned} P_n &> \exp\left(\frac{-n^2}{2N}\right)\exp\left(-2N^{-\frac{1}{5}}\right) \\ &= \left(1 + o(1)\right)\exp\left(\frac{-n^2}{2N}\right). \end{aligned}$$

We now return to the computation of $\mathrm{E}[\rho]$. From (4) we have

$$\mathrm{E}[\rho] = \sum_{n=0}^{\lfloor M \rfloor} P_n + \sum_{n=\lceil M \rceil}^{N-1} P_n + o(1) \tag{5}$$

where the error term comes from $nP_n < n\exp\left(\frac{-n^2}{2N}\right) = o(1)$ (we use $o(1)$ to denote any term whose absolute value tends to 0 as $N \to \infty$). The second sum is negligible, since

$$\begin{aligned} \sum_{n=\lceil M \rceil}^{N-1} P_n &< N\exp\Big(-\frac{M^2}{2N}\Big) \\ &= N\exp\Big(-\frac{1}{2}N^{-\frac{1}{5}}\Big) \\ &= o(1). \end{aligned} \tag{6}$$

For the first sum we have

$$\begin{aligned}\sum_{n=0}^{\lceil M\rceil} P_n &= \sum_{n=0}^{\lceil M\rceil}\Big(1+o(1)\Big)\exp\left(-\frac{n^2}{2N}\right)\\ &= \Big(1+o(1)\Big)\int_0^\infty e^{-\frac{x^2}{2N}}dx + O(1)\\ &= \Big(1+o(1)\Big)\sqrt{2N}\int_0^\infty e^{u^2}du + O(1)\\ &= \Big(1+o(1)\Big)\sqrt{2N}(\sqrt{\pi}/2)\\ &= \Big(1+o(1)\Big)\sqrt{\pi N/2}. \end{aligned} \tag{7}$$

Plugging (6) and (7) into (5) yields the desired result. □

Remark 9.4. One can similarly show $\mathrm{E}[\lambda] = \mathrm{E}[\sigma] = \frac{1}{2}\mathrm{E}[\rho] = \sqrt{\pi N/8}$, where $\sigma = \rho - \lambda$ is the length of the cycle.

In the baby-steps giant-steps algorithm (BSGS), if we assume that the discrete logarithm is uniformly distributed over $[1, N]$, then we should use $\sqrt{N/2}$ baby steps and expect to find the discrete logarithm after $\sqrt{N/2}$ giant steps, on average, using a total of $\sqrt{2N}$ group operations. But note that $\sqrt{\pi/2} \approx 1.25$ is less than $\sqrt{2} \approx 1.41$, so we may hope to compute discrete logarithms slightly faster than BSGS (on average) by simulating a random walk. Of course the worst-case running time for BSGS is better, since we will never need more than $\sqrt{2N}$ giant steps, but with a random walk the (very unlikely) worst case is N steps.

9.8 Pollard-ρ Algorithm

We now present the Pollard-ρ algorithm for computing $\log_\alpha \beta$, given $\beta \in \langle\alpha\rangle$; we should note that the assumption $\beta \in \langle\alpha\rangle$ which was not necessary in the baby-steps giant-steps algorithm is crucial here. As noted earlier, finding a collision in a random walk is useful to us only if we know how to express the colliding group elements as independent linear combinations of α and β. We thus extend the function $f\colon G \to G$ used to define our random walk to a function

$$f\colon \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z} \times G \to \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z} \times G,$$

which we require to have the property that if the input (a, b, γ) satisfies $a\alpha + b\beta = \gamma$, then $(a', b', \gamma') = f(a, b, \gamma)$ should satisfy $a'\alpha + b'\beta = \gamma'$.

There are several ways to define such a function f, one of which is the following. We first fix r distinct group elements $\delta_i = c_i\alpha + d_i\beta$ for some randomly chosen $c_i, d_i \in \mathbb{Z}/N\mathbb{Z}$. In order to simulate a random walk, we don't want r to be too small: empirically $r \approx 20$ works well [15]. We then define $f(a, b, \gamma) = (a + c_i, b + d_i, \gamma + \delta_i)$, where $i = h(\gamma)$ is determined by a randomly chosen *hash function*

$$h\colon G \to \{1, \ldots, r\}.$$

In practice we don't choose h randomly, we just need the preimages $h^{-1}(i)$ to partition G into r subsets of roughly equal size; for example, we might take the integer whose base-2 representation corresponds to the identifier $\mathsf{id}(\gamma) \in \{0,1\}^m$ and reduce it modulo r.[6]

[6]Note the importance of unique identifiers. We must be sure that γ is always hashed to the same value. Using a non-unique representation such as projective points on an elliptic curve will not achieve this.

To start our random walk, we pick random $a_0, b_0 \in \mathbb{Z}/N\mathbb{Z}$ and let $\gamma_0 = a_0\alpha + b_0\beta$. The walk defined by the iteration function f is known as an *r-adding* walk. Note that if $(a_{j+1}, b_{j+1}, \gamma_{j+1}) = f(a_j, b_j, \gamma_j)$, the value of γ_{j+1} depends only on γ_j, not on a_j or b_j, so the function f does define a walk in the same sense as before. We now give the algorithm.

Algorithm 9.5 (Pollard-ρ)**.** Given α, $N = |\alpha|$, $\beta \in \langle\alpha\rangle$, compute $\log_\alpha \beta$ as follows:

1. Compute $\delta_i = c_i\alpha + d_i\beta$ for $r \approx 20$ randomly chosen pairs $c_i, d_i \in \mathbb{Z}/N\mathbb{Z}$.
2. Compute $\gamma_0 = a_0\alpha + b_0\beta$ for randomly chosen $a_0, b_0 \in \mathbb{Z}/N\mathbb{Z}$.
3. Compute $(a_j, b_j, \gamma_j) = f(a_{j-1}, b_{j-1}, \gamma_{j-1})$ for $j = 1, 2, 3, \ldots$, until $\gamma_k = \gamma_j$ with $k > j$.
4. The collision $\gamma_k = \gamma_j$ implies $a_j\alpha + b_j\beta = a_k\alpha + b_k\beta$. Provided that $b_k - b_j$ is invertible in $\mathbb{Z}/N\mathbb{Z}$, we return $\log_\alpha \beta = \frac{a_j - a_k}{b_k - b_j} \in \mathbb{Z}/N\mathbb{Z}$; otherwise start over at step 1.

Note that if $N = |\alpha|$ is a large prime, it is extremely likely that $b_k - b_j$ will be invertible. In any case, by restarting we ensure that the algorithm terminates with probability 1, since it is certainly possible to have $\gamma_0 = x\alpha$ and $\gamma_1 = \beta$, where $x = \log_\alpha \beta$, for example. With this implementation the Pollard rho algorithm is a Las Vegas algorithm, even though it is often referred to in the literature as a Monte Carlo algorithm, due to the title of [8].

The description above does not specify how we should detect collisions. A simple method is to store all the γ_j as they are computed and look for a collision during each iteration. However, this implies a space complexity of ρ, which we expect to be on the order of $\sqrt{N}$. But we can use dramatically less space than this.

The key point is that once the walk enters a cycle, it will remain inside this cycle forever, and *every* step inside the cycle produces a collision. It is thus not necessary to detect a collision at the exact moment we enter the cycle, we can afford a slight delay. We now consider two space-efficient methods for doing this.

9.9 Floyd's cycle detection method

Floyd's cycle detection method [5, Ex. 3.1.6, p. 7] minimizes the space required: it keeps track of just two triples (a_j, b_j, γ_j) and (a_k, b_k, γ_k) that correspond to vertices of the walk (of course it also needs to store c_i, d_i, γ_i for $0 \le i < r$). The method is typically described in terms of a tortoise and a hare that are both traveling along the ρ-shaped walk. They start with the same γ_0, but in each iteration the hare takes two steps, while the tortoise takes just one. We thus modify step 3 of Algorithm 9.5 to compute

$$(a_j, b_j, \gamma_j) = f(a_{j-1}, b_{j-1}, \gamma_{j-1})$$
$$(a_k, b_k, \gamma_k) = f(f(a_{k-1}, b_{k-1}, \gamma_{k-1})).$$

The triple $(a_j, b_j\gamma_j)$ corresponds to the tortoise, and the triple (a_k, b_k, γ_k) corresponds to the hare. Once the tortoise enters the cycle, the hare (which must already be in the cycle) is guaranteed to collide with the tortoise within σ iterations, where σ is the length of the cycle (to see this, note that the hare gains on the tortoise by one step in each iteration and cannot pass the tortoise without landing on it). On average, we expect it to take $\sigma/2$ iterations for the hare to catch the tortoise and produce a collision, which we detect by testing whether $\gamma_j = \gamma_k$ after each iteration.

The expected number of iterations is thus $\mathrm{E}[\lambda + \sigma/2] = 3/4\,\mathrm{E}[\rho]$. But each iteration now requires three group operations, so the algorithm is actually slower by a factor of 9/4. Still, this achieves a time complexity of $O(\sqrt{N})$ group operations while storing just $O(1)$ group elements, which is a dramatic improvement.

9.10 The method of distinguished points

The "distinguished points" method (commonly attributed to Ron Rivest) uses slightly more space, say $O(\log^c N)$ group elements, for some constant c, but it detects cycles in essentially optimal time (within a factor of $1 + o(1)$ of the best possible), and uses just one group operation for each iteration, rather than the three required by Floyd's method.

The idea is to "distinguish" a certain subset of G by fixing a random boolean function $B\colon G \to \{0,1\}$ and calling the elements of $B^{-1}(1)$ *distinguished points*. We don't want the set of distinguished points to be too large, since we will store all the distinguished points we encounter during our walk, but we want our walk to contain many distinguished points; say $(\log N)^c$, on average, for some constant $c > 0$. This means we should choose B so that

$$\#B^{-1}(1) \approx \sqrt{N}(\log N)^c.$$

One way to define such a function B is to hash group elements to bit-strings of length k via a hash function $\tilde{h}\colon G \to \{0,1\}^k$, and then let $B(\gamma) = 1$ if and only if $\tilde{h}(\gamma)$ is the zero vector. If we set $k = \frac{1}{2}\log_2 N - c\log_2\log N$ then $B^{-1}(1)$ will have the desired cardinality. An easy and very efficient way to construct the hash function $\tilde{h}$ is to use the k least significant bits of the bit-string that uniquely represents the group element. For points on elliptic curves, we should use bits from the x-coordinate, since this will allow us to detect collisions of the form $\gamma_j = \pm\gamma_k$ (we can determine the sign by checking y-coordinates).

Algorithm 9.6 (Pollard-ρ using distinguished points)**.**

1. Pick random $c_i, d_i, a_0, b_0 \in \mathbb{Z}/N\mathbb{Z}$, compute $\delta_i = c_i\alpha + d_i\beta$ and $\gamma_0 = a_0\alpha + b_0\beta$, and initialize $D \leftarrow \emptyset$.
2. For $j = 1, 2, 3, \ldots$:
   a. Compute $(a_j, b_j, \gamma_j) = f(a_{j-1}, b_{j-1}, \gamma_{j-1})$.
   b. If $B(\gamma_j) = 1$ then
      i. If there exists $(a_k, b_k, \gamma_k) \in D$ with $\gamma_j = \gamma_k$ then return $\log_\alpha \beta = \frac{a_j - a_k}{b_k - b_j}$ if $\gcd(b_k - b_j, N) = 1$ and restart at step 1 otherwise.
      ii. If not, replace D by $D \cup \{(a_j, b_j, \gamma_j)\}$ and continue.

A key feature of the distinguished points method is that it is well-suited to a massively parallel implementation, which is critical for any large-scale discrete logarithm computation. Suppose we have many processors all running the same algorithm independently. If we have, say, $\sqrt{N}$ processors, then after just one step there is a good chance of a collision, and in general if we have m processors we expect to get a collision within $O(\sqrt{N}/m)$ steps. We can detect this collision as soon as the processors involved in the collision reach a distinguished point. However, the individual processors cannot realize this themselves, since they only know the distinguished points they have seen, not those seen by other processors. Whenever a processor encounters a distinguished point, it sends the corresponding triple to a central server that is responsible for detecting collisions. This scenario is also called a λ-*search*, since the collision typically occurs between paths with different starting points that then follow the same trajectory (forming the shape of the letter λ, rather than the letter ρ).

There is one important detail that must be addressed: if there are no distinguished points in the cycle then Algorithm 9.6 will never terminate!

The solution is to let the distinguished set S grow with time. We begin with $S = \tilde{h}^{-1}(\mathbf{0})$, where $\tilde{h}\colon G \to \{0,1\}^k$ with $k = \frac{1}{2}\log_2 N - c\log_2\log N$. Every $\sqrt{\pi N/2}$ iterations, we

decrease k by 1. This effectively doubles the number of distinguished points, and when k reaches zero we consider every point to be distinguished. This guarantees termination, and the expected space is still just $O(\log^c N)$ group elements.

9.11 Current ECDLP records

The current record for computing discrete logarithms on elliptic curves over finite fields involves a cyclic group with 117-bit prime order on an elliptic curve $E/\mathbb{F}_q$ with $q = 2^{127}$ and was set in 2016. The computation was run on 576 XC6SLX150 FPGAs and took about 200 days [1]. The current record for elliptic curves over prime fields was set in 2017 using the curve $E : y^2 = x^3 + 3$ over the 114-bit prime field $\mathbb{F}_{11957518425389075254535992784167879}$ with $\#E(\mathbb{F}_p)$ prime. This computation took advantage of the extra automorphisms of this curve and took the equivalent of 81 days running on 2000 Intel cores [4]. The record for elliptic curves over prime fields without extra automorphisms was set in 2009 using a 112-bit prime order group on an elliptic curve $E/\mathbb{F}_p$ with $p = (2^{128} - 3)/(11 \cdot 6949)$; this computation was run on a cluster of 200 PlayStation 3 consoles and took 180 days [3]. All of these records were set using a parallel Pollard-rho search and the method of distinguished points.

We should note that for elliptic curves over non-prime fields the non-generic methods we will discuss in the next lecture (index calculus) can be applied. This changes the situation dramatically, and it is now practical to solve the discrete logarithm problem on elliptic curves over $\mathbb{F}_q$ for suitably composite q with thousands of bits. But for elliptic curves over prime fields we know of no methods other than generic algorithms.

This claim holds even for quantum computers: there are very efficient algorithms for solving the discrete logarithm problem on an elliptic curve over a prime field, but these algorithms are generic in the sense that they apply to any group for which the group operation can be effectively implemented on a quantum computer using unique representations of group elements, an assumption that is already implicit in our black box model.

9.12 Computing discrete logarithms via the hidden subgroup problem

While we won't discuss quantum computing in this course (take 18.435J), let us briefly describe an efficient generic algorithm for solving the discrete logarithm on a quantum computer. As first proposed by Peter Shor [9] for computing discrete logarithms in $\mathbb{F}_p^\times$ and then generalized by others, this involves a reduction to what is now known as the *hidden subgroup problem* (HSP). We are given a finite group G containing a subgroup H along with a function $f\colon G \to S$ that is constant on cosets of H and maps each coset to a distinct element of S; here S can be any finite set, but for us S will actually be the group we want to compute a discrete logarithm in.

The hidden subgroup problem is to compute a set of generators for the unknown group H using f and group operations in G. There is an efficient polynomial-time algorithm to solve this problem on a quantum computer when H is abelian[7] assuming the group operation in G can be efficiently implemented on a quantum computer.[8] We won't describe the quantum algorithm for solving the hidden subgroup problem here, our aim is simply to show how it can be used to easily solve the discrete logarithm problem.

[7] The hidden subgroup problem for non-abelian groups is still open; even for dihedral groups we do not have a quantum polynomial-time algorithm.

[8] One can encapsulate this assumption by postulating a "quantum black box" that is used by "quantum generic group algorithms", just as we did for classical generic group algorithms above.

To compute the discrete logarithm problem of $\beta = \alpha^x$ in the cyclic group $\langle\alpha\rangle$ of order N one defines G, H, S, and f as follows:

$$G \coloneqq \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}, \qquad H \coloneqq \langle(x,1)\rangle, \qquad S \coloneqq \langle\alpha\rangle, \qquad \begin{aligned} f &\colon G \to S \\ (a,b) &\mapsto b\beta - a\alpha \end{aligned}$$

The computation of f only requires the inputs α, β and operations in the group $\langle\alpha\rangle$, it does not require knowledge of H or the discrete logarithm x we are tying to compute. We can use the standard double-and-add algorithm to compute f using $O(n)$ group operations. Given any set of generators for H we can easily recover x. All we need is an element $(a,b) \in H$ with $b \perp N$, since $x = ab^{-1} \bmod N$; if N is prime any nonzero element of H will do, and in general we can easily construct such an element as a linear combination of whatever set of generators our quantum computer gives us.

9.13 A generic lower bound for the discrete logarithm problem

We will now prove an essentially tight lower bound for solving the discrete logarithm problem with a generic group algorithm. We will show that if p is the largest prime divisor of N, then any generic group algorithm for the discrete logarithm problem must use $\Omega(\sqrt{p})$ group operations. In the case that the group order $N = p$ is prime this bound is tight, since we have already seen that the problem can be solved with $O(\sqrt{N})$ group operations using the baby-steps giant-steps method, and the Pohlig-Hellman complexity bound $O(n \log n + n\sqrt{p})$ shows that it is tight in general, up to logarithmic factors.

This lower bound applies not only to deterministic algorithms, but also to randomized algorithms: a generic Monte Carlo algorithm for the discrete logarithm problem must use $\Omega(\sqrt{p})$ group operations in order to be correct with probability bounded above $1/2$, and the expected running time of any generic Las Vegas algorithm is $\Omega(\sqrt{p})$ group operations.

The following theorem due to Shoup [10] generalizes an earlier result of Nechaev [7]. Our presentation here differs slightly from Shoup's and gives a sharper bound, but the proof is essentially the same. Recall that in our generic group model, each group element is uniquely represented as a bit-string via an injective map $\mathsf{id}\colon G \hookrightarrow \{0,1\}^m$, where $m = O(\log|G|)$.

Theorem 9.7 (Shoup)**.** *Let $G = \langle\alpha\rangle$ be a group of order N. Let $\mathcal{B}$ be a black box for G supporting the operations* `identity`, `inverse`, *and* `compose`, *using a random identification map $\mathsf{id}\colon G \hookrightarrow \{0,1\}^m$. Let $\mathcal{A}\colon \{0,1\}^m \times \{0,1\}^m \to \mathbb{Z}/N\mathbb{Z}$ be a randomized generic group algorithm that makes at most $s - 4\lceil\lg N\rceil$ calls to $\mathcal{B}$, for some integer s, and let x denote a random element of $\mathbb{Z}/N\mathbb{Z}$. Then*

$$\Pr_{x,\mathsf{id},\tau}[\mathcal{A}(\mathsf{id}(\alpha), \mathsf{id}(x\alpha)) = x] < \frac{s^2}{2p},$$

where τ denotes the random coin-flips made by $\mathcal{A}$ and p is the largest prime factor of N.

Note that $\mathcal{A}$ can generate random elements of G by computing $z\alpha$ for random $z \in \mathbb{Z}/N\mathbb{Z}$ (using at most $2\lg N$ group operations). We assume that $\mathcal{A}$ is given the group order N (this only makes the theorem stronger). The theorem includes deterministic algorithms as a special case where $\mathcal{A}$ does not use any of the random bits in τ. Bounding the number of calls $\mathcal{A}$ makes to $\mathcal{B}$ might appear to make the theorem inapplicable to Las Vegas algorithms, but we can convert a Las Vegas algorithm to a Monte Carlo algorithm by forcing it to halt and generate a random output if it exceeds its expected running time by some constant factor.

In order to simplify the presentation we will only prove Theorem 9.7 in the case $N = p$ is prime; the proof for composite N is an easy generalization of the prime order case, which in some sense the only case that matters given our $O(n \log n + n\sqrt{p})$ upper bound ($n = \log N$).

Proof of Theorem 9.7 for $N = p$ prime. To simplify the proof, we will replace $\mathcal{A}$ by an algorithm $\mathcal{A}'$ that does the following:

1. Use $\mathcal{B}$ to compute $\mathsf{id}(N\alpha) = \mathsf{id}(0)$.
2. Simulate $\mathcal{A}$, using $\mathsf{id}(0)$ to replace `identity` operations, to get $y = \mathcal{A}(\mathsf{id}(\alpha), \mathsf{id}(x\alpha))$.
3. Use $\mathcal{B}$ to compute $\mathsf{id}(y\alpha)$.

In the description above we assume that the inputs to $\mathcal{A}$ are $\mathsf{id}(\alpha)$ and $\mathsf{id}(x\alpha)$; the behavior of $\mathcal{A}'$ when this is not the case is irrelevant. Note that steps 1 and 3 each require at most $2\lceil \log_2 N \rceil - 1$ calls to $\mathcal{B}$ using double-and-add, so $\mathcal{A}'$ makes at most $s - 2$ calls to $\mathcal{B}$.

Let $\gamma_1 = \mathsf{id}(\alpha)$ and $\gamma_2 = \mathsf{id}(x\alpha)$. Without loss of generality we may assume that every interaction between $\mathcal{A}'$ and $\mathcal{B}$ is of the form $\gamma_k = \gamma_i \pm \gamma_j$, with $1 \le i, j < k$, where γ_i and γ_j are identifiers of group elements that are either inputs or values previously returned by $\mathcal{B}$ (here the notation $\gamma_i \pm \gamma_j$ means that $\mathcal{A}'$ is using $\mathcal{B}$ to add or subtract the group elements identified by γ_i and γ_j). Note that $\mathcal{A}'$ can invert γ_j by computing $\mathsf{id}(0) - \gamma_j$.

The number of such interactions is clearly a lower bound on the number of calls made by $\mathcal{A}'$ to $\mathcal{B}$. To further simplify matters, we will assume that the execution of $\mathcal{A}'$ is padded with operations of the form $\gamma_k = \gamma_1 + \gamma_1$ as required until k reaches s.

For $k = 1, \ldots, s$ define $F_k = a_k X + b_k \in \mathbb{Z}/p\mathbb{Z}[X]$ via:

$$F_1 := 1, \qquad F_2 := X, \qquad F_k := F_i \pm F_j \qquad (2 < k \le s).$$

Each F_k is a linear polynomial in X that satisfies

$$F_k(x) \equiv \log_{\gamma_1} \gamma_k \bmod p,$$

where we are abusing notation by writing $\gamma_k = \mathsf{id}(g_k)$ in place of $g_k \in G$.

Let us now consider the following game, which models the execution of $\mathcal{A}'$. At the start of the game we set $F_1 = 1$, $F_2 = X$, and γ_1 and γ_2 to distinct random values in $\{0,1\}^m$. For rounds $k = 2, 3, \ldots, s$, the algorithm $\mathcal{A}'$ and the black box $\mathcal{B}$ play the game as follows:

1. $\mathcal{A}'$ chooses a pair of integers i and j, with $1 \le i, j < k$, and a sign $\pm$ that determines $F_k = F_i \pm F_j$, and then asks $\mathcal{B}$ for the value of $\gamma_k = \gamma_i \pm \gamma_j$.
2. $\mathcal{B}$ sets $\gamma_k = \gamma_{k'}$ if $F_k = F_{k'}$ for some $k' < k$, and otherwise $\mathcal{B}$ sets γ_k to a random bit-string in $\{0,1\}^m$ that is distinct from $\gamma_{k'}$ for all $k' < k$.

After the sth round we pick $t \in \mathbb{Z}/p\mathbb{Z}$ at random and say that $\mathcal{A}'$ wins if $F_i(t) = F_j(t)$ for any $F_i \ne F_j$; otherwise $\mathcal{B}$ wins. Notice that the group G also plays no role in the game, it just involves bit-strings, but the constraints on $\mathcal{B}$'s choice of γ_k ensure that the bit strings $\gamma_1, \ldots, \gamma_s$ can be assigned to group elements in a consistent way. We now claim that

$$\Pr_{x,\mathsf{id},\tau}[\mathcal{A}(\mathsf{id}(\alpha), \mathsf{id}(x\alpha)) = x] \le \Pr_{t,\mathsf{id},\tau}[\mathcal{A}' \text{ wins the game}], \tag{8}$$

where the id function on the right represents an injective map $G \hookrightarrow \{0,1\}^m$ that is compatible with the choices made by $\mathcal{B}$ during the game, in other words, there exists a sequence of group elements $\alpha = \alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_s$ such that $\mathsf{id}(\alpha_i) = \gamma_i$ and $\alpha_k = \alpha_i \pm \alpha_j$, where i, j, and the sign $\pm$ correspond to the values chosen by $\mathcal{A}'$ in the kth round.

Any triple (x, id, τ) for which $\mathcal{A}(\mathsf{id}(\alpha), \mathsf{id}(x\alpha)) = x$ is also a triple (t, id, τ) for which $\mathcal{A}'$ wins the game; here we use the fact that $\mathcal{A}'$ always computes $y\alpha$, where $y = \mathcal{A}(\mathsf{id}(\alpha), \mathsf{id}(x\alpha))$, so $\mathcal{A}'$ forces a collision to occur whenever $\mathcal{A}$ outputs the correct value of x even if $\mathcal{A}$ did not actually encounter a collision (maybe $\mathcal{A}$ just made a lucky guess). Thus (8) holds.

We now bound the probability that $\mathcal{A}'$ wins the game. Consider any particular execution of the game, and let $F_{ij} = F_i - F_j$. We claim that for all i and j such that $F_{ij} \neq 0$,

$$\Pr_t[F_{ij}(t) = 0] \leq \frac{1}{p}. \tag{9}$$

We have $F_{ij}(X) = aX + b$ for some $a, b \in \mathbb{Z}/p\mathbb{Z}$ with a and b not both zero. If a is zero then $F_{ij}(t) = b \neq 0$ for all $t \in \mathbb{Z}/p\mathbb{Z}$ and (9) holds. Otherwise the map $[a]\colon t \mapsto at$ is a bijection, and in either case there is at most one value of t for which $F_{ij}(t) = 0$, which proves (9).

If $\mathcal{A}'$ wins the game then there must exist an $F_{ij} \neq 0$ for which $F_{ij}(t) = 0$. Furthermore, since $F_{ij}(t) = 0$ if and only if $F_{ji}(t) = 0$, we may assume $i < j$. Thus

$$\begin{aligned}
\Pr_{t,\mathsf{id},\tau}[\mathcal{A}' \text{ wins the game}] &\leq \Pr_{t,\mathsf{id},\tau}[F_{ij}(t) = 0 \text{ for some } F_{ij} \neq 0 \text{ with } i < j] \\
&\leq \sum_{i<j, F_{ij}\neq 0} \Pr_t[F_{ij}(t) = 0] \\
&\leq \binom{s}{2}\frac{1}{p} < \frac{s^2}{2p},
\end{aligned}$$

where we have used the union bound ($\Pr[A \cup B] \leq \Pr(A) + \Pr(B)$) to obtain the sum. □

Corollary 9.8. *Let G be a cyclic group of prime order N. Every deterministic generic algorithm for the discrete logarithm problem in G uses at least $(\sqrt{2}+o(1))\sqrt{N}$ group operations.*

The baby-steps giant-steps algorithm uses $(2 + o(1))\sqrt{N}$ group operations in the worst case, so this lower bound is tight up to a constant factor, but there is a slight gap. In fact, the baby-steps giant-steps method is not quite optimal; the constant factor 2 in the upper bound $(2 + o(1))\sqrt{N}$ can be improved via [2] (but this still leaves a small gap).

Let us now extend Theorem 9.7 to the case where the black box also supports the generation of random group elements for a cost of one group operation. We first note that having the algorithm generate random elements itself by computing $z\alpha$ for random $z \in \mathbb{Z}/N\mathbb{Z}$ does not change the lower bound significantly if only a small number of random elements are used; this applies to all of the algorithms we have considered.

Corollary 9.9. *Let G be a cyclic group of prime order N. Every generic Monte Carlo algorithm for the discrete logarithm problem in G that uses $o(\sqrt{N}/\log N)$ random group elements uses at least $(1 + o(1))\sqrt{N}$ group operations.*

This follows immediately from Theorem 9.7, since a Monte Carlo algorithm is required to succeed with probability bounded above $1/2$. In the Pollard-ρ algorithm, assuming it behaves like a truly random walk, the number of steps required before the probability of a collision exceeds $1/2$ is $\sqrt{2\log 2} \approx 1.1774$, so there is again only a small gap in the constant factor between the lower bound and the upper bound.

In the case of a Las Vegas algorithm, we can obtain a lower bound by supposing that the algorithm terminates as soon as it finds a non-trivial collision (in the proof, this corresponds to a nonzero F_{ij} with $F_{ij}(t) = 0$). Ignoring the $O(\log N)$ additive term, this occurs within m

steps with probability at most $m^2/(2p)$. Summing over m from 1 to $\sqrt{2p}$ and supposing that the algorithm terminates in exactly m steps with probability $(m^2-(m-1)^2)/(2p)$, the expected number of steps is $2\sqrt{2p}/3 + o(\sqrt{p})$.

Corollary 9.10. *Let G be a cyclic group of prime order N. Every generic Las Vegas algorithm for the discrete logarithm problem in G that generates an expected $o(\sqrt{N}/\log N)$ random group elements uses at least $(2\sqrt{2}/3 + o(1))\sqrt{N}$ expected group operations.*

Now let us consider a generic algorithm that generates a large number of random elements, say $R = N^{1/3+\delta}$ for some $\delta > 0$. The cost of computing $z\alpha$ for R random values of z can be bounded by $2R + O(N^{1/3})$. If we let $e = \lceil \lg N/3 \rceil$ and precompute $c\alpha$, $c2^e\alpha$, and $c2^{2e}\alpha$ for $c \in [1, 2^e]$, we can then compute $z\alpha$ for any $z \in [1, N]$ using just 2 group operations. We thus obtain the following corollary, which applies to every generic group algorithm for the discrete logarithm problem.

Corollary 9.11. *Let G be a cyclic group of prime order N. Every generic Las Vegas algorithm for the discrete logarithm problem in G uses an expected $\Omega(\sqrt{N})$ group operations.*

In fact, we can be more precise: the implied constant factor is at least $\sqrt{2}/2$.

References

[1] Daniel J. Bernstein, Susanne Engles, Tanja Lange, Ruben Niederhagen, Christof Paar, Peter Schwabe, and Ralf Zimmermann, *Faster elliptic curve discrete logarithms on FPGAs*, Cryptology eprint Archive, Report 2016/382, 2016.

[2] Daniel J. Bernstein and Tanja Lange, *Two giants and a grumpy baby*, in Proceedings of the Tenth Algorithmic Number Theory Symposium (ANTS X), Open Book Series **1**, Mathematical Sciences Publishers, 2013, 87–111.

[3] Joppe W. Bos, Marcelo E. Kaihara, Thorsten Kleinjung, Arjen K. Lenstra, and Peter L. Montgomery, *PlayStation 3 computing breaks 2^{60} barrier, 112-bit ECDLP solved*, EPFL Laboratory for Cryptologic Algorithms, 2009.

[4] Takuya Kusaka, Sho Joichi, Ken Ikuta, Md. Aal-Amin Khandaker, Yasuyuki Nogami, Satoshi Uehara, Nariyoshi Yamai, *Solving 114-bit ECDLP for a Barreto–Naehrig curve*, Information Security and Cryptology – ICISC 2017, LNCS **10779** (2018), 231–244.

[5] Donald E. Knuth, *The Art of Computer Programming, vol. II: Seminumerical Algorithms*, third edition, Addison-Wesley, 1998.

[6] Stephen C. Pohlig and Martin E. Hellman, *An improved algorithm for computing logarithms over GF(p) and its cryptographic significance*, IEEE Transactions on Information Theory **24** (1978), 106–110.

[7] V.I. Nechaev, *Complexity of a determinate algorithm for the discrete logarithm*, Mathematical Notes **55** (1994), 165–172.

[8] J.M. Pollard, *Monte Carlo methods for index computation (mod p)*, Mathematics of Computation **143** (1978), 918–924.

[9] Peter Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM J. Computing **26** (1997), 1484–1509.

[10] Victor Shoup, *Lower bounds for discrete logarithms and related problems*, Proceedings of Eurocrypt '97, LNCS **1233** (1997), 256–266, revised version available at `https://www.shoup.net/papers/dlbounds1.pdf`.

[11] Victor Shoup, *A Computational Introduction to Number Theory and Algebra*, Cambridge University Press, 2005.

[12] Andrew V. Sutherland, *Order computations in generic groups*, PhD thesis, Massachusetts Institute of Technology, 2007.

[13] Andrew V. Sutherland, *Structure computation and discrete logarithms in finite abelian p-groups*, Mathematics of Computation **80** (2011), 501–538.

[14] David C. Terr, *A modification of Shanks baby-step giant-step method*, Mathematics of Computation **69** (2000), 767–773.

[15] Edlyn Teske, *On random walks for Pollard's rho method*, Mathematics of Computation **70** (2001), 809–825.

10 Index calculus, smooth numbers, and factoring integers

Having explored generic algorithms for the discrete logarithm problem in some detail, we now consider a non-generic algorithm based on *index calculus*.[1] This algorithm depends critically on the distribution of *smooth numbers* (integers with small prime factors), which naturally leads to a discussion of two algorithms for factoring integers that also depend on smooth numbers: the Pollard $p-1$ method and the elliptic curve method (ECM).

10.1 Index calculus

Index calculus is a method for computing discrete logarithms in the multiplicative group of a finite field. This might not seem directly relevant to the elliptic curve discrete logarithm problem, but as we shall see when we discuss pairing-based cryptography, these two problems are not completely unrelated. Moreover, index calculus based methods can be applied to the discrete logarithm problem on elliptic curves over non-prime finite fields, as well as abelian varieties of higher dimension (even over prime fields); see [8, 9, 10].[2]

We will restrict our attention to the simplest case, a finite field of prime order $\mathbb{F}_p \simeq \mathbb{Z}/p\mathbb{Z}$, and let us fix the set of integers in $[0, N]$ with $N = p-1$ as a set of coset representatives for $\mathbb{Z}/p\mathbb{Z}$. Index calculus exploits the fact that we "lift" elements of $\mathbb{Z}/p\mathbb{Z}$ to their representatives in $[0, N] \cap \mathbb{Z}$.

$$\mathbb{Z} \overset{\longrightarrow}{\longleftarrow} \mathbb{Z}/p\mathbb{Z} \simeq \mathbb{F}_p$$

The map $\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ is the canonical quotient map given by reduction modulo p, and it is a ring homomorphism. The "lifting" map from $\mathbb{Z}/p\mathbb{Z}$ to $\mathbb{Z}$ is a section of the quotient map, which is an injective map of sets but is not a ring homomorphism.[3] Nevertheless, if we lift elements from $\mathbb{Z}/p\mathbb{Z}$ to $\mathbb{Z}$, perform a sequence of ring operations in $\mathbb{Z}$, and then reduce modulo p, we will get the same result as if we had performed the entire sequence of ring operations in $\mathbb{Z}/p\mathbb{Z} \simeq \mathbb{F}_p$. A key feature of working in $\mathbb{Z}$ is that we can uniquely factor integers in $[1, N]$ into prime powers, something that makes no sense in the field $\mathbb{Z}/p\mathbb{Z}$ where every nonzero element is a unit and there are no nontrivial prime ideals.

Let us fix a generator α for $(\mathbb{Z}/p\mathbb{Z})^\times$, and let $\beta \in \langle\alpha\rangle$ be the element whose discrete logarithm we wish to compute. For any integer e, we may consider the prime factorization of the integer $\alpha^e\beta^{-1} \in [1, N] \subseteq \mathbb{Z}$; here we are implicitly lifting $\alpha^e\beta^{-1} \in \mathbb{Z}/p\mathbb{Z}$ to its unique coset representative in $[1, N]$, as we will continue to do without further comment. When $e = \log_\alpha \beta$ this prime factorization will be trivial, but in general we will have

$$\prod p_i^{e_i} = \alpha^e\beta^{-1},$$

where the p_i vary over primes and the exponents e_i are nonnegative integers. Multiplying both sides by β and taking discrete logarithms with respect to α yields

$$\sum e_i \log_\alpha p_i + \log_\alpha \beta = e,$$

[1] If α is a generator for $\mathbb{F}_p^\times$ then the discrete logarithm of $\beta \in \mathbb{F}_p^\times$ with respect to α is also called the *index* of β (with respect to α), whence the term *index calculus*.

[2] The two are related: if E is an elliptic curve over a finite field $\mathbb{F}_{q^n}$ for some prime-power q, there is an associated abelian variety of dimension n over $\mathbb{F}_q$ known as the *Weil restriction* of E.

[3] Indeed, there are no homomorphisms from rings of positive characteristic to rings of characteristic zero (note that the zero ring has positive characteristic).

which determines $\log_\alpha \beta$ as a linear expression in the discrete logarithms $\log_\alpha p_i$, where $\log_\alpha p_i$ denotes the discrete logarithm of the image of p_i under the quotient map $\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$. This doesn't immediately help us, since we don't know the values of $\log_\alpha p_i$. However, if we repeat this procedure using many different values of e, we may obtain a system of linear equations that we can try to solve for $\log_\alpha \beta$.

In order to make this feasible, we need to restrict the primes p_i to lie in a reasonably small set. We thus fix a *smoothness bound*, say B, and define the *factor base*

$$P_B = \{p : p \le B \text{ is prime}\} = \{p_1, p_2, \dots, p_b\},$$

where $b = \pi(B)$ is the number of primes up to B (of which there are approximately $B/\log B$). Not all choices of e will yield an integer $\alpha^e \beta^{-1} \in [1, N] \subseteq \mathbb{Z}$ that we can factor over our factor base P_B, in fact most will not. But some choices will work, and for those that do we obtain a linear equation of the form

$$e_1 \log_\alpha p_1 + e_2 \log_\alpha p_2 + \cdots + e_b \log_\alpha p_b + \log_\alpha \beta = e,$$

in which at most $\lfloor \lg N \rfloor$ of the e_i are nonzero. We may not know any of the discrete logarithms that appear in this relation, but we can view

$$e_1 x_1 + e_2 x_2 + \cdots + e_b x_b + x_{b+1} = e$$

as a linear equation in $b+1$ variables $x_1, x_2, \dots, x_{b+1}$ over the ring $\mathbb{Z}/N\mathbb{Z}$. This equation has a solution, namely, $x_i = \log_\alpha p_i$, for $1 \le i \le b$, and $x_{b+1} = \log_\alpha \beta$. If we collect $b+1$ such equations by choosing random values of e and discarding those for which $\alpha^e \beta^{-1}$ is not B-smooth, the resulting linear system may determine a unique value x_{b+1}, the discrete logarithm we wish to compute.

This system will typically be under-determined; indeed, many of the variables x_i may not appear in any of our relations. But it is quite likely that the value of x_{b+1}, which is present in every equation, will be uniquely determined. We will not attempt to prove this (to give a rigorous proof one really needs more than $b+1$ equations, say, $b \log b$), but it is empirically true.[4] This suggests the following algorithm to compute $\log_\alpha \beta$.

Algorithm 10.1 (Index calculus)**.** Given $\beta \in \langle \alpha \rangle = (\mathbb{Z}/p\mathbb{Z})^\times$, compute $\log_\alpha \beta$ as follows:

1. Pick a smoothness bound B, compute the factor base $P_B \coloneqq \{p_1, \dots, p_b\}$ with $b \coloneqq \pi(B)$, and let $N \coloneqq p - 1$.
2. Generate $b+1$ random relations $R_i = (e_{i,1}, e_{i,2}, \dots, e_{i,b}, 1, e_i)$ by picking $e \in [1, N]$ at random and attempting to factor $\alpha^e \beta^{-1} \in [1, N]$ over the factor base P_B. Each successful factorization yields a relation R_i with $e_i = e$ and $\alpha^{e_i} \beta^{-1} = \prod p_j^{e_{i,j}}$.
3. Attempt to solve the system defined by the relations $R_1, \dots, R_{b+1}$ for $x_{b+1} \in \mathbb{Z}/N\mathbb{Z}$ using linear algebra (row reduce the corresponding matrix).
4. If $x_{b+1} = \log_\alpha \beta$ is uniquely determined, return this value, otherwise go to step 2.

It remains to determine the choice of B in step 1, but let us first make the following remarks.

Remark 10.2. It is not actually necessary to start over from scratch when x_{b+1} is not uniquely determined, typically adding just a few more relations will be enough.

[4] When considering potential attacks on a cryptographic system, one should always be willing to make any reasonable heuristic assumption that helps the attacker.

Remark 10.3. As noted above, the relations $R_1, \ldots, R_{b+1}$ will be very sparse (at most $\lfloor \lg N \rfloor + 1$ nonzero coefficients in each), which can speed up the linear algebra step significantly.

Remark 10.4. While solving the system $R_1, \ldots, R_{b+1}$ we are likely to encounter non-invertible elements of $\mathbb{Z}/N\mathbb{Z}$ (for example, 2 is never invertible, since $N = p - 1$ is even). Whenever this happens we can use a GCD computation to obtain a non-trivial factorization $N = N_1 N_2$ with N_1 and N_2 relatively prime. We then proceed to work in $\mathbb{Z}/N_1\mathbb{Z} \times \mathbb{Z}/N_2\mathbb{Z}$, using the CRT to recover the value of x_{b+1} in $\mathbb{Z}/N\mathbb{Z}$ (recurse as necessary).

Remark 10.5. Solving the system of relations will determine not only $x_{b+1} = \log_\alpha \beta$, but also many $x_i = \log_\alpha p_i$ for $p_i \in P_B$, which do not depend on β. If we are computing discrete logarithms for many different β with respect to the same base α, after the first computation the number of relations we need is just one more than the number of $x_i = \log_\alpha p_i$ that have yet to be determined. If we are computing discrete logarithms for $\Omega(b)$ values of β, we expect to compute just $O(1)$ relations per discrete logarithm, on average.

An integer whose prime factors are all bounded by B is said to be *B-smooth*. A large value of B will make it more likely that $\alpha^e \beta^{-1}$ is B-smooth, but it also makes it more difficult to determine whether this is in fact the case, since we need to determine all he prime factors of $\alpha^e \beta^{-1}$ up to B. We want to balance the cost of smoothness testing against the number of smoothness tests we expect to need in order to get $b + 1$ relations (note that b depends on B). Let us suppose for the moment that the cost of the linear algebra step is negligible by comparison (which turns out to be the case, at least in terms of time complexity). If we choose $e \in [1, N]$ uniformly at random then α^e, and therefore $\alpha^e \beta^{-1}$, will be uniformly distributed over $(\mathbb{Z}/p\mathbb{Z})^\times$, uniquely represented by the set of integers in $[1, N]$. To determine the optimal value of B, we need to know the probability that a random integer in $[1, N]$ is B-smooth.

10.2 The Canfield-Erdös-Pomerance Theorem

For positive real numbers x and y, let $\psi(x, y)$ count the y-smooth integers in $[1, x]$. The probability that a random integer $m \in [1, x]$ is y-smooth is then approximately $\frac{1}{x}\psi(x, y)$. We want our smoothness bound y to vary as a function of x, so it is standard to define

$$u := \frac{\log x}{\log y}$$

and replace y by $x^{1/u}$.

Theorem 10.6 (Canfield-Erdős-Pomerance)**.** *The asymptotic bound*

$$\frac{1}{x}\psi(x, x^{1/u}) = u^{-u+o(u)}$$

holds uniformly as $u, x \to \infty$, provided that $u < (1 - \epsilon) \log x / \log \log x$ for some $\epsilon > 0$.

For a proof on this result along with many other interesting facts about smooth numbers, we recommend the survey article by Granville [13].

10.3 Optimizing the smoothness bound

Let us assume that generating relations in step 2 dominates the overall complexity of Algorithm 10.1, and for the moment suppose that we simply use trial-division to attempt to factor $\alpha^e\beta^{-1}$ over P_B (we will see a more efficient method for smoothness testing shortly). The expected running time of Algorithm 10.1 is then approximately

$$(b+1)\cdot u^u \cdot b \cdot \mathsf{M}(\log N), \tag{1}$$

where $u = \log N/\log B$. The four factors in (1) are:

- $b+1$: the number of relations R_i that we need;
- u^u: the expected number of random exponents e we need to try in order to obtain a B-smooth integer $m \coloneqq \alpha^e\beta^{-1} \in [1, N]$;
- b: the number of trial divisions to test whether m is B-smooth (and factor it if it is);
- $\mathsf{M}(\log N)$: the time for each trial division.

We have $b = \pi(B) \sim B/\log B$, and if we ignore logarithmic factors we can replace both $b+1$ and b by B and drop the $\mathsf{M}(\log N)$ factor. We wish to choose u to minimize the quantity

$$B^2u^u = N^{2/u}u^u, \tag{2}$$

where we have used $B^u = N$ to eliminate B. Taking logarithms, it suffices to minimize

$$f(u) = \log(N^{2/u}u^u) = \frac{2}{u}\log N + u\log u,$$

so we want to consider solutions to

$$f'(u) = -\frac{2}{u^2}\log N + \log u + 1 = 0.$$

Ignoring the asymptotically negligible constant 1, we would like to pick u so that

$$u^2\log u \approx 2\log N.$$

For

$$u = 2\sqrt{\log N/\log\log N}, \tag{3}$$

we have

$$u^2\log u = \frac{4\log N}{\log\log N}\cdot\left(\log 2 + \frac{1}{2}(\log\log N - \log\log\log N)\right) = 2\log N + o(\log N),$$

as desired. The choice of u in (3) implies that we should use the smoothness bound

$$\begin{aligned} B = N^{1/u} &= \exp\left(\frac{1}{u}\log N\right)\\ &= \exp\left(\frac{1}{2}\sqrt{\log N\log\log N}\right)\\ &= L_N[1/2, 1/2]. \end{aligned}$$

Here we have used the asymptotic notation

$$L_N[\alpha, c] \coloneqq \exp\big((c + o(1))(\log N)^{\alpha}(\log\log N)^{1-\alpha}\big),$$

which is commonly used to denote complexity bounds of this form. Note that

$$L_N[0, c] = \exp((c + o(1)\log\log N) = (\log N)^{c+o(1)}$$

is polynomial in $\log N$, whereas

$$L_N[1, c] = \exp((c + o(1))\log N) = N^{c+o(1)}$$

is exponential in $\log N$. For $0 < \alpha < 1$ the bound $L_N[\alpha, c]$ is *subexponential* (in $\log N$).

We also have $u^u = \exp(u \log u) = L_N[1/2, 1]$, thus the total expected running time is

$$B^2 u^u = L_N[1/2, 1/2]^2 \cdot L_N[1/2, 1] = L_N[1/2, 2].$$

The cost of the linear algebra step is certainly no worse than $\widetilde{O}(b^3)$, which is $\widetilde{O}(B^3)$, In our subexponential notation this is $L_N[1/2, 3/2]$, which is dominated by the bound above, so our assumption that the cost of generating relations dominates the running time is justified. In fact, if we take advantage of the sparseness of the system noted in Remark 10.3, the cost of the linear algebra step can be bounded by $\widetilde{O}(b^2)$. However, in large computations the linear algebra step is often a limiting factor in practice because it is memory intensive and not as easy to parallelize as relation finding.

Remark 10.7. As noted earlier, if we are computing many (say at least $L_N[1/2, \sqrt{2}/2]$) discrete logarithms with respect to the same base α, we just need $O(1)$ relations per β, on average. In this case we should choose $B = N^{1/u}$ to minimize Bu^u rather than B^2u^u. This yields an average expected running time of $L_N[1/2, \sqrt{2}]$ per discrete logarithm.

A simple version of Algorithm 10.1 using trial-division for smoothness testing is implemented in this Sage notebook.

10.4 Improvements

Using the elliptic curve factorization method (ECM) described in the next section, the cost of testing and factoring B-smooth integers can be made subexponential in B and polynomial in $\log N$. This effectively changes B^2u^u in (2) to Bu^u, and the optimal smoothness bound becomes $B = L_N[1/2, 1/\sqrt{2}]$, yielding a heuristic expected running time of

$$L_N[1/2, \sqrt{2}].$$

There is a batch smoothness testing algorithm due to Bernstein [3] that for a sufficiently large set of integers yields an average time per integer that is actually polynomial in $\log N$, but this does not change the complexity in a way that is visible in our $L_N[\alpha, c]$ notation.

Using more advanced techniques, analogous to those used in the *number field sieve* for factoring integers, one can achieve a heuristic expected running time of the form

$$L_N[1/3, c]$$

for computing discrete logarithms in $\mathbb{F}_p^\times$ (again using an index calculus approach); see [12].

In finite fields of small characteristic $\mathbb{F}_{p^n} \simeq \mathbb{F}_p[x]/(f(x))$, one uses the *function field sieve*, where now the factor base consists of low degree polynomials in $\mathbb{F}_p[x]$ that represent elements of $\mathbb{F}_{p^n}$ when reduced modulo $f(x)$. This also yields an $L_N[1/3, c]$ bound (with a smaller value of c). Under heuristic assumptions, such a bound holds for all finite fields [16].

But this is far from the end of the story. In 2013 Antoine Joux announced an index calculus approach for finite fields of the form $\mathbb{F}_{q^k}$ with $q \approx k$ that heuristically achieves an $L_N[1/4 + o(1), c]$ time complexity [14]. Shortly thereafter a recursive variant of Joux's approach was used to obtain a heuristically quasi-polynomial-time complexity of $k^{O(\log k)}$, which in terms of $N = q^k$ is bounded by $L_N[\epsilon, c]$ for every $\epsilon, c > 0$. At first glance the assumption $q \approx k$ might seem restrictive, but even for finite fields of the form $\mathbb{F}_{2^k}$ with k prime it suffices to compute discrete logarithms in the extension field $\mathbb{F}_{2^{kr}}$ with $r = \lceil \lg k \rceil$, which for $q = 2^r \approx k$ has the desired form $\mathbb{F}_{q^k}$. Even though we are now working in a larger field, the $k^{O(\log k)}$ bound is still quasi-polynomial in the input size k, and as a function of $N = 2^k$ it is dominated by $L_N[\epsilon, c]$ for all $\epsilon, c > 0$, hence quasi-polynomial-time.

As of March 2021 the record for computing discrete logarithms in finite fields was set in the field $\mathbb{F}_{2^{30750}}$, using about 2900 core-years in 2019 [11]. The record for prime degree finite fields was set in 2014 in the field $\mathbb{F}_{2^{1279}}$, using less than 4 core years [15] (this record could surely be improved), and the record for "safe" prime fields $\mathbb{F}_p$ (where $(p-1)/2$ is prime), was set in 2019 for a 795-bit prime p using about 3100 core years [6].

The recent dramatic improvements in computing discrete logarithms in finite fields of small characteristic has effectively eliminated interest in pairing-based elliptic curve cryptography over such fields. As discussed in Lecture 1, in pairing-based cryptography one needs to consider the difficulty of the discrete logarithm problem both in the group of rational points on an elliptic curve over a finite field $\mathbb{F}_q$ and in the multiplicative group of a low degree extension of $\mathbb{F}_q$. None of these results have had any impact on the prospects of pairing-based cryptography over prime fields.[5]

10.5 The Pollard $p-1$ method

In 1974, Pollard introduced a Monte Carlo algorithm for factoring integers [19] that works astonishingly well when the integer $p-1$ is extremely smooth (but in the worst case is no better than trial division). The algorithm takes as input an integer N to be factored and a smoothness bound B.

Algorithm 10.8 (Pollard $p-1$ factorization)**.**
Input: An integer N to be factored and a smoothness bound B.
Output: A proper divisor of N or `failure`.

1. Pick a random integer $a \in [1, N-1]$.
2. If $d = \gcd(a, N)$ is not 1 then return d.
3. Set $b = a$ and for each prime $\ell \le B$:
   a. Set $b = b^{\ell^e} \bmod N$, where $\ell^{e-1} < N \le \ell^e$. If $b = 1$ then return `failure`.
   b. If $d = \gcd(b-1, N)$ is not 1 then return d.
4. Return `failure`

[5] Quantum computers are a potential threat, but this is a separate issue; the attacks based on Joux's breakthrough all use classical models of computation.

Rather than using a fixed bound B, we could simply let the algorithm keep running through primes ℓ until it either succeeds or fails in step 3b. But in practice one typically uses a very small smoothness bound B and switches to a different algorithm if the $p-1$ method fails. In any case, it is convenient to have B fixed for the purposes of analysis.

Example 10.9. Let $N = 899$ and suppose we pick $a = 2$ in step 1. Then $d = 1$ in step 2, and the table below illustrates the situation at the end of each iteration of step 3.

ℓ	e	b	d
2	10	605	1
3	7	690	1
5	5	683	31

The algorithm finds the factor 31 of $N = 29 \cdot 31$ when $\ell = 5$ because $\#(\mathbb{Z}/31)^\times = 30 = 2 \cdot 3 \cdot 5$ is 5-smooth but $\#(\mathbb{Z}/29)^\times = 28 = 2^2 \cdot 7$ is not: if we put $m = 2^{10} \cdot 3^7 \cdot 5^5$ then m is divisible by $\#(\mathbb{Z}/31\mathbb{Z})^\times$ but not by $\#(\mathbb{Z}/29\mathbb{Z})^\times$, and it follows that we always have $a^m \equiv 1 \bmod 31$, but for most choices of a we will have $a^m \not\equiv 1 \bmod 29$, leading to $d = \gcd(a^m - 1, 29 \cdot 31) = 31$.

If we had instead used $N = 31 \cdot 41$ we would have found $d = N$ when $\ell = 5$ and failed because $\#(\mathbb{Z}/41\mathbb{Z})^\times = 40 = 2^3 \cdot 5$ has the same largest prime factor as $\#(\mathbb{Z}/31\mathbb{Z})^\times$.

Theorem 10.10. *Let p and q be prime divisors of N, and let ℓ_p and ℓ_q be the largest prime divisors of $p-1$ and $q-1$, respectively. If $\ell_p \le B$ and $\ell_p < \ell_q$ then Algorithm 10.8 succeeds with probability at least $1 - \frac{1}{\ell_q}$.*

Proof. If $a \equiv 0 \bmod p$ then the algorithm succeeds in step 2, so we may assume $a \perp p$. When the algorithm reaches $\ell = \ell_p$ in step 3 we have $b = a^m$, where $m = \prod_{\ell \le \ell_p} \ell^e$ is a multiple of $p-1$. By Fermat's little theorem, $b = a^m \equiv 1 \bmod p$ and therefore p divides $b-1$. But ℓ_q does not divide m, so with probability at least $1 - \frac{1}{\ell_q}$ we have $b \not\equiv 1 \bmod q$, in which case $1 < \gcd(b-1, N) < N$ in step 3b and the algorithm succeeds. □

For almost all values of N, Algorithm 10.8 will succeed with very high probability given the smoothness bound $B = \sqrt{N}$. But if N is a prime power, or if the largest prime dividing $p-1$ is the same for every prime factor p of N it will still fail, no matter what value of a is chosen. In the best case, the algorithm can succeed very quickly. As demonstrated in this Sage notebook, if $N = p_1 p_2$ where p_1 and p_2 are 512-bit primes, if $p_1 - 1$ happens to be very smooth then Algorithm 10.8 can factor N within a few seconds; no other algorithm currently known can factor this integer N in a reasonable amount of time. However, in the worst-case the running time is $O(\pi(B)\,\mathsf{M}(\log N)\log N)$, and with $B = \sqrt{N}$ the complexity is $O(\sqrt{N}\,\mathsf{M}(\log N))$, the same as trial division (and as noted above, success is not guaranteed).

But rather than focusing on factoring a single integer N, let us consider a slightly different problem. Suppose we have a large set of composite integers (for example, a list of RSA moduli[6]), and our goal is to factor any one of them. How long would this take if we simply applied the $p-1$ method to each integer one-by-one?

For a given value of B, the expected time for the algorithm to achieve a success is

$$\frac{O(\pi(B)\,\mathsf{M}(\log N)\log N)}{\Pr[\texttt{success}]}. \tag{4}$$

[6] In fact, many RSA key generation algorithms incorporate specific measures to prevent the type of attack we consider here. In any case, current RSA keys are necessarily large enough (2048 bits) to be quite safe from the $L_N[1/2, \sqrt{2}]$ algorithm considered here.

Let p be a prime factor of N. The algorithm is very likely to succeed if $p-1$ is B-smooth, since it is very unlikely that all the other prime factors q of N have $q-1$ with exactly the same largest prime factor as $p-1$. Let us heuristically assume that integers of the form $p-1$ are at least as likely to be smooth as a random integer of similar size.

By the Canfield-Pomerance-Erdős Theorem, the probability that a random integer less than N is B-smooth is $u^{-u+o(u)}$, where $u = \log N/\log B$. If we ignore the $o(u)$ error term and factors that are polynomial in $\log N$ (which will be bounded by $o(u)$ in any case), we may simplify (4) to

$$N^{1/u}u^u. \tag{5}$$

This is minimized (up to asymptotically negligible factors) for $u = \sqrt{2\log N/\log\log N}$, thus we should use the smoothness bound

$$B = N^{1/u} = \exp\left(\left(1/\sqrt{2}+o(1)\right)\sqrt{\log N\log\log N}\right) = L_N[1/2, 1/\sqrt{2}],$$

where the $o(1)$ term incorporates the $o(u)$ error term and the factors polynomial in $\log N$ that we have ignored. We also have $u^u = \exp(u\log u) = L_N[1/2, 1/\sqrt{2}]$, and the total expected running time is therefore

$$N^{1/u}u^u = L_N[1/2, 1/\sqrt{2}]L_N[1/2, 1/\sqrt{2}] = L_N[1/2, \sqrt{2}].$$

Thus even though the $p-1$ method has an exponential worst-case running time, if we apply it to a sequence of random integers we achieve a (heuristically) subexponential running time. But this isn't much help if there is a particular integer N that we want to factor.

10.6 The elliptic curve method for factoring integers (ECM)

Using elliptic curves we can effectively achieve the randomized scenario envisioned above while keeping N fixed. The Pollard $p-1$ algorithm works in the group $(\mathbb{Z}/N\mathbb{Z})^\times$, but we can also think of it as performing simultaneous computations in the groups $(\mathbb{Z}/p\mathbb{Z})^\times$ for primes $p|N$; it succeeds when one of these groups has smooth order. If we instead take an elliptic curve $E/\mathbb{Q}$ defined by an integral equation $y^2 = x^3 + Ax + B$ that we can reduce modulo N, we have an opportunity to factor N whenever $E(\mathbb{F}_p)$ has smooth order, for some prime $p|N$. The key difference is that we can vary the curve E while keeping N fixed; we get a new group $E(\mathbb{F}_p)$ each time we change E. This is the basis of the elliptic curve method (ECM), introduced by Hendrik Lenstra [17] in the mid 1980s.

The algorithm is essentially the same as Pollard's $p-1$ method. Rather than exponentiating a random element of $(\mathbb{Z}/N\mathbb{Z})^\times$ to a large smooth power and hoping that it becomes the identity modulo some prime p dividing N, we instead multiply a random point on an elliptic curve by a large smooth scalar and hope that it becomes the identity modulo some prime p dividing N. If this doesn't happen we switch to a different curve and try again.

As in Pollard's $p-1$ algorithm, we don't know the primes p dividing N *a priori*, so we work modulo N and use GCD's to find a factor of N. If P is a point on $E/\mathbb{Q}$ and $mP = (Q_x : Q_y : Q_z)$ is a multiple of P that reduces to 0 modulo a prime p dividing N, then p divides $\gcd(Q_z, N)$. Notice that even though we are working with points on an elliptic curve over $\mathbb{Q}$, we only care about their reductions modulo primes dividing N, so we can keep the coordinates reduced modulo N throughout the algorithm.

In order to get a proper divisor of N we also need $\gcd(Q_z, N) \neq N$. This is very likely to be the case, so long as P is not a torsion point of $E(\mathbb{Q})$; if P is a torsion point

it will have the same order modulo every prime divisor of N and we will always have $\gcd(Q_z, N) = N$ whenever the gcd is non-trivial. Given an elliptic curve $E/\mathbb{Q}$, it is generally hard to find non-torsion points in $E(\mathbb{Q})$, in fact there may not be any.[7] Instead we pick integers $x_0, y_0, a \in [1, N-1]$ and let $b = y_0^2 - x_0^3 - ax_0$. This guarantees that $P = (x_0, y_0)$ is a rational point on the elliptic curve $E/\mathbb{Q}$ defined by $y^2 = x^3 + ax + b$. The probability that P is a torsion point is negligible.[8] We now give the algorithm, which takes not only an integer N and a smoothness bound B, but also a bound M on the largest prime factor of N that we seek to find (as discussed below, this is useful for smoothness testing).

Algorithm 10.11 (ECM)**.**
Input: An integer N to be factored, a smoothness bound B, and a prime bound M.
Output: A proper divisor of N or `failure`.

1. Pick random integers $a, x_0, y_0 \in [0, N-1]$ and set $b = y_0^2 - x_0^3 - ax_0$.
2. If $d = \gcd(4a^3 + 27b^2, N)$ is not 1 then return d if $d < N$ or `failure` if $d = N$.
3. Let $Q = P = (x_0 : y_0 : 1)$.
4. For all primes $\ell < B$:
   a. Set $Q = \ell^e Q \bmod N$, where $\ell^{e-1} \le (\sqrt{M}+1)^2 < \ell^e$.
   b. If $d = \gcd(Q_z, N)$ is not 1 then return d if $d < N$ or `failure` if $d = N$.
5. Return `failure`.

The scalar multiplication in step 4a is performed using projective coordinates, and while it is defined in terms of the group operation in $E(\mathbb{Q})$, we only keep track of the coordinates of Q modulo N; the projective coordinates are integers and there are no inversions involved, so all of the arithmetic can be performed in $\mathbb{Z}/N\mathbb{Z}$.

Theorem 10.12. *Assume $4a^3+27b^2$ is not divisible by N, and let P_1 and P_2 be the reductions of P modulo distinct primes p_1 and p_2 dividing N, with $p_1 \le M$. Suppose $|P_1|$ is ℓ_1-smooth and $|P_2|$ is not, for some prime $\ell_1 \le B$. Then Algorithm 10.11 succeeds.*

Proof. When the algorithm reaches step 4b with $\ell = \ell_1$ we must have $Q = mP$, where $m = \prod_{\ell \le \ell_1} \ell^e$ is a multiple of $|P_1|$, since $|P_1|$ is ℓ_1-smooth and $|P_1| \le (\sqrt{p_1}+1)^2 \le (\sqrt{M}+1)^2$. So $Q \equiv 0 \bmod p_1$, but $Q \not\equiv 0 \bmod p_2$, since $|P_2|$ is not ℓ_1-smooth. Therefore Q_z is divisible by p_1 but not p_2 and a proper factor $d = \gcd(Q_z, N)$ of N will be found in step 4b. □

If the algorithm fails, we can simply try again. Heuristically, provided N is not a perfect power and has a prime factor $p \le M$, we will eventually succeed. Factoring perfect powers can be efficiently handled by the algorithm developed in Problem 1 of Problem Set 3. Provided N is not a prime power and has a prime factor $p < M$, Algorithm 10.11 is very likely to succeed whenever it picks a triple (x_0, y_0, a) that yields an elliptic curve whose reduction modulo p has B-smooth order. So the number of times we expect to run the algorithm before we succeed depends on the probability that $\#E(\mathbb{F}_p)$ is B-smooth.

The integer $\#E(\mathbb{F}_p)$ must lie in the Hasse interval $[p+1-2\sqrt{p}, p+1+2\sqrt{p}]$, which is unfortunately too narrow for us to apply any theorems on the density of B-smooth integers

[7] There are standard parameterizations that are guaranteed to produce a curve $E/\mathbb{Q}$ with a known point $P \in E(\mathbb{Q})$ of infinite order; see [1], for example. Here we just generate random E and P at random.

[8] This follows (for example) from the Lutz–Nagell theorem [20, Theorem 8.7], which implies that if y_0 is nonzero then y_0^2 must divide $4a^3 + 27b^2 = 4a^3 + 27(x_0^3 + ax_0)^2$, which is extremely unlikely.

(we cannot even prove that this interval contains any primes, and smooth numbers are much rarer than primes). So to analyze the complexity of Algorithm 10.11 (and to optimize the choice of B), we resort to the heuristic assumption that, at least when $\#E(\mathbb{F}_p)$ lies in the narrower interval $[p+1-\sqrt{p}, p+1+\sqrt{p}]$, the probability the $\#E(\mathbb{F}_p)$ is B-smooth is comparable to the probability that a random integer in the interval $[p, 2p]$ is B-smooth.[9]

One can prove that the probability that $\#E(\mathbb{F}_p)$ lies in $[p+1-\sqrt{p}, p+1+\sqrt{p}]$ is at least $1/2$ (this is implied, asymptotically, by the Sato–Tate theorem), and further that the probability that $\#E(\mathbb{F}_p)$ takes on any particular value in this interval is $\Omega(1/(\sqrt{p}\log p))$. These facts are both proved in Lenstra's paper [17], and we will be able to prove them ourselves once we have covered the theory of complex multiplication. This means that we can make our heuristic assumption independent of any facts about elliptic curves, we simply need to assume that a random integer in the interval $[p+1-\sqrt{p}, p+1+\sqrt{p}]$ has roughly the same probability of being B-smooth as a random integer in the interval $[p, 2p]$.

Under our heuristic assumption, the analysis of the algorithm follows the analysis of the Pollard $p-1$ method. This algorithm takes $O(\pi(B)(\log M)\,\mathsf{M}(\log N))$ time per elliptic curve, and if N has a prime factor $p \le M$, it will need to try an average of $O(u^u)$ curves before it finds a factor. As in §10.5, this implies that the optimal value of B is $L_M[1/2, 1/\sqrt{2}]$, and with this value of B the expected time to factor N is $L_M[1/2, \sqrt{2}]\,\mathsf{M}(\log N)$. In general, we may not know a bound M on the smallest prime factor p of N *a priori*, but if we simply start with a small choice of M and periodically double it, we can achieve a running time of

$$L_p[1/2, \sqrt{2}]\,\mathsf{M}(\log N),$$

where p is the smallest prime factor of N.

A crucial point is that this running time depends almost entirely on p rather than N, a property that distinguishes ECM from all other factorization algorithms with heuristically subexponential running times. There are factorization algorithms such as the quadratic sieve and the number field sieve that are heuristically faster when all of the prime factors of N are large, but in practice one first uses ECM to look for any relatively small prime factors before resorting to these heavyweight algorithms.

The fact that the complexity of ECM depends primarily on the size of the smallest prime divisor of N also makes it a very good algorithm for smoothness testing. Testing whether a given integer N is $L_N[1/2, c]$-smooth using ECM takes just

$$\begin{aligned} L_{L_N[1/2,c]}\left[1/2, \sqrt{2}\right] &\approx \exp\left(\sqrt{2\log(\exp(c\sqrt{\log N \log\log N})\log\log(\exp(c\sqrt{\log N \log\log N})}\right) \\ &= \exp\left(\sqrt{2c\sqrt{\log N \log\log N}(1/2+o(1))\log\log N}\right) \\ &= \exp\left((\sqrt{c}+o(1))(\log N)^{1/4}(\log\log N)^{3/4}\right) \\ &= L_N\left[1/4, \sqrt{c}\right] \end{aligned}$$

expected time, which is faster than any other method known.[10]

10.7 Efficient implementation

Algorithm 10.11 spends essentially all of its time performing elliptic curve scalar multiplications modulo N, so it is worth choosing the elliptic curve representation and the coordinate

[9] Asymptotically, this is the same as the probability that a random integer in $[1, p]$ is B-smooth.

[10] As noted earlier, for batch smoothness testing, Bernstein's algorithm [3] is faster.

system to optimize this operation. Edwards curves, which we saw in Lecture 2, are an excellent choice; see [4] for a detailed discussion of how to efficiently implement ECM using Edwards curves. Another popular choice is Montgomery curves [18]; as explained in [5], there is a close relationship between Montgomery curves and Edwards curves. These were originally introduced specifically for the purpose of optimizing the elliptic curve factorization method but are now used in many other applications of elliptic curves, including primality proving and cryptography.

10.8 Montgomery Curves

A *Montgomery curve* is an elliptic curve defined by an equation of the form

$$By^2 = x^3 + Ax^2 + x, \tag{6}$$

where $B \neq 0$ and $A \neq \pm 2$. To convert this to Weierstrass form, let $u = Bx$ and $w = B^2y$. Substituting $x = u/B$ and $y = w/B^2$ in (6) and multiplying by B^3 yields

$$w^2 = u^3 + ABu^2 + B^2u,$$

which is in the form of a general Weierstrass equation. To obtain a short Weierstrass equation, we assume our base field has characteristic different from 3 and complete the cube by letting $v = u + \frac{AB}{3}$. We then obtain

$$\begin{aligned}
w^2 &= u^3 + ABu^2 + B^2u \\
w^2 &= \left(v - \frac{AB}{3}\right)^3 + AB\left(v - \frac{AB}{3}\right)^2 + B^2\left(v - \frac{AB}{3}\right) \\
w^2 &= v^3 - ABv^2 + \frac{A^2B^2}{3}v - \frac{A^3B^3}{27} + ABv^2 - \frac{2A^2B^2}{3}v + \frac{A^3B^3}{9} + B^2v - \frac{AB^3}{3} \\
w^2 &= v^3 + \left(B^2 - \frac{A^2B^2}{3}\right)v + \left(\frac{2A^3B^3}{27} - \frac{AB^3}{3}\right).
\end{aligned}$$

In order to check that (6) actually defines an elliptic curve, we should verify that it is nonsingular. We could do these using the coefficients of the curve in short Weierstrass form, but it is easier to do this directly. We need to determine whether there are any points $(x : y : z)$ on the projective curve $By^2z = x^3 + Ax^2z + xz^2$ at which all three partial derivatives vanish. For any such point we must have

$$\frac{\partial}{\partial x}\colon 3x^2 + 2Axz + z^2 = 0, \qquad \frac{\partial}{\partial y}\colon 2Byz = 0, \qquad \frac{\partial}{\partial z}\colon By^2 - (Ax^2 + 2xz) = 0.$$

We assume we are working in a field of characteristic not equal to 2 or 3. Suppose that $y \neq 0$. Then the equation for $\frac{\partial}{\partial y}$ gives $z = 0$, and from $\frac{\partial}{\partial x}$, we get $x = 0$. But this is a contradiction, since the equation for $\frac{\partial}{\partial z}$ is not satisfied. On the other hand, if $y = 0$, then $z = -\frac{A}{2}x \neq 0$. We have $3x^2 - A^2x^2 + \frac{A^2}{4}x^2 = 0$, and therefore $3 - \frac{3}{4}A^2 = 0$, since $x \neq 0$. Thus $A^2 = 4$, but we require $A \neq \pm 2$ in (6), so this cannot be the case.

10.9 Montgomery curve group law

The transformation of a Montgomery curve to Weierstrass form is a linear transformation that preserves the symmetry about the y-axis, so the geometric view of the group law

remains the same: three points on a line sum to zero, which is the point at infinity. To add points P_1 and P_2 we construct the line $\overline{P_1P_2}$ (using a tangent when $P_1 = P_2$), find the third intersection point with the curve, and then reflect over the y-axis to obtain $P_3 = P_1 + P_2$. In this section we compute explicit algebraic formulas for this operation, just as we did for curves in Weierstrass form earlier in the course.

The cases involving inverses and the point at infinity are easy (we have $P - P = 0$ and $P + 0 = 0 + P = P$), so let $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ be two (possibly equal but not opposite) affine points on the curve whose sum $P_3 = (x_3, y_3)$ we wish to compute. We first compute the slope m of the line $\overline{P_1P_2}$.

$$m = \begin{cases} \dfrac{y_1 - y_2}{x_1 - x_2} & \text{if } P_1 \neq P_2, \\ \dfrac{3x_1^2 + 2Ax_1 + 1}{2By_1} & \text{if } P_1 = P_2. \end{cases} \tag{7}$$

Now we want to intersect the line $y - y_1 = m(x - x_1)$ with the curve equation (6). Substituting $m(x - x_1) + y_1$ in for y, we get

$$B\left(m(x - x_1) + y_1\right)^2 = x^3 + Ax^2 + x. \tag{8}$$

We know x_1, x_2, and x_3 are the three roots of this cubic equation, since P_1, P_2, and $-P_3$ all lie on the curve and the line $\overline{P_1P_2}$. Thus the coefficient of x^2 in (8) must be equal to $x_1 + x_2 + x_3$. We get a Bm^2x^2 term on the left side of (8) and an Ax^2 term on the right, so we have $x_1 + x_2 + x_3 = Bm^2 - A$. Solving for x_3 and using the equation for $\overline{P_1P_2}$ to compute $-y_3$, we obtain

$$\begin{aligned} x_3 &= Bm^2 - (A + x_1 + x_2) \\ y_3 &= m(x_1 - x_3) - y_1. \end{aligned} \tag{9}$$

These formulas closely resemble the formulas for a curve in short Weierstrass form, but with an extra B and A in the equation for x_3. However, they have the key property that they allow us to completely eliminate the y-coordinate from consideration. This is useful because the y-coordinate is not needed in many applications; we do not need to know the y-coordinate of a point P in order to determine whether $mP = 0$ for a given integer m. This makes the y-coordinate superfluous in applications such as ECM and ECPP.

Let us consider the doubling case first. Plugging in the expression for m given by (7) in the case $P_1 = P_2 = (x_1, y_1)$ into (9) and remembering the curve equation $By^2 = x^3 + Ax^2 + x$,

$$\begin{aligned} x_3 &= B\frac{(3x_1^2 + 2Ax_1 + 1)^2}{4B^2y_1^2} - (A + 2x_1) \\ &= \frac{(3x_1^2 + 2Ax_1 + 1)^2 - 4(A + 2x_1)(x_1^3 + Ax_1^2 + x_1)}{4(x_1^3 + Ax_1^2 + x_1)} \\ &= \frac{(x_1^2 - 1)^2}{4x_1(x_1^2 + Ax_1 + 1)}, \end{aligned}$$

thus we can derive x_3 from x_1 without needing to know y_1. In projective coordinates,

$$\begin{aligned} &= \frac{(x_1^2 - z_1^2)^2}{4x_1z_1(x_1^2 + Ax_1z_1 + z_1^2)} \\ &= \frac{(x_1^2 - z_1^2)^2}{4x_1z_1\big((x_1 - z_1)^2 + (A + 2)x_1z_1\big)}. \end{aligned}$$

Thus we may write

$$\begin{aligned} x_3 &= (x_1+z_1)^2(x_1-z_1)^2 \\ 4x_1z_1 &= (x_1+z_1)^2-(x_1-z_1)^2 \\ z_3 &= 4x_1z_1\big((x_1-z_1)^2+C(4x_1z_1)\big). \end{aligned} \tag{10}$$

where $C=(A+2)/4$. Notice that these formulas do not involve y_1 and they only require 5 multiplications: 3 to compute x_3, none to compute $4x_1z_1$, and 2 more to compute z_3. One of these is a multiplication by the constant C, which may take negligible time if we can arrange for C to be small.

Now let us do the same thing for addition:

$$\begin{aligned} x_3 &= B\frac{(y_1-y_2)^2}{(x_1-x_2)^2}-(A+x_1+x_2) \\ x_3(x_1-x_2)^2 &= B(y_1-y_2)^2-(A+x_1+x_2)(x_1-x_2)^2 \\ &= By_1^2+By_2^2-2By_1y_2-(A+x_1+x_2)(x_1-x_2)^2 \\ &= -2By_1y_2+2x_1x_2(A+x_1+x_2)+x_1+x_2 \\ &= -2By_1y_2+x_2(x_1^2+Ax_1+1)+x_1(x_2^2+Ax_2+1) \\ &= -2By_1y_2+\frac{x_2}{x_1}By_1^2+\frac{x_1}{x_2}By_2^2 \\ &= B\frac{(x_2y_1-x_1y_2)^2}{x_1x_2} \end{aligned} \tag{11}$$

This gives us an equation for x_3 in $P_3=P_1+P_2$, but it still involves the y-coordinates of P_1 and P_2. To address this, let us also compute the x-coordinate x_4 of $P_4=P_1-P_2$. The hard work is already done, we just need to negate y_2 in the equation for x_3. Thus

$$x_4(x_1-x_2)^2=B\frac{(x_2y_1+x_1y_2)^2}{x_1x_2}. \tag{12}$$

Multiplying equations (11) and (12) yields

$$\begin{aligned} x_3x_4(x_1-x_2)^4 &= \frac{B^2(x_2^2y_1^2-x_1^2y_2^2)^2}{x_1^2x_2^2}=\frac{(x_2^2By_1^2-x_1^2By_2^2)^2}{x_1^2x_2^2} \\ &= \frac{\big(x_2^2(x_1^3+Ax_1^2+x_1)-x_1^2(x_2^3+Ax_2^2+x_2)\big)^2}{x_1^2x_2^2} \\ &= \big(x_2(x_1^2+Ax_1+1)-x_1(x_2^2+Ax_2+1)\big)^2 \\ &= (x_2x_1^2-x_1x_2^2+x_2-x_1)^2 \\ &= \big((x_1-x_2)(x_1x_2-1)\big)^2. \end{aligned}$$

Canceling a factor of $(x_1-x_2)^2$ from both sides gives

$$x_3x_4(x_1-x_2)^2=(x_1x_2-1)^2, \tag{13}$$

which does not involve y_1 or y_2 (but does require us to know x_4).

We now switch to projective coordinates:

$$\frac{x_3}{z_3} \cdot \frac{x_4}{z_4}\left(\frac{x_1}{z_1} - \frac{x_2}{z_2}\right)^2 = \left(\frac{x_1x_2}{z_1z_2} - 1\right)^2$$
$$\frac{x_3}{z_3} = \frac{z_4}{x_4} \cdot \frac{(x_1x_2 - z_1z_2)^2}{(x_1z_2 - x_2z_1)^2},$$

which yields

$$x_3 = z_4\left[(x_1 - z_1)(x_2 + z_2) + (x_1 + z_1)(x_2 - z_2)\right]^2 \tag{14}$$
$$z_3 = x_4\left[(x_1 - z_1)(x_2 + z_2) - (x_1 + z_1)(x_2 - z_2)\right]^2$$

These formulas require just 6 multiplications, but they assume that we already know the x-coordinate x_4/z_4 of $P_1 - P_2$. But if we structure the double-and-add algorithm for scalar multiplication appropriately, we can use the formulas in (10) and (14) to efficiently compute the x-coordinate of the scalar multiple mP using what is known as a *Montgomery ladder*. We assume points are represented simply as projective pairs $(x : z)$ that omit the y-coordinate.

Algorithm 10.13 (Montgomery Ladder)**.**
Input: A point $P = (x_1 : z_1)$ on a Montgomery curve and a positive integer m.
Output: The point $mP = (x_m : z_m)$.

1. Let $m = \sum_{i=0}^{k} m_i 2^i$ be the binary representation of m.
2. Set $Q[0] = P$ and compute $Q[1] = 2P$ (note that $P = Q[1] - Q[0]$).
3. For $i = k - 1$ down to 0:
   a. $Q[1 - m_i] \leftarrow Q[1] + Q[0]$ (Using $P = Q[1] - Q[0]$)
   b. $Q[m_i] \leftarrow 2Q[0]$
4. Return $Q[0]$.

The Montgomery ladder is the usual double-and-add algorithm, augmented to ensure that $Q[1] - Q[0] = P$ is invariant throughout. A nice feature of the algorithm is that every iteration of the loop is essentially the same: a Montgomery addition followed by a Montgomery doubling. This makes the algorithm resistant to side-channel attacks. If we assume that the input point P is in affine form $(x_1 : 1)$, then $z_1 = z_4 = 1$ in the addition formulas in (14), which saves one multiplication. This yields a total cost of $\big(10+o(1)\big)\log_2 m$ field multiplications for Algorithm 10.13, or only $\big(9+o(1)\big)\log_2 m$ if the constant C is small enough to make the multiplications by C negligible. This is faster than using Edwards' curves (at least in a side-channel resistant configuration where one is not using optimized doubling formulas).

An implementation of Algorithms 10.11 and 10.13 can be found in this Sage notebook.

10.10 Torsion on a Montgomery Curve

Every Montgomery curve has $(0, 0)$ as a rational point of order 2 (as with curves in short Weierstrass form, the points of order 2 are precisely those with y-coordinate 0). This tells us that not every elliptic curve can be put in Montgomery form, since not every elliptic curve has a rational point of order 2. In fact, more is true.

Theorem 10.14. *The Montgomery curve E/k defined by $By^2 = x^3 + Ax^2 + x$ has either three rational points of order 2 or a rational point of order 4 (possibly both).*

Proof. The cubic $x^3 + Ax^2 + x$ has either one or three rational roots, and these roots are distinct, since the curve is nonsingular. If it has three roots, then there are three rational points of the form $(x, 0)$, all of which have order 2.

If it has only one root, then $x^2 + Ax + 1$ has no roots, so $A^2 - 4 = (A+2)(A-2)$ is not a quadratic residue. Therefore one of $A+2$ and $A-2$ is a quadratic residue (and the other is not), so either $\frac{A+2}{B}$ or $\frac{A-2}{B}$ is a quadratic residue. We will use this fact to find a point of order 4 that doubles to the 2-torsion point $(0, 0)$, which is the unique point on the curve whose x-coordinate is 0.

To get $x_3 = 0$ in the doubling formulas (10), we must have $x_1 = \pm z_1$, equivalently, $x_1/z_1 = \pm 1$. Plugging this into the curve equation, we seek a solution to either $By^2 = A+2$ or $By^2 = A-2$. But we have already shown that either $\frac{A+2}{B}$ or $\frac{A-2}{B}$ is a quadratic residue, so one of these equations has a solution and there is a rational point of order 4. □

Thus, like Edwards curves, the torsion subgroup of a Montgomery curve always has order divisible by 4. For the purposes of the ECM algorithm this is actually a feature, since it slightly increases the likelihood that the group order will be smooth. In fact, most implementations use specific parameterizations to generate curves $E/\mathbb{Q}$ that are guaranteed to have even larger torsion subgroups, typically isomorphic to either $\mathbb{Z}/12\mathbb{Z}$ or $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/8\mathbb{Z}$; see [1, 4, 18] for examples (the $\mathbb{Z}/12\mathbb{Z}$ case is illustrated in the example implementation).

The converse of Theorem 10.14 does not hold; there are elliptic curves with three rational points of order 2 that cannot be put in Montgomery form. However, every elliptic curve with a rational point of order 4 can be put in Montgomery form.

Theorem 10.15. *Let $E\colon y^2 = x^3 + ax + b$ be an elliptic curve over a field k. Suppose $E(k)$ contains a point P of order 4, and let $2P = (x_0, 0)$. Then $3x_0^2 + a$ is a square in k and E can be put in Montgomery form $E'\colon By^2 = x^3 + Ax^2 + x$ by setting $B = 1/\sqrt{3x_0^2 + a}$ and $A = 3x_0B$; the map $(x, y) \mapsto (B(x - x_0), By)$ defines an isomorphism from E to E'.*

Proof. Let $P = (u, v)$. From the elliptic curve doubling formula, we have

$$\begin{aligned} x_0 &= \left(\frac{3u^2 + a}{2v}\right)^2 - 2u \\ &= \frac{(9u^4 + 6au^2 + a^2) - 8u(u^3 + au + b)}{4(u^3 + au + b)} \\ &= \frac{u^4 - 2au^2 - 8bu + a^2}{4(u^3 + au + b)}. \end{aligned}$$

Therefore u satisfies

$$u^4 - 4x_0u^3 - 2au^2 - (4ax_0 + 8b)u - 4bx_0 + a^2 = 0.$$

We have $0^2 = x_0^3 + ax_0 + b$, so we can replace b by $-x_0^3 - ax_0$, yielding

$$u^4 - 4x_0u^3 - 2au^2 + (8x_0^3 + 4ax_0)u + 4x_0^4 + 4ax_0^2 + a^2 = 0.$$

The LHS is a perfect square. If we put $u = z + x_0$ we can write this as

$$(z^2 - (3x_0^2 + a))^2 = 0.$$

Now $z = u - x_0 \in k$, so $z^2 - (3x_0^2 + a)$ must have a root in k. Thus $3x_0^2 + a$ is a square, as claimed, and it is nonzero because x_0 is not a repeated root of $x_0^3 + ax_0 + b$. Now let $B = 1/\sqrt{3x_0^2 + a}$ and $A = 3x_0B$ be as in the theorem and let $E'\colon By^2 = x^3 + Ax^2 + x$.

To check that $(x, y) \mapsto (B(x - x_0), By)$ defines an isomorphism from $E \to E'$, we plug $(B(x - x_0), By)$ into the equation for E' and note that

$$\begin{aligned}
B(By)^2 &= (B(x - x_0))^3 + A(B(x - x_0))^2 + B(x - x_0)\\
B^2y^2 &= B^2(x^3 - 3x_0x^2 + 3x_0^2x - x_0^3) + 3x_0B^2(x^2 - 2x_0x + x_0^2) + x - x_0\\
y^2 &= x^3 - 3x_0^2x + 2x_0^3 + (x - x_0)(3x_0^2 + a)\\
y^2 &= x^3 + ax - x_0^3 - ax_0\\
y^2 &= x^3 + ax + b.
\end{aligned}$$

This also shows that E' is not singular, since E is not (so we must have $A^2 \neq 4$). □

References

[1] A.O.L. Atkin and François Morain, *Finding suitable curves for the elliptic curve method of factorization*, Mathematics of Computation **60** (1993), 399–405.

[2] Razvan Barbulescu, Pierrick Gaudry, Antoine Joux, Emmanuel Thomé, *A heuristic quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic*, in Advances in Cryptology — EUROCRYPT 2014, LNCS **8441** (2014), Springer, 1–16.

[3] Daniel J. Bernstein, *How to find smooth parts of integers*, unpublished preprint, 2004.

[4] Daniel J. Bernstein, Peter Birkner, Tanja Lange, and C. Peters, *ECM using Edwards curves*, Mathematics of Computation **82** (2013), 1139–1179.

[5] Daniel J. Bernstein and Tanja Lange, *Montgomery curves and the Montgomery ladder*, Cryptology ePrint Archive, Report 2017/293, 2017.

[6] Fabrice Boudot, Pierrick Gaudry, Aurore Guillevic, Nadia Heninger, Emmanuel Thomé, and Paul Zimmermann, *795-bit factoring and discrete logarithms*, NMBRTHRY listserv posting, December 2, 2019.

[7] E. R. Canfield, Paul Erdős, and Carl Pomerance, *On a problem of Oppenheim concerning "factorisatio numerorum"*, Journal of Number Theory **17** (1983), 1–28.

[8] Andreas Enge, *Discrete logarithms in curves over finite fields*, Finite fields and applications, Contemporary Mathematics **461**, AMS, 2008, 119–139 (arXiv:0712.3916).

[9] Andreas Enge and Pierrick Gaudry, *A general framework for subexponential discrete logarithm algorithms*, Acta Arithmetica **102** (2002), 83–103.

[10] Pierrick Gaudry, *Index calculus for abelian varieties of small dimension and the elliptic curve discrete logarithm problem*, J. Symbolic Computation **44** (2009), 1690–1702.

[11] Robert Granger, Thorsten Kleinjung, Arjen Lenstra, Benjamin Wesolowski, Jens Zumbrägel, *Discrete logarithms in $GF(2^{30750})$*, NMBRTHRY listserv posting, July 10, 2019.

[12] Daniel M. Gordon, *Discrete Logarithms in $GF(p)$ using the number field sieve*, SIAM J. Discrete Math **6** (1993), 124–138.

[13] Andrew Granville, *Smooth numbers, computational number theory and beyond*, in Algorithmic Number Theory: Lattices, Number Fields, Curves and Cryptography (MSRI Workshop), Mathematical Sciences Research Institute Publications **44**, 2008, 267–324.

[14] Antoine Joux, *A new index calculus algorithm with complexity $L(1/4 + o(1))$ in very small characteristic*, in Selected Areas in Cryptography — SAC 2013, LNCS **8282** (2014), Springer, 355–379.

[15] Thorsten Kleinjung, *Discrete logarithms in $GF(2^{1279})$*, NMBRTHRY listserv posting, October 17, 2014.

[16] Antoine Joux, Reynald Lercier, Nigel Smart, and Frederik Vercauteren, *The number field sieve in the medium prime case*, Advances in Cryptology — CRYPTO 2006, LNCS **4117** (2006), Springer, 326–344.

[17] Hendrik Lenstra, *Factoring integers with elliptic curves*, Annals of Mathematics **126** (1987), 649–673

[18] Peter L. Montgomery, *Speeding the Pollard and elliptic curve methods of factorization*, Mathematics of Computation **48** (1987), 243–264.

[19] J. M. Pollard, *Theorems of factorization and primality testing*, Proceedings of the Cambridge Philosophical Society **76** (1974): 521–528

[20] Lawrence C. Washington, *Elliptic curves: Number theory and cryptography*, second edition, Chapman and Hall/CRC, 2008.

[21] Paul Zimmermann and Bruce Dodson, *20 years of ECM*, Algorithmic Number Theory 7th International Symposium (ANTS VII), LNCS 4076 (2006), 525–542.

11 Primality proving

In this lecture, we consider the question of how to efficiently determine whether a given integer N is prime. This question is intimately related to the problem of factoring N; without a method for determining primality, we have no way of knowing when we have completely factored N. This is an important issue for probabilistic factorization algorithms such as the elliptic curve method (ECM): if we attempt to factor a prime with ECM, the algorithm will never terminate.

Even if we are able to guarantee termination, there is still the issue of correctness. If a Monte Carlo algorithm claims that an integer N is the product of two primes p and q, it is easy to verify that $N = pq$, but how do we know that this is the *complete* factorization of N? We need to be able to *prove* that p and q are both prime, and we would like to do so in a way that can be efficiently verified. Factoring is a lot harder than multiplication, and we might similarly expect that proving an integer is prime is harder than verifying the result, provided the prover can provide a "paper trail" that can be easily verified. This leads to the notion of a *certificate* for primality, and these can be constructed using elliptic curves.

11.1 Classical primality tests

The most elementary approach to primality proving is trial division: we attempt to divide N by every integer $p \le \sqrt{N}$. If no such p divides N, then N is prime. This takes $O(\sqrt{N}\,\mathsf{M}(\log N))$, which is impractical for large N, but it serves as a useful base case for more sophisticated recursive methods that we will consider.

Remark 11.1. This complexity bound can be slightly improved. Using fast sieving techniques [8, Alg. 3.2.2], we can enumerate the primes p up to $\sqrt{N}$ in $O(\sqrt{N}\log N/\log\log N)$ time and then perform trial divisions by just the primes $p \le \sqrt{N}$, rather than every integer $p \le \sqrt{N}$. Applying the prime number theorem and the Schönhage-Strassen bound, the sieving time dominates the cost of the divisions and the overall complexity of trial division is then $O(\sqrt{N}\log N/\log\log N)$.

Many classical primality tests are based on Fermat's little theorem.

Theorem 11.2 (Fermat)**.** *If N is prime, then for all $a \in \mathbb{Z}/N\mathbb{Z}$:*

$$a^N = a.$$

This implies that if $a^N \neq a$ for some $a \in \mathbb{Z}/N\mathbb{Z}$, then N cannot be prime. This gives us a way to efficiently prove that certain integers are composite. For example, $N = 91$ is not prime because

$$2^{91} \equiv 37 \bmod 91.$$

But this does not always work. For example, $341 = 11 \cdot 31$ is not prime, but

$$2^{341} \equiv 2 \bmod 341.$$

In this case, using a different value of a will work. If we take $a = 3$ we find that

$$3^{341} \equiv 168 \bmod 341,$$

which proves that 341 is not prime.

However, for certain composite integers N there is *no* choice of a that will work. Thus even if $a^N \equiv a \bmod N$ for every integer a, we cannot be sure that N is prime.

Definition 11.3. A *Carmichael number* is a composite integer N such that $a^N \equiv a \bmod N$ for every integer a.

The first four Carmichael numbers are $561, 1105, 1729$, and 2821; see sequence A002997 in the On-Line Encyclopedia of Integer Sequences (OEIS) for more examples, or this site for statistics regarding the 20,138,200 Carmichael numbers less than 10^{21}. The largest known Carmichael number has about 300 billion decimal digits and more than 10 billion distinct prime factors [5]. The question of whether or not there are infinitely many Carmichael numbers was open for more than 80 years and finally settled in 1994.

Theorem 11.4 (Alford-Granville-Pomerance)**.** *The set of Carmichael numbers is infinite.*

Proof. See [6]. □

The infinitude of Carmichael numbers implies that any approach based on Fermat's little theorem is doomed to fail for an infinite set of integers. We would like a criterion that holds if, *and only if*, N is prime. One candidate is the following theorem, which uses the Euler function $\phi(N) = \#(\mathbb{Z}/N\mathbb{Z})^\times$, which we recall is multiplicative (meaning that $\phi(ab) = \phi(a)\phi(b)$ for all $a \perp b$), by the Chinese remainder theorem.

Theorem 11.5. *A positive integer N is prime if and only if $\phi(N) = N - 1$.*

Proof. If N is prime every nonzero residue class in $\mathbb{Z}/N\mathbb{Z}$ is invertible and $\phi(N) = N - 1$. Otherwise there is a nonzero residue class that is not invertible and $\phi(N) \le N - 2$. □

One approach suggested by this theorem is to simply compute $\phi(N)$ and check whether it is equal to $N - 1$. However, computing $\phi(N)$ is very difficult, in general.[1] Fortunately, we can use Theorem 11.5 in a less obvious way, via the following lemma. We restrict our attention to odd integers $N > 1$, since it is easy to tell whether an even integer is prime or not (and 1 is not prime).

Lemma 11.6. *Let $p = 2^s t + 1$ be prime, with t odd, and let a be an integer that is nonzero modulo p. Exactly one of the following holds:*

(i) $a^t \equiv 1 \bmod p$;

(ii) $a^{2^i t} \equiv -1 \bmod p$, *for some* $0 \le i < s$.

Proof. Consider the endomorphism $\varphi\colon x \mapsto x^t$ of the cyclic group $(\mathbb{Z}/p\mathbb{Z})^\times$ of order $2^s t$; the kernel and image of φ are cyclic subgroups of orders t and 2^s, respectively. For each $a \in (\mathbb{Z}/p\mathbb{Z})^\times$, either $a \in \ker\varphi$, in which case (i) holds, or $\varphi(a) = a^t$ has order 2^k for some $0 < k \le s$, in which case $a^{2^{k-1}t}$ has order 2 and must be equal to -1, the unique element of order 2 in $(\mathbb{Z}/p\mathbb{Z})^\times$, so (ii) holds with $i = k - 1$. □

Definition 11.7. Let $N = 2^s t + 1$ be an odd integer, with t odd. An integer $a \not\equiv 0 \bmod N$ is a *witness* for (the compositeness of) N if both of the following hold:

$$\text{(i) } a^t \not\equiv 1 \bmod N \qquad \text{(ii) } a^{2^i t} \not\equiv -1 \bmod N \text{ for } 0 \le i < s.$$

[1] If N is the product of two primes, it is easy to show that computing $\phi(N)$ is as hard as factoring N, and under the Extended Riemann Hypothesis, this is true in general [13].

If a is a witness for an odd integer $N > 1$, then Lemma 11.6 implies that N is composite. Prime numbers clearly have no witnesses. It is not immediately clear that every odd composite integer N necessarily has a witness, but this is true. In fact, if we pick a at random it is quite likely to be a witness, as independently proved by Monier [14] and Rabin [18].

Theorem 11.8 (Monier–Rabin)**.** *Let N be an odd composite integer. The probability that a random integer $a \in [1, N-1]$ is a witness for N is at least $3/4$.*

The theorem suggests that if N is composite and we pick, say, 100 random integers $a \in [1, N-1]$, then we are almost certainly going to find a witness for N. On the other hand, if N is prime then we will not find a witness. This doesn't actually *prove* that N is prime (unless we try more than $1/4$ of all $a \in [1, N-1]$), but we can at least view it as strongly supporting this possibility.

Proof.[2] Let $N = 2^s t + 1$ be an odd composite number with t odd, and let $N = q_1 \cdots q_r$ be its unique factorization into prime powers q_j. Let $b := a^t$ and let $b_j := b \bmod q_j$. If a is not a witness then either $b \equiv 1 \bmod N$, in which case $b_j \equiv 1 \bmod q_j$ for all j, or $b^{2^i} \equiv -1 \bmod N$ for some $0 \le i < s$, in which case $b_j^{2^i} \equiv -1 \bmod q_j$ for all j. If we put $i := -1$ in the first case, then each b_j is an element of order 2^{i+1} in the 2-Sylow subgroup S_j of $(\mathbb{Z}/q_j\mathbb{Z})^\times$.

We will bound the probability that every b_j is an element of S_j of order 2^{i+1} by $1/4$. Note that b_j need not be uniformly distributed modulo q_j, so some care is required.

Case 1: N is divisible by a square. Then some $q_j = p^k$ with $k > 1$. Since p is odd, the group $(\mathbb{Z}/p^k\mathbb{Z})^\times$ is cyclic of order $\phi(p^k) = p^{k-1}(p-1)$, and t is coprime to p (since it is coprime to N), so the probability that b_j lies in S_j at most $1/p^{k-1}$; this is at most $1/4$ except when $p^k = 9$. For $q_j = p^k = 9$ we have $b_j \in S_j = \{\pm 1\}$ if and only if $a \bmod q_j \equiv \pm 1$, since $3 \nmid t \,|\, (N-1)$. This occurs with probability at most $2/8 = 1/4$, since each of the 8 nonzero residues modulo 9 is equally likely.

Case 2: N is a product of $r \ge 3$ distinct primes q_j. Each 2-Sylow subgroup S_j is a cyclic of order 2^{k_j}, for some $k_j > 1$, and at most half the elements in S_j can have any particular order. If we assume each b_j actually lies in S_j then they are uniformly distributed in S_j (since t is odd), and the probability they all have the same order is at most $1/4$.[3]

Case 3: $N = q_1 q_2$ is a product of 2 distinct primes. Let $q_1 = 2^{s_1} t_1 + 1$, and $q_2 = 2^{s_2} t_2 + 1$, with $s_1 \ge s_2$ and $t_1, t_2 \perp 2$. Define the random variable X_j to be -1 if b_j does not lie in S_j, otherwise let $X_j = i$ where b_j has order 2^i in S_j. We wish to show $\Pr[X_1 = X_2 \ge 0] \le 1/4$.

Suppose $s_1 > s_2$. Half the elements in S_1 have order $2^{s_1} > 2^{s_2}$, so $\Pr[0 \le X_1 \le s_2] \le 1/2$, and $\Pr[X_2 = X_1 | 0 \le X_1 \le s_2] \le 1/2$; therefore $\Pr[X_1 = X_2 \ge 0] \le 1/4$.

Now suppose that $s_1 = s_2$. We have

$$2^s t = N - 1 = q_1 q_2 - 1 = (q_1 - 1)(q_2 - 1) + (q_1 - 1) + (q_2 - 1) = 2^s t_1 t_2 + 2^{s_1} t_1 + 2^{s_2} t_2,$$

thus if t_1 divides t then it also divides t_2, and conversely. If t_1 and t_2 both divide t, then $t_1 = t_2$ and $q_1 = q_2$, a contradiction. So assume $t_1 \nmid t$. Then $t_1 \ne 1$ must be divisible by a power of an odd prime $\ell \ge 3$ that does not divide t. It follows that $\Pr[X_1 \ge 0] \le 1/3$, and we also have $\Pr[X_1 = X_2 | X_1 \ge 0] \le 1/2$, therefore $\Pr[X_1 = X_2 \ge 0] \le 1/6 < 1/4$. □

Theorem 11.8 yields the following probabilistic primality test, due to Gary Miller [13] and Michael Rabin [18]

[2] The proof we give here is a bit different (and more elementary) than the proofs of Monier and Rabin.

[3] This rules out all Carmichael numbers, since they all have at least 3 distinct prime factors.

Algorithm 11.9 (Miller-Rabin)**.** Given an odd integer $N > 1$:

1. Pick a random integer $a \in [1, N-1]$.
2. Write $N = 2^s t + 1$, with t odd, and compute $b = a^t \bmod N$.
   If $b \equiv \pm 1 \bmod N$, return **true** (a is not a witness, N could be prime).
3. For i from 1 to $s-1$:
   a. Set $b \leftarrow b^2 \bmod N$.
   b. If $b \equiv -1 \bmod N$, return **true** (a is not a witness, N could be prime).
4. Return **false** (a is a witness, N is definitely not prime).

Example 11.10. For $N = 561$ we have $561 = 2^4 \cdot 35 + 1$, so $s = 4$ and $t = 35$, and for $a = 2$ we find that

$$2^{35} \equiv 263 \bmod 561,$$

which is not $\pm 1 \bmod 561$ so we continue and compute

$$263^2 \equiv 166 \bmod 561,$$
$$166^2 \equiv 67 \bmod 561,$$
$$67^2 \equiv 1 \bmod 561.$$

None of these values is congruent to -1, so $a = 2$ is a witness for $N = 561$ and we return **false**, meaning that 561 is definitely not prime. Note the contrast with the Fermat test, which jumps immediately to the last value computed above and does not detect that 561 is composite.

The Miller-Rabin test is a Monte Carlo algorithm with 1-sided error. If N is prime the algorithm will always correctly output **true**, and if N is composite the algorithm will correctly output **false** with probability at least $3/4$. The running time of the algorithm is $O(n\,\mathsf{M}(n))$, quasi-quadratic in $n = \log N$. This makes it extremely efficient, and it is the most widely used method for testing primality. In practical implementations, one performs several iterations of the Miller-Rabin test (choosing a new random integer a each time), and if they all return **true**, conclude that N is "probably prime".

But we should be careful how we interpret this. Any particular integer N is either prime or not; it makes no sense to say that N is prime with some probability. But if N is a uniformly distributed over some interval, then it does make sense to ask for the probability that N is prime, given that it passed a Miller-Rabin test. If N is selected from a large interval, say $[1, e^{1000}]$, then the probability that N is prime is quite small, approximately $1/1000$. In this situation, we need to be careful, since false positives are more likely than primes. It might appear to require several Miller-Rabin tests before we could say with better than 50% confidence that a large random integer N is prime. However, the Miller-Rabin test is far more powerful than Theorem 11.8 suggests.

Theorem 11.11 (Damgård-Landrock-Pomerance)**.** *Let N be a random odd integer in the interval $[2^{k-1}, 2^k]$ and let a be a random integer in $[1, N-1]$. Then*

$$\Pr[N \text{ is prime} \,|\, a \text{ is not a witness for } N] \geq 1 - k^2 \cdot 4^{2-\sqrt{k}}.$$

Proof. See [9, Thm. 2]. □

For large N, Theorem 11.11 gives excellent bounds on the probability that a random integer N is prime, given that it passes a single Miller-Rabin test. For example:

$$k = 256: \qquad 1 - k^2 \cdot 4^{2-\sqrt{k}} = 1 - 2^{-12},$$
$$k = 4096: \qquad 1 - k^2 \cdot 4^{2-\sqrt{k}} = 1 - 2^{-100}.$$

Thus when k is large it only takes a few successful Miller-Rabin tests to become astronomically confident that a randomly chosen integer N is prime.

11.2 Elliptic Curve Primality Proving

We now consider a method to unequivocally *prove* that a given integer N is prime or composite using elliptic curves. Elliptic curve primality proving (ECPP) was introduced by Goldwasser and Kilian in 1986 [10]. Like Lenstra's elliptic curve method (ECM) for integer factorization [11] which appeared at roughly the same time, it takes advantage of the fact that elliptic curves provide a way to generate abelian groups of varying orders over a fixed finite field. To simplify the statement of the Goldwasser-Kilian theorem, we first make the following definitions.

Definition 11.12. Let $P = (P_x : P_y : P_z)$ be a projective point on an elliptic curve $E/\mathbb{Q}$, with $P_x, P_y, P_z \in \mathbb{Z}$ and $\gcd(P_x, P_y, P_z) = 1$, and let N be a nonzero integer. If $P_z \equiv 0 \bmod N$ then P is *zero mod* N; otherwise, P is *nonzero mod* N. If $\gcd(P_z, N) = 1$ then P is *strongly nonzero mod* N.

Note that if P is strongly nonzero mod N, then P is nonzero mod p for every prime $p|N$. When N is prime, the notions of nonzero and strongly nonzero coincide. We now state the theorem, using $\Delta(E) := -16(4A^3 + 27B^2)$ to denote the discriminant of an elliptic curve $E\colon y^2 = x^3 + Ax + B$ in short Weierstrass form.

Theorem 11.13 (Goldwasser-Kilian)**.** *Let $E/\mathbb{Q}$ be an elliptic curve, and let $M, N > 1$ be integers with $M > (N^{1/4} + 1)^2$ and $N \perp \Delta(E)$, and let $P \in E(\mathbb{Q})$. If MP is zero mod N and $(M/\ell)P$ is strongly nonzero mod N for every prime $\ell|M$ then N is prime.*

Proof. Suppose for the sake of contradiction that the hypothesis holds and N is composite. Then N has a prime divisor $p \le \sqrt{N}$, and E has good reduction at p since $N \perp \Delta(E)$. Let M_p be the order of the reduction of P on E modulo p. The point MP is zero mod N and therefore zero mod p, so $M_p|M$; and we must have $M_p = M$, since $(M/\ell)P$ is strongly nonzero mod N and therefore nonzero mod p, for every prime $\ell|M$. Thus P has order M on the reduction of E modulo p, and by the Hasse bound, $M \le (\sqrt{p} + 1)^2$. But we also have $M > (N^{1/4} + 1)^2 \ge (p^{1/2} + 1)^2$, which is our desired contradiction. □

In order to apply the theorem, we need to know the prime factors q of M. In particular, we need to be sure that these q are actually prime! To simplify matters, we restrict ourselves to the case that $M = q$ is prime, and introduce the notion of a *primality certificate*.

Definition 11.14. A *primality certificate* for p is a tuple of integers

$$(p, A, B, x_1, y_1, q),$$

where $P = (x_1 : y_1 : 1)$ is a point on the elliptic curve $E\colon y^2 = x^3 + Ax + B$ over $\mathbb{Q}$, the integer $p > 1$ is prime to $\Delta(E)$, and qP is zero mod p with $q > (p^{1/4} + 1)^2$.

Note that $P = (x_1 : y_1 : 1)$ is strongly nonzero mod p, since its z-coordinate is 1. Theorem 11.13 implies that if there exists a primality certificate $(p, \ldots, q)$ for $N = p$ in which $M = q$ is prime, then p is prime. Thus a primality certificate $(p, \ldots, q)$ reduces the question of p's primality to the question of q's primality. Using a chain of such certificates, we can reduce to a case in which q is so small that we are happy to test its primality via trial division. This leads to the following recursive algorithm.

Algorithm 11.15 (Goldwasser-Kilian ECPP)**.** Given an odd integer p (a candidate prime), and a bound b, with $p > b > 5$, either construct a primality certificate (p, A, B, x_1, y_1, q) with $q \le (\sqrt{p}+1)^2/2$ or prove that p is composite.

1. Pick random integers $A, x_0, y_0 \in [0, p-1]$, and set $B = y_0^2 - x_0^3 - Ax_0$.
   Repeat until $\gcd(4A^3 + 27B^2, p) = 1$, then define $E\colon y^2 = x^3 + Ax + B$.
2. Use Schoof's algorithm to compute the number of points m on the reduction of E modulo p, assuming that p is prime. If anything goes wrong (which it might if p is actually composite), or if $m \notin \mathcal{H}(p)$, then return **composite**.
3. Write $m = cq$, where c is b-smooth and q is b-coarse (all prime factors greater than b). If $c = 1$ or $q \le (p^{1/4}+1)^2$, then go to step 1.
4. Perform a Miller-Rabin test on q. If it returns **false** then go to step 1.
5. Compute $P = (P_x : P_y : P_z) = c \cdot (x_0 : y_0 : 1)$ on E, working modulo p.
   If $\gcd(P_z, p) \ne 1$, go to step 1, else let $x_1 \equiv P_x/P_z \bmod p$ and $y_1 \equiv P_y/P_z \bmod p$.
6. Compute $Q = (Q_x : Q_y : Q_z) = q \cdot (x_1 : y_1 : 1)$ on E, working modulo p.
   If $Q_z \not\equiv 0 \bmod p$ then return **composite**.
7. If $q > b$, then recursively verify that q is prime using inputs q and b; otherwise, verify that q is prime by trial division. If q is found to be composite, go to step 1.
8. Output the certificate $(p, A, \tilde{B}, x_1, y_1, q)$, where $\tilde{B} \equiv B \bmod p$ is chosen so that we have $y_1^2 = x_1^3 + Ax_1 + \tilde{B}$ (over $\mathbb{Z}$ not just modulo p).

Note that step 4 is not strictly necessary, a composite q would eventually be detected in the recursive call, but it greatly reduces the probability that we will waste time in the recursive call, which speeds up the algorithm.

When the input to Algorithm 11.15 is prime, it will output a sequence of certificates, one for each recursive call, that reduce the question of p's primality to that of a prime $q < b$ that has been proved prime via trial division. Taken together, the sequence of primality certificates constitutes a *primality proof* for p. The complexity of this algorithm, and the complexity of verifying the primality proof it generates, are considered in the problem set, under the heuristic assumption that the integer m behaves like a random integer of similar size in terms of its factorization into b-smooth and b-coarse parts.

Without any heuristic assumptions, Goldwasser and Kilian proved that for almost all inputs p of a given size (all but a subexponentially small fraction), the expected running time of this algorithm is polynomial in $\log p$. Heuristically, this is believed to be true for all inputs, but we cannot prove this. Adleman and Huang later came up with a clever work-around to this problem that yielded an algorithm with a provably polynomial expected running time for all inputs [4]. Their strategy is to "reduce" the problem of proving the primality of the given input p to that of proving the primality of a *larger* prime $p' \approx p^2$. The key point is that the prime p' is obtained in a random way that makes it very likely that the Goldwasser-Kilian algorithm can prove its primality within a polynomial time bound (and

if this does not happen we can always generate a different p' and try again). In practice the algorithm of Adleman and Huang is never used, since it is believed that, in fact, it is always faster to just use the original Goldwasser-Kilian algorithm, no matter what p is, and the correctness of the Goldwasser-Kilian algorithm is guaranteed. But the Adleman-Huang result was theoretically significant, because it proved that primes could be recognized in polynomial time by a randomized algorithm (of course we can now do so deterministically, as discussed below, but this was a major open question at the time).

Remark 11.16. In [4] Adleman and Huang obtain the prime p' as the order of a randomly chosen abelian variety J_C of dimension 2 that is associated to a genus 2 curve C over $\mathbb{F}_p$ (assuming that p is prime). The abelian variety J_C is called the *Jacobian* of the curve C. It is analogous to the group of points on an elliptic curve (an abelian variety of dimension 1), except that when C has genus 2 the "points" on J_C actually correspond to pairs of points on the curve C. There is a generalization of Hasse's theorem due to Weil that implies that the cardinality of $J_C(\mathbb{F}_p)$ is on the order of p^2 and lies within an interval of width $\approx 8p^{3/2}$. This interval is large enough (relative to p^2) that we can prove that it contains many primes, roughly as many as implied by the prime number theorem. Adleman and Huang show that for a random curve C, the cardinality of $J_C(\mathbb{F}_p)$ is reasonably likely to be any one of a large subset of these primes, yielding a prime p' that is very likely to be one that the Goldwasser-Kilian algorithm can certify in polynomial time. In order to make this all work, Adleman and Huang modify the Goldwasser-Kilian algorithm slightly to make the proportion of bad inputs even smaller, and they also use the fact that $\#J_C(\mathbb{F}_p)$ can be computed in polynomial time using an analog of Schoof's algorithm due to Pila [16].

In fact, the original algorithm of Goldwasser-Kilian is no longer used; there is a much faster ECPP algorithm due to Atkin and Morain that uses the CM method to construct an elliptic curve E modulo p with suitable order m (assuming that p is prime), eliminating the need to generate many random curves, and use of Schoof's algorithm [3]. Like the Goldwasser-Kilian algorithm, this algorithm has not been proved to run in expected polynomial time, but in practice it is very fast. When combined with a further optimization due to Shallit [15], its expected running time is heuristically believed to be $\tilde{O}(n^4)$, where $n = \log p$. This makes it the current method of choice for general purpose primality proving. We will examine the Atkin-Morain algorithm more closely after we have studied the theory of complex multiplication.

We should note that there is now a deterministic polynomial-time algorithm for proving primality due to Agrawal, Kayal, and Saxena [2]. This is an important theoretical result, but it is not used in practice. The time bound proved in [2] is $\tilde{O}(n^{10.5})$; this can be improved to $\tilde{O}(n^6)$ (see [12]), but even with this improvement it is still much slower than the $\tilde{O}(n^4)$ heuristic complexity of ECPP. There is a randomized version of the AKS algorithm due to Bernstein [7] that runs in $\tilde{O}(n^4)$ time, but the constant factors appear to make it slower than ECPP, and it requires substantially more memory. The certificates it produces also take longer to verify.

The current record for general purpose primality proving is for the 40,000 digit partition number $p(1289844341)$ (the number of ways one can write 1289844341 as a sum of positive integers), which, as long suspected and now proved, happens to be prime. This record was set by Paul Underwood using an optimized version of the ECPP algorithm in February 2020 (see [17] for an up-to-date list of ECPP records). There are of course much larger integers that have been proved prime (for example, the 24 million digit Mersenne prime $2^{82589933}-1$), but these are all of a form that permits specialized $\tilde{O}(n^2)$-time algorithms to be used. There

are also specialized forms of elliptic curve primality proving that run in $\tilde{O}(n^2)$-time and these have been used to prove the primality of some large primes that no non-elliptic curve based method can feasibly handle [1].

References

[1] Alexander Abatzoglou, Alice Silverberg, Andrew V. Sutherland, and Angela Wong, *A framework for deterministic primality proving using elliptic curves with complex multiplication*, Mathematics of Computation **85** (2016), 1461–1483.

[2] Manindra Agrawal, Neeraj Kayal, Nitin Saxena, *Primes is in P*, Annals of Math. **160** (2004), 781–793.

[3] A. O. L. Atkin and François Morain, *Elliptic curves and primality proving*, Mathematics of Computation **61** (1993),29–68.

[4] Leonard M. Adleman and Ming-Deh A. Huang, *Primality testing and abelian varieties over finite fields*, Lecture Notes in Mathematics **1512**, Springer, 1992.

[5] W. R. Alford, Jon Grantham, Steven Hayman, and Andrew Shallue, *Constructing Carmichael numbers through improved subset-product algorithms*, Mathematics of Computation **83** (2014), 889–915.

[6] W. R. Alford, Andrew Granville, and Carl Pomerance, *There are infinitely many Carmichael numbers*, Annals of Mathematics **140** (1994), 703–722.

[7] Daniel J. Bernstien, *Proving primality in essentially quartic random time*, Mathematics of Computation **76** (2007), 389–403.

[8] Richard Crandall and Carl Pomerance, *Prime numbers: A computational perspective*, 2nd edition, Springer, 2005.

[9] Ivan Damgård, Peter Landrock, and Carl Pomerance, *Average case error estimates for the strong probable prime test*, Mathematics of Computation **61** (1993), 177–194.

[10] Shafi Goldwasser and Joseph Kilian, *Almost all primes can be quickly certified*, Proceedings of the Eighteenth ACM Symposium on the Theory of Computing (1986), 316–329.

[11] Hendrik Lenstra, *Factoring integers with elliptic curves*, Annals of Math. **126** (1987), 649–673

[12] Hendrik Lenstra and Carl Pomerance, *Primality testing with Gaussian periods*, Journal of the European Mathematical Society, to appear.

[13] Gary L. Miller, *Riemann's hypothesis and tests for primality*, Journal of Computer and System Sciences **13** (1976), 300-317.

[14] Louis Monier, *Evaluation and comparison of two efficient probabilistic primality testing algorithms*, Theoretical Computer Science **12** (1980), 97–108.

[15] François Morain, *Implementing the asymptotically fast version of the elliptic curve primality proving algorithm*, Mathematics of Computation **76** (2007), 493–505.

[16] Jonathan Pila, *Frobenius maps of abelian varieties and finding roots of unity in finite fields*, Math. Comp. **55** (1990), 745–763.

[17] Chris K. Caldwell, *The top twenty elliptic curve primality proofs*, Prime Pages website, accessed March 16, 2019.

[18] Michael O. Rabin *Probabilistic algorithm for testing primality*, Journal of Number Theory **12** (1980), 128–138.

12 Endomorphism algebras

The key to improving the efficiency of elliptic curve primality proving (and many other algorithms) is the ability to directly construct an elliptic curve $E/\mathbb{F}_q$ with a specified number of rational points, rather than generating curves at random until a suitable curve is found. To do this we need to develop the theory of *complex multiplication*. As a first step in this direction we introduce the endomorphism algebra of an elliptic curve and classify the possible endomorphism algebras of an elliptic curve.

Recall from Lecture 6 that the endomorphism ring $\mathrm{End}(E)$ of an elliptic curve E/k consists of the isogenies from E to itself, together with the zero morphism; addition is defined point-wise and multiplication is composition. The ring $\mathrm{End}(E)$ is not necessarily commutative, but its center (elements that commute with every other element of the ring) always contains the multiplication-by-n maps $[n]$; these form a subring of $\mathrm{End}(E)$ isomorphic to $\mathbb{Z}$. We will identify this subring with $\mathbb{Z}$, and may write n rather than $[n]$ without risk of confusion: note that $n\phi = \phi + \cdots + \phi$ is the same as $[n] \circ \phi$. We thus have $\mathbb{Z} \subseteq \mathrm{End}(E)$, but this inclusion is not necessarily an equality. The following facts about $\mathrm{End}(E)$ were proved in Lecture 6:

- $\mathrm{End}(E)$ has no zero divisors;
- $\deg\colon \mathrm{End}(E) \to \mathbb{Z}_{\geq 0}$ defined by $\alpha \mapsto \deg \alpha$ is multiplicative (with $\deg 0 \coloneqq 0$);
- $\deg n = n^2$ for all $n \in \mathbb{Z} \subseteq \mathrm{End}(E)$;
- each $\alpha \in \mathrm{End}(E)$ has a *dual* $\hat{\alpha} \in \mathrm{End}(E)$ with $\alpha\hat{\alpha} = \hat{\alpha}\alpha = \deg \alpha = \deg \hat{\alpha}$, and $\hat{\hat{\alpha}} = \alpha$;
- $\hat{n} = n$ for all $n \in \mathbb{Z} \subseteq \mathrm{End}(E)$;
- $\widehat{\alpha + \beta} = \hat{\alpha} + \hat{\beta}$ and $\widehat{\alpha\beta} = \hat{\beta}\hat{\alpha}$ for all $\alpha, \beta \in \mathrm{End}(E)$;
- $\mathrm{tr}\,\alpha \coloneqq \alpha + \hat{\alpha}$ satisfies $\mathrm{tr}\,\alpha = \mathrm{tr}\,\hat{\alpha}$ and $\mathrm{tr}(\alpha + \beta) = \mathrm{tr}\,\alpha + \mathrm{tr}\,\beta$;
- $\mathrm{tr}\,\alpha = \deg \alpha + 1 - \deg(\alpha - 1) \in \mathbb{Z}$ for all $\alpha \in \mathrm{End}(E)$;
- α and $\hat{\alpha}$ are the roots of the characteristic equation $x^2 - (\mathrm{tr}\,\alpha)x + \deg \alpha \in \mathbb{Z}[x]$.

These facts imply that the map $\varphi \mapsto \hat{\varphi}$ is an *involution* of $\mathrm{End}(E)$.

Definition 12.1. An *anti-homomorphism* $\varphi\colon R \to S$ of rings is a homomorphism of their additive groups that satisfies $\varphi(1_R) = 1_S$ and $\varphi(\alpha\beta) = \varphi(\beta)\varphi(\alpha)$ for all $\alpha, \beta \in R$. An *involution* (or *anti-involution*) is an anti-homomorphism $\varphi\colon R \to R$ that is its own inverse: $\varphi \circ \varphi$ is the identity map.

A nontrivial involution of a commutative ring is an automorphism of order 2.

12.1 The endomorphism algebra of an elliptic curve

The additive group of $\mathrm{End}(E)$, like all abelian groups, is a $\mathbb{Z}$-module. Recall that if R is a commutative ring, an *R-module* M is an (additively written) abelian group that admits a scalar multiplication by R compatible with its structure as an abelian group. This means that for all $\alpha, \beta \in M$ and $r, s \in R$ we have

$$(r+s)\alpha = r\alpha + s\alpha, \qquad r\alpha + r\beta = r(\alpha + \beta), \qquad r(s\alpha) = (rs)\alpha, \qquad 1\alpha = \alpha$$

(one can check these conditions also imply $0\alpha = 0$ and $(-1)\alpha = -\alpha$).

The ring $\mathrm{End}(E)$ is not only a $\mathbb{Z}$-module. Like all rings, it has a multiplication that is compatible with its structure as a $\mathbb{Z}$-module, making it a $\mathbb{Z}$-algebra. For any commutative ring R, an (associative unital) *R-algebra* A is a (not necessarily commutative) ring equipped with a ring homomorphism $R \to A$ that maps R into the center of A.[1] In our situation the map $\mathbb{Z} \to \mathrm{End}(E)$ sending n to $[n]$ is injective and we simply view $\mathbb{Z}$ as a subring of $\mathrm{End}(E)$ that necessarily lies in its center. When we have a ring A with an involution that is also an R-algebra, we typically require the involution to fix R, so that we may view it as an R-algebra involution; this holds for the involution $\alpha \mapsto \hat{\alpha}$ on our $\mathbb{Z}$-algebra $\mathrm{End}(E)$.

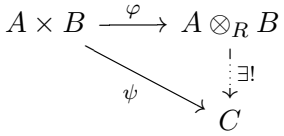
We now want to "upgrade" our $\mathbb{Z}$-algebra $\mathrm{End}(E)$ to a $\mathbb{Q}$-algebra (in other words, a $\mathbb{Q}$-vector space with a multiplication that is compatible with its structure as a vector space), To do this we take the tensor product of $\mathrm{End}(E)$ with $\mathbb{Q}$.

Definition 12.2. The *endomorphism algebra* of E is $\mathrm{End}^0(E) := \mathrm{End}(E) \otimes_{\mathbb{Z}} \mathbb{Q}$.

Recall that for a commutative ring R, the *tensor product* $A \otimes_R B$ of two R-modules A and B can be defined as the R-module generated by the formal symbols $\alpha \otimes \beta$ with $\alpha \in A$ and $\beta \in B$, subject to the relations

$$(\alpha_1+\alpha_2)\otimes\beta = \alpha_1\otimes\beta+\alpha_2\otimes\beta, \quad \alpha\otimes(\beta_1+\beta_2) = \alpha\otimes\beta_1+\alpha\otimes\beta_2, \quad r\alpha\otimes\beta = \alpha\otimes r\beta = r(\alpha\otimes\beta),$$

for $\alpha_1, \alpha_2 \in A$, $\beta_1, \beta_2 \in B$ and $r \in R$. The elements of $A \otimes_R B$ are finite sums of *pure tensors* $\alpha \otimes_R \beta$. We can use the relations above to simplify these sums. In general not every element of $A \otimes_R B$ can be reduced to a pure tensor, but in our situation this is in fact the case (see Lemma 12.5 below). The tensor product behaves quite differently than the direct product (for example, $A \times 0 = A$ but $A \otimes_R 0 = 0$), but we do have a canonical R-bilinear map $\varphi\colon A \times B \to A \otimes_R B$ defined by $(\alpha, \beta) \mapsto \alpha \otimes \beta$. This map is universal in the sense that every R-bilinear map of R-modules $\psi\colon A \times B \to C$ can be written uniquely as a composition



This *universal property* can also be taken as a definition of the tensor product (without guaranteeing its existence).

When A and B are not only R-modules but R-algebras, we give the tensor product $A \otimes_R B$ the structure of an R-algebra by defining multiplication of pure tensors

$$(\alpha_1 \otimes \beta_1)(\alpha_2 \otimes \beta_2) = \alpha_1\alpha_2 \otimes \beta_1\beta_2$$

and extending linearly; this means we can compute $(\sum_i \alpha_i \otimes \beta_i)(\sum_j \alpha_j \otimes \beta_j)$ using the distributive law. The multiplicative identity is necessarily $1_A \otimes 1_B$. The R-algebras A and B can be canonically mapped to $A \otimes_R B$ via $\alpha \mapsto \alpha \otimes 1_B$ and $\beta \mapsto 1_A \otimes \beta$. These maps need not be injective; indeed, $A \otimes_R B$ may be the zero ring even when A and B are not.

Example 12.3. The tensor product $\mathbb{Z}/2\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/3\mathbb{Z}$ is the zero ring. To see why, note that for any pure tensor $\alpha \otimes \beta$ we have

$$\alpha \otimes \beta = \alpha \otimes -2\beta = 2\alpha \otimes -\beta = 0 \otimes -\beta = 0 \otimes 0 = 0.$$

[1]Here we consider only associative unital algebras; one can define a more general notion of an R-algebra that is not necessarily a ring (Lie algebras, for example).

Example 12.4. If V is a k-vector space with basis $(v_1, \dots, v_n)$ and L/k is any field extension, then $V \otimes_k L$ is an L-vector space with basis $(v_1 \otimes 1, \dots, v_n \otimes 1)$; multiplication by scalars in L takes place on the RHS of each pure tensor. This implies that if V is a k-algebra of k-dimension n, then $V \otimes_k L$ is an L-algebra of L-dimension n.

Lemma 12.5. *Let R be an integral domain with fraction field B, and let A be an R-algebra. Every element of $A \otimes_R B$ can be written as a pure tensor $\alpha \otimes \beta$.*

Proof. It suffices to show that $\alpha_1 \otimes \beta_1 + \alpha_2 \otimes \beta_2$ can be written as $\alpha_3 \otimes \beta_3$. Let $\beta_1 = r_1/s_1$ and $\beta_2 = r_2/s_2$ with $r_1, r_2, s_1, s_2 \in R$. Then

$$\begin{aligned}
\alpha_1 \otimes \beta_1 + \alpha_2 \otimes \beta_2 &= \alpha_1 \otimes \frac{r_1}{s_1} + \alpha_2 \otimes \frac{r_2}{s_2} \\
&= \alpha_1 \otimes \frac{r_1 s_2}{s_1 s_2} + \alpha_2 \otimes \frac{r_2 s_1}{s_1 s_2} \\
&= (r_1 s_2 \alpha_1) \otimes \frac{1}{s_1 s_2} + (r_2 s_1 \alpha_2) \otimes \frac{1}{s_1 s_2} \\
&= (r_1 s_2 \alpha_1 + r_2 s_1 \alpha_2) \otimes \frac{1}{s_1 s_2},
\end{aligned}$$

so we may take $\alpha_3 = r_1 s_2 \alpha_1 + r_2 s_1 \alpha_2$ and $\beta_3 = 1/(s_1 s_2)$. □

The lemma implies that every element of $\mathrm{End}^0(E) = \mathrm{End}(E) \otimes_{\mathbb{Z}} \mathbb{Q}$ can be written as $\phi \otimes r$ for some $\phi \in \mathrm{End}(E)$ and $r \in \mathbb{Q}$; to simplify notation we will simply use $r\phi$ to denote $\phi \otimes r$. Note that this representation is not unique (if $r' = r/n$ and $\phi' = n\phi$ then $r'\phi' = r\phi$). The only difference between $r\phi$, with $r \in \mathbb{Q}$, and $n\phi$, with $n \in \mathbb{Z}$, is that the former is not necessarily an endomorphism, but if we multiply $r\phi$ by the denominator of r we will get an element of $\mathrm{End}(E)$ that corresponds to an endomorphism.

The canonical homomorphisms $\mathrm{End}(E) \to \mathrm{End}^0(E)$ and $\mathbb{Q} \to \mathrm{End}^0(E)$ are injective, because $\mathrm{End}(E)$ and $\mathbb{Q}$ are torsion-free $\mathbb{Z}$-algebras, so we may identify both $\mathrm{End}(E)$ and $\mathbb{Q}$ with corresponding subrings of $\mathrm{End}^0(E)$ that intersect in $\mathbb{Z}$. Every element of $\mathrm{End}^0(E)$ has an integer multiple that lies in the subring $\mathrm{End}(E)$, and the subring $\mathbb{Q}$ lies in the center of $\mathrm{End}^0(E)$, which makes $\mathrm{End}^0(E)$ a $\mathbb{Q}$-algebra. We also note that $\mathrm{End}^0(E)$ has no zero divisors: if $(r\phi)(r'\phi') = rr'\phi\phi' = 0$ then either $rr' = 0$ or $\phi\phi' = 0$, so one of r, r', ϕ, ϕ' is zero (since $\mathbb{Q}$ and $\mathrm{End}(E)$ have no zero divisors); this implies that one of $r\phi$ or $r'\phi'$ is zero.

12.2 The Rosati involution and the reduced norm and trace

We now extend the involution $\alpha \mapsto \hat{\alpha}$ on $\mathrm{End}(E)$ to $\mathrm{End}^0(E)$ by defining $\widehat{r\alpha} = r\hat{\alpha}$ for all $r \in \mathbb{Q}$. This implies that $\hat{r} = r$ for all $r \in \mathbb{Q}$ (take $\alpha = 1$), and therefore $\hat{\hat{\alpha}} = \alpha$ holds for all $\alpha \in \mathrm{End}^0(E)$. We also have $\widehat{\alpha\beta} = \hat{\beta}\hat{\alpha}$ and $\widehat{\alpha + \beta} = \hat{\alpha} + \hat{\beta}$ for all $\alpha, \beta \in \mathrm{End}^0(E)$, since these hold for elements of $\mathrm{End}(E)$ and scalars are fixed by $\alpha \mapsto \hat{\alpha}$ and commute. Thus the map $\alpha \mapsto \hat{\alpha}$ is an involution of the $\mathbb{Q}$-algebra $\mathrm{End}^0(E)$, and it is known as the *Rosati involution*.

The Rosati involution allows us to extend the notions of degree and trace on $\mathrm{End}(E)$ to a norm and a trace defined on all of $\mathrm{End}^0(E)$.

Definition 12.6. Let $\alpha \in \mathrm{End}^0(E)$. The (reduced) *norm* of α is $\mathrm{N}\alpha = \alpha\hat{\alpha}$ and the (reduced) *trace* of α is $\mathrm{T}\alpha = \alpha + \hat{\alpha}$.[2]

[2] $\mathrm{N}\alpha$ and $\mathrm{T}\alpha$ are often called the reduced norm and reduced trace and may be denoted $\mathrm{Nrd}\,\alpha$ and $\mathrm{Trd}\,\alpha$ to distinguish them from the more general notion of norm and trace in a $\mathbb{Q}$-algebra, which involve taking the determinant or trace of the $\mathbb{Q}$-linear transformation $\beta \mapsto \alpha\beta$ (this coincides with the reduced norm and trace when $\dim_{\mathbb{Q}} \mathrm{End}^0(E) = 2$, but not otherwise). We shall only consider the reduced norm and trace.

We now show that $\mathrm{N}\alpha$ and $\mathrm{T}\alpha$ lie in $\mathbb{Q}$, and prove some other facts we will need.

Lemma 12.7. *For all $\alpha \in \mathrm{End}^0(E)$ we have $\mathrm{N}\alpha \in \mathbb{Q}_{\geq 0}$, with $\mathrm{N}\alpha = 0$ if and only if $\alpha = 0$. We also have $\mathrm{N}\hat{\alpha} = \mathrm{N}\alpha$ and $\mathrm{N}(\alpha\beta) = \mathrm{N}\alpha\mathrm{N}\beta$ for all $\alpha, \beta \in \mathrm{End}^0(E)$.*

Proof. Write $\alpha = r\phi$, with $r \in \mathbb{Q}$, $\phi \in \mathrm{End}(E)$. Then $\mathrm{N}\alpha = \alpha\hat{\alpha} = r^2 \deg\phi \geq 0$. If r or ϕ is zero then $\alpha = 0$ and $\mathrm{N}\alpha = 0$, and otherwise $\mathrm{N}\alpha > 0$. We have $\alpha\mathrm{N}\hat{\alpha} = \alpha\hat{\alpha}\alpha = (\mathrm{N}\alpha)\alpha = \alpha\mathrm{N}\alpha$, so $\mathrm{N}\hat{\alpha} = \mathrm{N}\alpha$ when $\alpha \neq 0$ (since $\mathrm{End}^0(E)$ has no zero divisors), and $\mathrm{N}\hat{\alpha} = \mathrm{N}\alpha = 0$ when $\alpha = 0$. Finally, for any $\alpha, \beta \in \mathrm{End}^0(E)$ we have

$$\mathrm{N}(\alpha\beta) = \alpha\beta\widehat{\alpha\beta} = \alpha\beta\hat{\beta}\hat{\alpha} = \alpha(\mathrm{N}\beta)\hat{\alpha} = \alpha\hat{\alpha}\mathrm{N}\beta = \mathrm{N}\alpha\mathrm{N}\beta. \qquad \square$$

Corollary 12.8. *Every nonzero $\alpha \in \mathrm{End}^0(E)$ has a multiplicative inverse α^{-1}.*

Proof. If we put $\beta = \hat{\alpha}/\mathrm{N}\alpha$, then $\alpha\beta = \mathrm{N}\alpha/\mathrm{N}\alpha = 1$ and $\beta\alpha = \mathrm{N}\hat{\alpha}/\mathrm{N}\alpha = 1$, so $\beta = \alpha^{-1}$. $\square$

The corollary implies that $\mathrm{End}^0(E)$ is a *division ring*; it satisfies all the field axioms except that multiplication need not be commutative. This means that $\mathrm{End}^0(E)$ is a field if and only if it is commutative.

Lemma 12.9. *For all $\alpha \in \mathrm{End}^0(E)$ we have $\mathrm{T}\hat{\alpha} = \mathrm{T}\alpha \in \mathbb{Q}$. For any $r \in \mathbb{Q}$, $\alpha, \beta \in \mathrm{End}^0(E)$ we have $\mathrm{T}(\alpha + \beta) = \mathrm{T}\alpha + \mathrm{T}\beta$, and $\mathrm{T}(r\alpha) = r\mathrm{T}\alpha$.*

Proof. We first note that $\mathrm{T}\hat{\alpha} = \hat{\alpha} + \hat{\hat{\alpha}} = \hat{\alpha} + \alpha = \alpha + \hat{\alpha} = \mathrm{T}\alpha$, and

$$\mathrm{T}\alpha = \alpha + \hat{\alpha} = 1 + \alpha\hat{\alpha} - (1-\alpha)(1-\hat{\alpha}) = 1 + \mathrm{N}\alpha - \mathrm{N}(1-\alpha) \in \mathbb{Q}.$$

We also have

$$\mathrm{T}(\alpha+\beta) = \alpha + \beta + \widehat{\alpha+\beta} = \alpha + \beta + \hat{\alpha} + \hat{\beta} = \alpha + \hat{\alpha} + \beta + \hat{\beta} = \mathrm{T}\alpha + \mathrm{T}\beta.$$

and

$$\mathrm{T}(r\alpha) = r\alpha + \widehat{r\alpha} = r\alpha + \hat{\alpha}\hat{r} = r\alpha + \hat{\alpha}r = r\alpha + r\hat{\alpha} = r(\alpha + \hat{\alpha}) = r\mathrm{T}\alpha,$$

since $\mathbb{Q}$ lies in the center of $\mathrm{End}^0(E)$ and is fixed by the Rosati involution. $\square$

Lemma 12.10. *Let $\alpha \in \mathrm{End}^0(E)$. Then α and $\hat{\alpha}$ are roots of the polynomial*

$$x^2 - (\mathrm{T}\alpha)x + \mathrm{N}\alpha \in \mathbb{Q}[x].$$

Proof. We have

$$0 = (\alpha - \alpha)(\alpha - \hat{\alpha}) = \alpha^2 - \alpha(\alpha + \hat{\alpha}) + \alpha\hat{\alpha} = \alpha^2 - (\mathrm{T}\alpha)\alpha + \mathrm{N}\alpha,$$

and similarly for $\hat{\alpha}$, since $\mathrm{T}\hat{\alpha} = \mathrm{T}\alpha$ and $\mathrm{N}\hat{\alpha} = \mathrm{N}\alpha$. $\square$

Corollary 12.11. *For any nonzero $\alpha \in \mathrm{End}^0(E)$, if $\mathrm{T}\alpha = 0$ then $\alpha^2 = -\mathrm{N}\alpha < 0$. An element $\alpha \in \mathrm{End}^0(E)$ is fixed by the Rosati involution if and only if $\alpha \in \mathbb{Q}$.*

Proof. The first statement follows immediately from $\alpha^2 - (\mathrm{T}\alpha)\alpha + \mathrm{N}\alpha = 0$. For the second, we have $\hat{r} = r$ for $r \in \mathbb{Q}$, and if $\hat{\alpha} = \alpha$ then $\mathrm{T}\alpha = \alpha + \hat{\alpha} = 2\alpha$, so $\alpha = (\mathrm{T}\alpha)/2 \in \mathbb{Q}$. $\square$

12.3 Quaternion algebras

Before we can give a complete classification of the possible endomorphism algebras $\mathrm{End}^0(E)$ that can arise, we need to introduce quaternion algebras.

Definition 12.12. A *quaternion algebra* over a field k is a k-algebra that has a k-basis of the form $\{1, \alpha, \beta, \alpha\beta\}$, with $\alpha^2, \beta^2 \in k^\times$ and $\alpha\beta = -\beta\alpha$.

Let H be a quaternion algebra over a field k. Then H is a 4-dimensional k-vector space with basis $\{1, \alpha, \beta, \alpha\beta\}$, and we may distinguish the subspace $k \subseteq H$ spanned by 1, which does not depend on the choice of α and β. The complementary subspace H_0 (spanned by $\alpha, \beta, \alpha\beta$) is the space of *pure quaternions*. Every $\gamma \in H$ has a unique decomposition of the form $a + \gamma_0$ with $a \in k$ and $\gamma_0 \in H_0$. The element $\hat{\gamma} := a - \gamma_0$ is the *conjugate* of γ. If γ is a pure quaternion then $\hat{\gamma} = -\gamma$, and for $\gamma \in k$ we have $\hat{\gamma} = \gamma$.

The map $\gamma \mapsto \hat{\gamma}$ is an involution of the k-algebra H, and we define the (reduced) trace $\mathrm{T}\gamma := \gamma + \hat{\gamma}$ and (reduced) norm $\mathrm{N}\gamma := \gamma\hat{\gamma}$, both of which lie in k. It is easy to check that $\mathrm{T}\gamma = \mathrm{T}\hat{\gamma}$ and $\mathrm{N}\gamma = \mathrm{N}\hat{\gamma}$, the trace is additive, the norm is multiplicative, and for $a \in k$ we have $\mathrm{T}a = 2a$ and $\mathrm{N}a = a^2$.

Lemma 12.13. *A quaternion algebra is a division ring if and only if* $\mathrm{N}\gamma = 0$ *implies* $\gamma = 0$.

Proof. Let γ be a nonzero element of a quaternion algebra H. Then $\hat{\gamma} \neq 0$ (since $\hat{0} = 0 \neq \gamma$) If H is a division ring, then γ has an inverse γ^{-1} and $\gamma^{-1}\mathrm{N}\gamma = \gamma^{-1}\gamma\hat{\gamma} = \hat{\gamma} \neq 0$, so $\mathrm{N}\gamma \neq 0$. Conversely, if $\mathrm{N}\gamma \neq 0$ then $\gamma(\hat{\gamma}/\mathrm{N}\gamma) = 1$ and $(\hat{\gamma}/\mathrm{N}\gamma)\gamma = 1$, so γ has an inverse $\hat{\gamma}/\mathrm{N}\gamma$, which implies that H is a division ring. □

Example 12.14. The most well known example of a quaternion algebra is the ring of *Hamilton quaternions* (or *Hamiltonians*) $\mathbb{H}$: the $\mathbb{R}$-algebra with basis $\{1, i, j, ij\}$, where $i^2 = j^2 = -1$ and $ij = -ji$ (the product ij is often denoted k). This was the first example of a noncommutative division ring and has many applications in mathematics and physics.

Remark 12.15. The polynomial $x^2 + 1$ has infinitely many roots in $\mathbb{H}$ (one can take any $x = bi + cj + dk$ with $b^2 + c^2 + d^2 = 1$).

Example 12.16. Let $H = \mathrm{M}_2(k)$ be the ring of 2×2 matrices over a field k with

$$\alpha := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \beta := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \alpha\beta = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \beta\alpha = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

then $\alpha^2 = \beta^2 = 1 \in k^\times$ and $\alpha\beta = -\beta\alpha$, so H is a quaternion algebra, but it is not a division ring, by Lemma 12.13, since $\mathrm{N}(1+\alpha) = (1+\alpha)(1-\alpha) = 0$ but $1 + \alpha \neq 0$. Every quaternion algebra that is not a division ring arises in this way. Such quaternion algebras are said to be *split*, while those that are division rings are called *non-split*.

12.4 Classification theorem for endomorphism algebras

Theorem 12.17. *Let E/k be an elliptic curve. Then* $\mathrm{End}^0(E)$ *is isomorphic to one of:*

(i) *the field of rational numbers $\mathbb{Q}$;*

(ii) *an imaginary quadratic field $\mathbb{Q}(\alpha)$ with $\alpha^2 < 0$;*

(iii) *a quaternion algebra $\mathbb{Q}(\alpha, \beta)$ with $\alpha^2, \beta^2 < 0$.*

Proof. We always have $\mathbb{Q} \subseteq \mathrm{End}^0(E)$, and if $\mathbb{Q} = \mathrm{End}^0(E)$ we are in case (i).

Otherwise, let α be an element of $\mathrm{End}^0(E)$ not in $\mathbb{Q}$. By replacing α with $\alpha - \frac{1}{2}\mathrm{T}\alpha$, we may assume without loss of generality that $\mathrm{T}\alpha = 0$, since

$$\mathrm{T}\left(\alpha - \frac{1}{2}\mathrm{T}\alpha\right) = \mathrm{T}\alpha - \frac{1}{2}\mathrm{TT}\alpha = \mathrm{T}\alpha - \frac{1}{2}2\mathrm{T}\alpha = 0,$$

where $\mathrm{TT}\alpha = 2\mathrm{T}\alpha$ because $\mathrm{T}\alpha \in \mathbb{Q}$. Now $\alpha^2 < 0$, by Corollary 12.11, and $\mathbb{Q}(\alpha) \subseteq \mathrm{End}^0(E)$ is an imaginary quadratic field. If $\mathbb{Q}(\alpha) = \mathrm{End}^0(E)$ then we are in case (ii).

Otherwise, let β be an element of $\mathrm{End}^0(E)$ not in $\mathbb{Q}(\alpha)$. As with α, we may assume without loss of generality that $\mathrm{T}\beta = 0$ so that $\beta^2 < 0$. By replacing β with

$$\beta - \frac{\mathrm{T}(\alpha\beta)}{2\alpha^2}\alpha \tag{1}$$

we can also assume $\mathrm{T}(\alpha\beta) = 0$ (to check, multiply (1) by α and compute the trace; replacing β with (1) does not change its trace because $\mathrm{T}\alpha = 0$). Thus $\mathrm{T}\alpha = \mathrm{T}\beta = \mathrm{T}(\alpha\beta) = 0$. This implies $\alpha = -\hat{\alpha}$, $\beta = -\hat{\beta}$, and $\alpha\beta = -\widehat{\alpha\beta} = -\hat{\beta}\hat{\alpha}$. Substituting the first two equalities into the third yields $\alpha\beta = -\beta\alpha$. Applying this together with the fact that $\alpha^2 < 0$ and $\beta^2 < 0$ lie in $\mathbb{Q}$, it is clear that $\{1, \alpha, \beta, \alpha\beta\}$ spans $\mathbb{Q}(\alpha, \beta)$ as a $\mathbb{Q}$-vector space.

To show that $\mathbb{Q}(\alpha, \beta)$ is a quaternion algebra, we need to show that 1, α, β, and $\alpha\beta$ are $\mathbb{Q}$-linearly independent. By construction, 1, α, β are linearly independent: note $\beta \notin \mathbb{Q}(\alpha)$ implies $\alpha \notin \mathbb{Q}(\beta)$, since $\mathbb{Q}(\beta) = \{r + s\beta : r, s \in \mathbb{Q}\}$ (because $\beta^2 \in \mathbb{Q}$). Now suppose for the sake of contradiction that

$$\alpha\beta = a + b\alpha + c\beta,$$

for some $a, b, c \in \mathbb{Q}$. Taking traces on both sides yields $\mathrm{T}(a) = 0$, so $a = 0$ (since $a \in \mathbb{Q}$). But then $\alpha(\beta - b) = c\beta$, which is impossible for $\mathbb{Q}$-linearly independent α, β.

Thus $\mathbb{Q}(\alpha, \beta) \subseteq \mathrm{End}^0(E)$ is a quaternion algebra with $\alpha^2, \beta^2 < 0$. If $\mathbb{Q}(\alpha, \beta) = \mathrm{End}^0(E)$ then we are in case (iii).

Otherwise, let γ be an element of $\mathrm{End}^0(E)$ that does not lie in $\mathbb{Q}(\alpha, \beta)$. As with β, we may assume without loss of generality that $\mathrm{T}\gamma = 0$ and $\mathrm{T}(\alpha\gamma) = 0$, which implies $\alpha\gamma = -\gamma\alpha$. Then $\alpha\beta\gamma = -\beta\alpha\gamma = \beta\gamma\alpha$, so α commutes with $\beta\gamma$. By Lemma 12.18 below, $\beta\gamma \in \mathbb{Q}(\alpha)$. This implies $\gamma \in \mathbb{Q}(\alpha, \beta)$, contrary to our assumption that $\gamma \notin \mathbb{Q}(\alpha, \beta)$. □

Lemma 12.18. *If $\alpha, \beta \in \mathrm{End}^0(E)$ commute and $\alpha \notin \mathbb{Q}$ then $\beta \in \mathbb{Q}(\alpha)$.*

Proof. As in the proof of Theorem 12.17, we can transform α and β so that $\mathrm{T}\alpha = \mathrm{T}\beta = \mathrm{T}(\alpha\beta) = 0$, and therefore $\alpha\beta = -\beta\alpha$; this involves replacing α with $\alpha - r$ and then replacing β with $\beta - s - t\alpha$ for some $r, s, t \in \mathbb{Q}$; if α and β commute then so do all $\mathbb{Q}$-linear combinations, so the hypothesis still holds. We then have $\alpha\beta + \beta\alpha = 2\alpha\beta = 0$, which implies $\alpha = 0$ or $\beta = 0$, since $\mathrm{End}^0(E)$ has no zero divisors. We cannot have $\alpha = 0$, since $\alpha \notin \mathbb{Q}$, so $\beta = 0 \in \mathbb{Q}(\alpha)$. □

Remark 12.19. In the proofs of Theorem 12.17 and Lemma 12.18 we never used the fact that $\mathrm{End}^0(E)$ is the endomorphism algebra of an elliptic curve. Indeed, one can replace $\mathrm{End}^0(E)$ with any $\mathbb{Q}$-algebra A possessing an involution $\alpha \mapsto \hat{\alpha}$ that fixes $\mathbb{Q}$ such that the associated norm $\mathrm{N}\alpha = \alpha\hat{\alpha}$ maps nonzero elements of A to positive elements of $\mathbb{Q}$; all other properties of $\mathrm{End}^0(E)$ that we used can be derived from these.

Having classified the possible endomorphism algebras $\mathrm{End}^0(E)$, our next task is to classify the possible endomorphism rings $\mathrm{End}(E)$. We begin with the following corollary to Theorem 12.17.

Corollary 12.20. *Let E/k be an elliptic curve. The endomorphism ring $\mathrm{End}(E)$ is a free $\mathbb{Z}$-module of rank r, where $r = 1, 2, 4$ is the dimension of $\mathrm{End}^0(E)$ as a $\mathbb{Q}$-vector space.*

Recall that a free $\mathbb{Z}$-module of rank r is an abelian group isomorphic to $\mathbb{Z}^r$.

Proof. Let us pick a basis $\{e_1, \dots, e_r\}$ for $\mathrm{End}^0(E)$ as a $\mathbb{Q}$-basis with the property that $\mathrm{T}(e_i e_j) = 0$ unless $i = j$ (use the basis $\{1, \alpha\}$ when $\mathrm{End}^0(E) = \mathbb{Q}(\alpha)$ and $\{1, \alpha, \beta, \alpha\beta\}$ when $\mathrm{End}^0(E) = \mathbb{Q}(\alpha, \beta)$, where α and β are constructed as in the proof of Theorem 12.17). After multiplying by suitable integers if necessary, we can assume without loss of generality that $e_1, \dots e_r \in \mathrm{End}(E)$ (this doesn't change $\mathrm{T}(e_i e_j) = 0$ for $i \neq j$).

For any $\mathbb{Z}$-module $A \subseteq \mathrm{End}^0(E)$ we have an associated *dual* $\mathbb{Z}$-module

$$A^* := \{\alpha \in \mathrm{End}^0(E) : \mathrm{T}(\alpha\phi) \in \mathbb{Z} \ \forall \phi \in A\}.$$

Note that A^* is closed under addition and multiplication by integers (if $\mathrm{T}(\alpha\phi), \mathrm{T}(\beta\phi) \in \mathbb{Z}$ then $\mathrm{T}(m\alpha\phi + n\beta\phi) \in \mathbb{Z}$ for all $m, n \in \mathbb{Z}$), so A^* is also a $\mathbb{Z}$-module. It is clear from the definition that if A and B are any $\mathbb{Z}$-modules in $\mathrm{End}^0(E)$, then $A \subseteq B$ implies $B^* \subseteq A^*$ (making A bigger imposes a stronger constraint on A^*).

Now let A be the $\mathbb{Z}$-module spanned by $e_1, \dots, e_r \in \mathrm{End}(E)$. Then $A \subseteq \mathrm{End}(E)$, and therefore $\mathrm{End}(E)^* \subseteq A^*$. We also note that $\mathrm{End}(E) \subseteq \mathrm{End}(E)^*$, since $\mathrm{T}(\alpha\phi) \in \mathbb{Z}$ for all $\alpha, \phi \in \mathrm{End}(E)$. Thus

$$A \subseteq \mathrm{End}(E) \subseteq \mathrm{End}(E)^* \subseteq A^*.$$

We can write any $\alpha \in A^* \subseteq \mathrm{End}^0(E)$ as $a_1 e_1 + \cdots + a_r e_r$ for some $a_1, \dots, a_r \in \mathbb{Q}$ (since $e_1, \dots, e_r$ is a $\mathbb{Q}$-basis for $\mathrm{End}^0(E)$). For each e_i we then have

$$\mathrm{T}(\alpha e_i) = a_1 \mathrm{T}(e_1 e_i) + \cdots + a_r \mathrm{T}(e_r e_i) = a_i \mathrm{T}(e_i^2),$$

since $\mathrm{T}(e_i e_j) = 0$ for $i \neq j$, and $\mathrm{T}(\alpha e_i) = a_i \mathrm{T}(e_i^2) \in \mathbb{Z}$ since $\alpha \in A^*$ and $e_i \in A$. Thus a_i is an integer multiple of $1/\mathrm{T}(e_i^2)$, and it follows that $\{e_1/\mathrm{T}(e_1^2), \dots, e_r/\mathrm{T}(e_r^2)\}$ is a basis for A^* as a $\mathbb{Z}$-module, which is therefore a free $\mathbb{Z}$-module of rank r, as is A (both are torsion free because $\mathrm{End}^0(E)$ is torsion free). It follows that $\mathrm{End}(E)$ and $\mathrm{End}(E)^*$ both free $\mathbb{Z}$-modules of rank r, since they are both contained in and contain a free $\mathbb{Z}$-module of rank r (every subgroup of $\mathbb{Z}^r$ is isomorphic to $\mathbb{Z}^s$ for some $0 \leq s \leq r$).[3] □

Definition 12.21. An elliptic curve E for which $\mathrm{End}(E) \not\simeq \mathbb{Z}$ is said to have *complex multiplication*.

It follows from Theorem 12.17 that if E has complex multiplication then $\mathrm{End}^0(E)$ is either an imaginary quadratic field or a quaternion algebra. Each element of $\mathrm{End}(E)$ that does not lie in $\mathbb{Z}$ is the root of quadratic polynomial in $\mathbb{Z}[x]$ that has no real roots, which we could view as a complex number (an algebraic integer, in fact). Elements ϕ of $\mathrm{End}(E)$ that lie in $\mathbb{Z}$ correspond to multiplication by some integer n, and we may view elements of $\mathrm{End}(E)$ that do not lie in $\mathbb{Z}$ as "multiplication" by some complex number that corresponds to an algebraic integer that is a root of the characteristic polynomial of ϕ.

[3] More generally, if R is a principal ideal domain (PID) then every submodule of a free R-module of rank r is free of rank $s \leq r$. This fails when R is not a PID (submodules of a free module need not be free)

12.5 Orders in $\mathbb{Q}$-algebras

Definition 12.22. Let K be a $\mathbb{Q}$-algebra of finite dimension r as a $\mathbb{Q}$-vector space. An *order* $\mathcal{O}$ in K is a subring of K that is a free $\mathbb{Z}$-module of rank r. Equivalently, $\mathcal{O}$ is a subring of K that is finitely generated as a $\mathbb{Z}$-module and satisfies $K = \mathcal{O} \otimes_{\mathbb{Z}} \mathbb{Q}$.

Note that an order is required to be both a lattice (a free $\mathbb{Z}$-module of maximal rank) *and* a ring; in particular it must contain 1.

Example 12.23. The integers $\mathbb{Z}$ are the unique example of an order in $\mathbb{Q}$. Non-examples include the even integers, which is a lattice but not a ring, and the set $\{a/2^n : a, n \in \mathbb{Z}\}$, which is a ring but not a lattice (because it is not finitely generated as a $\mathbb{Z}$-module).

It follows from Corollary 12.20 that the endomorphism ring $\mathrm{End}(E)$ is an order in the $\mathbb{Q}$-algebra $\mathrm{End}^0(E)$. Note that if $\mathrm{End}^0(E) = \mathbb{Q}$, then we must have $\mathrm{End}(E) = \mathbb{Z}$, but in general there are infinitely many non-isomorphic possibilities for $\mathrm{End}(E)$.

Every order lies in some *maximal order* (an order that is not contained in any other); this follows from an application of Zorn's lemma, using the fact that elements of an order necessarily have monic minimal polynomials. In general, maximal orders need not be unique, but when the $\mathbb{Q}$-algebra K is a number field (a finite extension of $\mathbb{Q}$), this is the case. In view of Theorem 12.17, we are primarily interested in the case where K is an imaginary quadratic field, but it is just as easy to prove this for all number fields. We first need to recall a few standard results from algebraic number theory.

Definition 12.24. An *algebraic number* α is a complex number that is the root of a polynomial with coefficients in $\mathbb{Q}$. An *algebraic integer* is a complex number that is the root of a monic polynomial with coefficients in $\mathbb{Z}$.

Two fundamental results of algebraic number theory are (1) the set of algebraic integers in a number field form a ring, and (2) every number field has an *integral basis* (a basis whose elements are algebraic integers). The following theorem gives a more precise statement.

Theorem 12.25. *The set of algebraic integers $\mathcal{O}_K$ in a number field K form a ring that is a free $\mathbb{Z}$-module of rank r, where $r = [K : \mathbb{Q}]$ is the dimension of K as a $\mathbb{Q}$-vector space.*

Proof. See Theorem 2.1 and Corollary 2.30 in [1] (or Theorems 2.9 and 2.16 in [3]).[4] □

Theorem 12.26. *The ring of integers $\mathcal{O}_K$ of a number field K is its unique maximal order.*

Proof. The previous theorem implies that $\mathcal{O}_K$ is an order. To show that it is the unique maximal order, we need to show that every order $\mathcal{O}$ in K is contained in $\mathcal{O}_K$. It suffices to show that every $\alpha \in \mathcal{O}$ is an algebraic integer. Viewing $\mathcal{O}$ as a $\mathbb{Z}$-lattice of rank $r = [K : \mathbb{Q}]$, consider the sublattice generated by all powers of α. Let $[\beta_1, \ldots, \beta_r]$ be a basis for this sublattice, where each β_i is a $\mathbb{Z}$-linear combination of powers of α. Let n be an integer larger than any of the exponents in any of the powers of α that appear in any β_i. Then $\alpha^n = c_1\beta_1 + \cdots + c_r\beta_r$, for some $c_1, \ldots, c_r \in \mathbb{Z}$, and this determines a monic polynomial of degree n with α as a root. Therefore α is an algebraic integer. □

Finally, we characterize the orders in imaginary quadratic fields, which are the number fields we are most interested in.

[4] The proof of the second part of this theorem is essentially the same as the proof of Corollary 12.20; instead of the reduced trace in $\mathrm{End}^0(E)$, one uses the trace map from K to $\mathbb{Q}$, which has similar properties.

Theorem 12.27. *Let K be an imaginary quadratic field with ring of integers $\mathcal{O}_K$. The orders $\mathcal{O}$ in K are precisely the subrings $\mathbb{Z} + f\mathcal{O}_K$, where f is any positive integer.*

Proof. The maximal order $\mathcal{O}_K$ is a free $\mathbb{Z}$-module (a lattice) of rank 2 that contains 1, so it has a $\mathbb{Z}$-basis of the form $[1, \tau]$ for some $\tau \notin \mathbb{Z}$. Let $\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$. It is clear that $\mathcal{O}$ is a sub-lattice of $\mathcal{O}_K$ that properly contains $\mathbb{Z}$, hence it is of rank 2. The $\mathbb{Z}$-module $\mathcal{O}$ is a subset of the ring $\mathcal{O}_K$ and contains 1, so to show that $\mathcal{O}$ is a ring it suffices to show that it is closed under multiplication. So let $a + f\alpha$ and $b + f\beta$ be arbitrary elements of $\mathcal{O}$, with $a, b \in \mathbb{Z}$ and $\alpha, \beta \in \mathcal{O}_K$. Then

$$(a + f\alpha)(b + f\beta) = ab + af\beta + bf\alpha + f^2\alpha\beta = ab + f(a\beta + b\alpha + f\alpha\beta) \in \mathcal{O},$$

since $ab \in \mathbb{Z}$ and $(a\beta + b\alpha + f\alpha\beta) \in \mathcal{O}_K$. So $\mathcal{O}$ is a subring of K. To see that $\mathcal{O}$ is an order, note that $\mathcal{O} \otimes_{\mathbb{Z}} \mathbb{Q} = \mathcal{O}_K \otimes_{\mathbb{Z}} \mathbb{Q} = K$.

Now let $\mathcal{O}$ be any order in K. Then $\mathcal{O}$ is a rank-2 sub-lattice of $\mathcal{O}_K = [1, \tau]$ that contains 1, so $\mathcal{O}$ must contain an integer multiple of τ. Let f be the least positive integer for which $f\tau \in \mathcal{O}$. The lattice $[1, f\tau]$ lies in $\mathcal{O}$, and we claim that in fact $\mathcal{O} = [1, f\tau]$. Any element α of $\mathcal{O}$ must lie in $\mathcal{O}_K$ and is therefore of the form $\alpha = a + b\tau$ for some $a, b \in \mathbb{Z}$. The element $b\tau = \alpha - a$ then lies in $\mathcal{O}$, and the minimality of f implies that f divides b. Thus $\mathcal{O} = [1, f\tau] = \mathbb{Z} + f\mathcal{O}_K$. □

Remark 12.28. In the theorem above we never actually used the fact that the quadratic field K is imaginary; in fact, the theorem holds for real quadratic fields as well.

The integer f in Theorem 12.27 is called the *conductor* of the order $\mathcal{O} = \mathbb{Z} + f\mathcal{O}_K$. It is equal to the index $[\mathcal{O}_K : \mathcal{O}]$, which is necessarily finite.

References

[1] J. S. Milne, *Algebraic number theory*, course notes, version 3.06, 2014.

[2] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer 2009.

[3] Ian Stewart and David Tall, *Algebraic number theory and Fermat's last theorem*, third edition, A.K. Peters, 2002.

13 Ordinary and supersingular elliptic curves

Let E/k be an elliptic curve over a field of positive characteristic p. In Lecture 6 we proved that for any nonzero integer n, the multiplication-by-n map $[n]$ is separable if and only if n is not divisible by p. This implies that the separable degree of the multiplication-by-p map cannot be $p^2 = \deg[p]$, it must be either p or 1, meaning that its kernel $E[p]$ is either cyclic of order p or trivial. The terms *ordinary* and *supersingular* distinguish these two cases:

$$\begin{array}{lcl} E \text{ is ordinary} & \iff & E[p] \simeq \mathbb{Z}/p\mathbb{Z}. \\ E \text{ is supersingular} & \iff & E[p] = \{0\}. \end{array}$$

We now want to explore this distinction further, and relate it to our classification of endomorphism algebras. In the previous lecture we showed that $\mathrm{End}^0(E) := \mathrm{End}(E) \otimes_{\mathbb{Z}} \mathbb{Q}$ has dimension $1, 2$, or 4 as a $\mathbb{Q}$-vector space, depending on whether $\mathrm{End}^0(E)$ is isomorphic to $\mathbb{Q}$, an imaginary quadratic field, or a quaternion algebra.

Before we begin, let us recall some facts about isogenies proved in Lectures 5 and 6. We assume throughout that we are working in a field k of positive characteristic p.

1. Any isogeny α can be decomposed as $\alpha = \alpha_{\mathrm{sep}} \circ \pi^n$, where α_{sep} is separable, and π is the (purely inseparable) p-power Frobenius map $\pi\colon (x : y : z) \mapsto (x^p : y^p : z^p)$.
2. If $\alpha = \alpha_{\mathrm{sep}} \circ \pi^n$ then $\deg_s \alpha := \deg \alpha_{\mathrm{sep}}$, $\deg_i \alpha := p^n$, and $\deg \alpha = (\deg_s \alpha)(\deg_i \alpha)$.
3. We have $\#\ker \alpha = \deg_s \alpha$ (so E is supersingular if and only if $\deg_s[p] = 1$).
4. We have $\deg(\alpha \circ \beta) = (\deg \alpha)(\deg \beta)$, and similarly for $\deg_s$ and $\deg_i$.
5. A sum of inseparable isogenies is inseparable and the sum of a separable and an inseparable isogeny is separable (a sum of separable isogenies need not be separable).
6. The multiplication-by-n map $[n]$ is inseparable if and only if $p|n$.

Recall that an isogeny α is purely inseparable when $\deg_s \alpha = 1$, equivalently, when $\ker \alpha = \{0\}$. Thus an elliptic curve is supersingular if and only if the multiplication-by-p map $[p]$ is purely inseparable. This makes it clear that the property of being ordinary or supersingular is invariant under base change: if E/k is an elliptic curve over k and L/k is any field extension, the separable degree of $[p]$ on E_L does not depend on L.

Warning 13.1. As noted in the previous lecture, in this course the ring $\mathrm{End}(E)$ consists of endomorphisms defined over k; if we wish to refer to endomorphisms defined over $\bar{k}$ we will write $\mathrm{End}(E_{\bar{k}})$ or refer to the *geometric* endomorphism ring (or algebra). Many authors use $\mathrm{End}(E)$ to denote $\mathrm{End}(E_{\bar{k}})$, but this distinction is important.[1]

The property of being ordinary or supersingular is an isogeny invariant.

Theorem 13.2. *Let $\phi\colon E_1 \to E_2$ be an isogeny of elliptic curves. Then E_1 is supersingular if and only if E_2 is supersingular (and E_1 is ordinary if and only if E_2 is ordinary).*

[1] For example, there are algorithms that apply to any elliptic curve $E/\mathbb{F}_q$ for which $\mathrm{End}(E)$ is an imaginary quadratic field, but one often finds them written under the strictly stronger assumption that E is ordinary.

Proof. Let $p_1 \in \mathrm{End}(E_1)$ and $p_2 \in \mathrm{End}(E_2)$ denote the multiplication-by-p maps on E_1 and E_2, respectively. We have $p_2 \circ \phi = \phi + \cdots + \phi = \phi \circ p_1$, thus

$$\begin{aligned} p_2 \circ \phi &= \phi \circ p_1 \\ \deg_s(p_2 \circ \phi) &= \deg_s(\phi \circ p_1) \\ \deg_s(p_2)\deg_s(\phi) &= \deg_s(\phi)\deg_s(p_1) \\ \deg_s(p_2) &= \deg_s(p_1). \end{aligned}$$

The elliptic curve E_i is supersingular if and only if $\deg_s(p_i) = 1$; the theorem follows. □

In what follows we will often want to refer to the image of E under the p-power Frobenius isogeny $(x : y : z) \mapsto (x^p : y^p : z^p)$ which we shall denote $E^{(p)}$. When E is defined over $\mathbb{F}_p$ we will have $E^{(p)} = E$ and π will be the Frobenius endomorphism π_E, but in general $E^{(p)}$ is the elliptic curve obtained by taking an equation for E and raising each coefficient to the pth power (it does not matter which equation we pick, the curve $E^{(p)}$ is well-defined up to isomorphism). We similarly define $E^{(q)}$ to be the image of the q-power Frobenius isogeny. Note that $[p] = \pi\hat{\pi}$ is purely inseparable if and only if $\hat{\pi}$ is purely inseparable (since π is always purely inseparable), thus E is supersingular if and only if $\hat{\pi}$ is purely inseparable.

In order to simplify the presentation we will often assume $p > 3$ and use short Weierstrass equations $y^2 = x^3 + Ax + B$ to define our elliptic curves, but except for where explicitly noted otherwise, all results in this lecture also hold in characteristic 2 and 3. An advantage of using short Weierstrass equations is that it allows us to put isogenies in our standard form $\left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$, with $u, v, s, t \in k[x]$ chosen so that $u \perp v$ and $s \perp t$.

We also note that $[p] = \pi\hat{\pi}$, where $\hat{\pi}$ is the dual of the p-power Frobenius isogeny π. The multiplicativity of separable degrees implies that $[p]$ is purely inseparable if and only if $\hat{\pi}$ is (since π is always purely inseparable) and $\deg\hat{\pi} = p$ is prime, so $\hat{\pi}$ is purely inseparable if and only if it is inseparable. Thus E is supersingular if and only if $\hat{\pi}$ is inseparable, a fact we will use to shorten the proofs that follow.

13.1 Ordinary/supersingular elliptic curves over finite fields

Theorem 13.3. *An elliptic curve $E/\mathbb{F}_q$ is supersingular if and only if $\operatorname{tr}\pi_E \equiv 0 \bmod p$.*

Proof. If E is supersingular then $[p] = \pi\hat{\pi}$ is purely inseparable, in which case $\hat{\pi}$ is inseparable, as are $\hat{\pi}^n = \widehat{\pi^n} = \hat{\pi}_E$ and $\pi_E = \pi^n$. Their sum $[\operatorname{tr}\pi_E] = \pi_E + \hat{\pi}_E$ is then inseparable, so p must divide $\operatorname{tr}\pi_E$, equivalently, $\operatorname{tr}\pi_E \equiv 0 \bmod p$.

Conversely, if $\operatorname{tr}\pi_E \equiv 0 \bmod p$, then $[\operatorname{tr}\pi_E]$ is inseparable, as is $\hat{\pi}_E = [\operatorname{tr}\pi_E] - \pi_E$. This means that $\hat{\pi}^n$ and therefore $\hat{\pi}$ is inseparable which implies that E is supersingular. □

Corollary 13.4. *Let $E/\mathbb{F}_p$ be an elliptic curve over a field of prime order $p > 3$. Then E is supersingular if and only if $\operatorname{tr}\pi_E = 0$, equivalently, if and only if $\#E(\mathbb{F}_p) = p + 1$.*

Proof. By Hasse's theorem, $|\operatorname{tr}\pi_E| \le 2\sqrt{p}$, and $2\sqrt{p} < p$ for $p > 3$. □

Warning 13.5. Corollary 13.4 does not hold for $p \le 3$; there are supersingular curves over $\mathbb{F}_2$ and $\mathbb{F}_3$ with nonzero Frobenius traces.

This should convince you that supersingular curves over $\mathbb{F}_p$ are rare: there are $\approx 4\sqrt{p}$ possible values for $\operatorname{tr}\pi_E$, all but one of which correspond to ordinary curves.

Theorem 13.6. *Let E be an elliptic curve over a finite field $\mathbb{F}_q$ and suppose $\pi_E \notin \mathbb{Z}$. Then $\mathrm{End}^0(E) = \mathbb{Q}(\pi_E) \simeq \mathbb{Q}(\sqrt{D})$ is an imaginary quadratic field with $D = (\mathrm{tr}\,\pi_E)^2 - 4q$. This applies in particular whenever q is prime, and also whenever E is ordinary.*

Proof. The Frobenius endomorphism π_E is a root of its characteristic polynomial

$$x^2 - (\mathrm{tr}\,\pi_E)x + \deg \pi_E,$$

with discriminant $D = (\mathrm{tr}\,\pi_E)^2 - 4\deg\pi_E = (\mathrm{tr}\,\pi_E)^2 - 4q$, so $\mathbb{Q}(\pi_E) \simeq \mathbb{Q}(\sqrt{D})$. The assumption $\pi_E \notin \mathbb{Z}$ implies $\pi_E \notin \mathbb{Q}$, since π_E is an algebraic integer, and that $\mathrm{tr}(\pi_E)^2 \neq 4q$, so $D < 0$ (by the Hasse bound) and $\mathbb{Q}(\pi_E)$ is an imaginary quadratic field.

We can write any $\alpha \in \mathrm{End}^0(E)$ as $\alpha = s\phi$ with $s \in \mathbb{Q}$ and $\phi \in \mathrm{End}(E)$. Writing ϕ as $\phi(x,y) = (r_1(x), r_2(x)y)$ in standard form, we have

$$(\phi\pi_E)(x,y) = (r_1(x^q), r_2(x^q)y^q) = (r_1(x)^q, r_2(x)^q y^q) = (\pi_E\phi)(x,y),$$

thus ϕ, and therefore α, commutes with π_E. Therefore $\alpha \in \mathbb{Q}(\pi_E)$, by Lemma 12.18, so $\mathrm{End}^0(E) = \mathbb{Q}(\pi_E)$ as claimed. □

Corollary 13.7. *Let E be an elliptic curve over $\mathbb{F}_q$ with $q = p^n$. If n is odd or E is ordinary, then $\mathrm{End}^0(E) = \mathbb{Q}(\pi_E) \simeq \mathbb{Q}(\sqrt{D})$ is an imaginary quadratic field with $D = (\mathrm{tr}\,\pi_E)^2 - 4q$.*

Proof. If $\pi_E \in \mathbb{Z}$ then $D = (\mathrm{tr}\,\pi_E)^2 - 4\deg\pi_E = 0$ and $2\sqrt{q} = \pm\,\mathrm{tr}\,\pi_E \in \mathbb{Z}$, which is possible only if q is a square and $\mathrm{tr}\,\pi_E$ is a multiple of p, in which case n is even and E is supersingular. The corollary then follows from Theorem 13.6. □

If $E/\mathbb{F}_q$ is an ordinary elliptic curve, or more generally, whenever $\pi_E \notin \mathbb{Z}$, the subring $\mathbb{Z}[\pi_E]$ of $\mathrm{End}(E)$ generated by π_E is a lattice of rank 2. It follows that $\mathbb{Z}[\pi_E]$ is an order in the imaginary quadratic field $K \coloneqq \mathrm{End}^0(E)$, and is therefore contained in the maximal order $\mathcal{O}_K$ (the ring of integers of K). The endomorphism ring $\mathrm{End}(E)$ need not equal $\mathbb{Z}[\pi_E]$, but the fact that it contains $\mathbb{Z}[\pi_E]$ and is contained in $\mathcal{O}_K$ constrains $\mathrm{End}(E)$ to a finite set of possibilities. Recall from Theorem 12.27 that every order $\mathcal{O}$ in K is characterized by its *conductor* $[\mathcal{O}_K : \mathcal{O}]$.

Theorem 13.8. *Let $E/\mathbb{F}_q$ be an elliptic curve for which $\mathrm{End}^0(E)$ is an imaginary quadratic field K with ring of integers $\mathcal{O}_K$. Then*

$$\mathbb{Z}[\pi_E] \subseteq \mathrm{End}(E) \subseteq \mathcal{O}_K,$$

and the conductor of $\mathrm{End}(E)$ divides $[\mathcal{O}_K : \mathbb{Z}[\pi_E]]$.

Proof. Immediate from the discussion above. □

Remark 13.9. Theorem 13.8 implies that once we know $\mathrm{tr}\,\pi_E$ (which we can compute in polynomial time using Schoof's algorithm), which determines $\mathrm{End}^0(E) \simeq K = \mathbb{Q}(\sqrt{D})$ and the orders $\mathcal{O}_K$ and $\mathbb{Z}[\pi_E]$, we can constrain $\mathrm{End}(E)$ to a finite set of possibilities distinguished by the conductor $f \coloneqq [\mathcal{O}_K : \mathrm{End}(E)]$. No polynomial-time algorithm is known for computing the integer f, but there is a Las Vegas algorithm that has a heuristically subexponential expected running time [1]. This makes it feasible to compute f even when q is of cryptographic size (say $q \approx 2^{256}$).

Remark 13.10. It will often be convenient to identify $\mathrm{End}^0(E)$ with K and $\mathrm{End}(E)$ with an order $\mathcal{O}$ in K. But we should remember that we are actually speaking of isomorphisms. In the case of an imaginary quadratic field, there are two distinct choices for this isomorphism. This choice can be made canonically, see [3, Thm. II.1.1], however this is not particularly relevant to us, as we are going to be working in finite fields where we cannot distinguish the square roots of D in any case. We thus accept the fact that we are making an arbitrary choice when we fix an isomorphism of $\mathrm{End}^0(E)$ with K by identifying π_E with, say, $(t+\sqrt{D})/2$ (as opposed to $(t-\sqrt{D})/2$).

Before leaving the topic of ordinary and supersingular curves, we want to prove a remarkable fact: while over any algebraically closed field there are always infinitely many non-isomorphic elliptic curves, only a finite number can be supersingular. To prove this we first introduce the j-invariant, which will play a critical role in the lectures to come.

13.2 The j-invariant of an elliptic curve

As usual, we shall assume we are working over a field k whose characteristic is not 2 or 3, so that we can put our elliptic curves E/k in short Weierstrass form $y^2 = x^3 + Ax + B$.

Definition 13.11. The *j-invariant* of the elliptic curve $E\colon y^2 = x^3 + Ax + B$ is

$$j(E) = j(A,B) = 1728\frac{4A^3}{4A^3 + 27B^2}.$$

Note that the denominator of $j(E)$ is nonzero, since it is the discriminant of the cubic $x^3 + Ax + B$, which has no repeated roots. There are two special cases worth noting: if $A = 0$ then $j(A,B) = 0$, and if $B = 0$ then $j(A,B) = 1728$ (note that A and B cannot both be zero). The j-invariant can also be defined for elliptic curves in general Weierstrass form, which is necessary to address fields of characteristic 2 and 3; see [2, III.1].[2]

The key property of the j-invariant $j(E)$ is that it characterizes E up to isomorphism over $\bar{k}$. Before proving this we first note that every element of the field k is the j-invariant of an elliptic curve defined over k.

Theorem 13.12. *For every $j_0 \in k$ there is an elliptic curve E/k with j-invariant $j(E) = j_0$.*

Proof. We assume $\mathrm{char}(k) \neq 2, 3$; see [2, III.1.4.c] for a general proof. If j_0 is 0 or 1728 we may take E to be $y^2 = x^3 + 1$ or $y^2 = x^3 + x$, respectively. Otherwise, let E/k be the elliptic curve defined by $y^2 = x^3 + Ax + B$ where

$$A = 3j_0(1728 - j_0),$$
$$B = 2j_0(1728 - j_0)^2.$$

We claim that $j(A,B) = j_0$. We have

$$\begin{aligned} j(A,B) &= 1728\frac{4A^3}{4A^3 + 27B^2} \\ &= 1728\frac{4 \cdot 3^3 j_0^3 (1728 - j_0)^3}{4 \cdot 3^3 j_0^3 (1728 - j_0)^3 + 27 \cdot 2^2 j_0^2 (1728 - j_0)^4} \\ &= 1728\frac{j_0}{j_0 + 1728 - j_0} \\ &= j_0. \end{aligned}$$

□

[2] As noted in the errata, there is a typo on p. 42 of [2]; the equation $b_2 = a_1^2 - 4a_4$ should read $b_2 = a_1^2 - 4a_2$.

We now give a necessary and sufficient condition for two elliptic curves to be isomorphic. An isomorphism ϕ of elliptic curves is an invertible isogeny, equivalently, an isogeny of degree 1 (the dual isogeny gives an inverse isomorphism, since $\phi\hat{\phi} = \hat{\phi}\phi = 1$). Recall from Lecture 5 that an isogeny between elliptic curves that are defined over k is assumed to be defined over k (hence representable by rational functions with coefficients in k), and we say that two elliptic curves are isogenous over an extension L of k to indicate that the isogeny is defined over L (strictly speaking, it is an isogeny between the base changes of the elliptic curves to L). As we saw in problem 3 of Problem Set 1, elliptic curves that are isomorphic over $\bar{k}$ need not be isomorphic over k.

Theorem 13.13. *Elliptic curves $E\colon y^2 = x^3 + Ax + B$ and $E'\colon y^2 = x^3 + A'x + B'$ defined over k are isomorphic (over k) if and only if $A' = \mu^4 A$ and $B' = \mu^6 B$, for some $\mu \in k^\times$.*

Proof. Let $\phi\colon E \to E'$ be an isomorphism in standard form $\phi(x, y) = (r_1(x), r_2(x)y)$ with $r_1, r_2 \in k(x)$. Since ϕ is an isomorphism, its kernel is trivial, so r_1 and r_2 must be polynomials, by Lemma 4.27 and Corollary 4.28. Thus $r_1(x) = ax + b$ for some $a, b \in k$ with $a \neq 0$. Substituting into the curve equation for E', we have

$$\begin{aligned} r_2(x)^2 y^2 &= (ax+b)^3 + A'(ax+b) + B' \\ r_2(x)^2(x^3 + Ax + B) &= (ax+b)^3 + A'(ax+b) + B'. \end{aligned}$$

By comparing the degrees of the polynomials on both sides, we see that $r_2(x)$ must be constant, say $r_2(x) = c$. Comparing coefficients of x^2 shows that $b = 0$, and comparing coefficients of x^3 shows that $c^2 = a^3$; thus $a = (c/a)^2$ and $c = (c/a)^3$. If we let $\mu = c/a \in k^\times$ then we have

$$\mu^6(x^3 + Ax + B) = \mu^6 x^3 + A'\mu^2 x + B',$$

and it follows that $A' = \mu^4 A$ and $B' = \mu^6 B$ as claimed.

Conversely, if $A' = \mu^4 A$ and $B' = \mu^6 B$ for some $\mu \in k^*$, then the map $\phi\colon E \to E'$ defined by $\phi(x, y) = (\mu^2 x, \mu^3 y)$ is an isomorphism, since it is an isogeny of degree 1. □

We are now ready to prove the theorem stated at the beginning of this section.

Theorem 13.14. *Let E and E' be elliptic curves over k. Then E and E' are isomorphic over $\bar{k}$ if and only if $j(E) = j(E')$. If $j(E) = j(E')$ and the characteristic of k is not 2 or 3 then there is a field extension K/k of degree at most 6, 4, or 2, depending on whether $j(E) = 0$, $j(E) = 1728$, or $j(E) \neq 0, 1728$, such that E and E' are isomorphic over K.*

Remark 13.15. The first statement is true in characteristic 2 and 3 (see [2, III.1.4.b]), but the second statement is not; one may need to take K/k of degree up to 12 when k has characteristic 2 or 3.

Proof. We assume $\mathrm{char}(k) \neq 2, 3$. Suppose $E\colon y^2 = x^3 + Ax + B$ and $E'\colon y^2 = x^3 + A'x + B'$ are isomorphic over $\bar{k}$. For some $\mu \in \bar{k}^*$ we have $A' = \mu^4 A$ and $B' = \mu^6 B$, by Theorem 13.13. We then have

$$j(A', B') = 1728\frac{4(\mu^4 A)^3}{4(\mu^4 A)^3 + 27(\mu^6 B)^2} = 1728\frac{4A^3}{4A^3 + 27B^2} = j(A, B).$$

For the converse, suppose that $j(A, B) = j(A', B') = j_0$. If $j_0 = 0$ then $A = A' = 0$ and we may choose $\mu \in K^\times$, where K/k is an extension of degree at most 6, so that $B' = \mu^6 B$

(and $A' = \mu^4 A = 0$). Similarly, if $j_0 = 1728$ then $B = 0$ and we may choose $\mu \in K^\times$, where K/k is an extension of degree at most 4, so that $A' = \mu^4 A$ (and $B' = \mu^6 B = 0$). We may then apply Theorem 13.13 to show that E and E' are isomorphic over K (by extending the field of definition of E and E' from k to K).

We now assume $j_0 \neq 0, 1728$. Let $A'' = 3j_0(1728-j_0)$ and $B'' = 2j_0(1728-j_0)^2$, as in the proof of Theorem 13.12, so that $j(A'', B'') = j_0$. Plugging in $j_0 = 1728 \cdot 4A^3/(4A^3 + 27B^2)$, we have

$$\begin{aligned} A'' &= 3 \cdot 1728 \frac{4A^3}{4A^3+27B^2}\left(1728 - 1728\frac{4A^3}{4A^3+27B^2}\right) \\ &= 3 \cdot 1728^2 \frac{4A^3 \cdot 27B^2}{(4A^3+27B^2)^2} = \left(\frac{2^7 3^5 AB}{4A^3+27B^2}\right)^2 A, \\ B'' &= 2 \cdot 1728 \frac{4A^3}{4A^3+27B^2}\left(1728 - 1728\frac{4A^3}{4A^3+27B^2}\right)^2 \\ &= 2 \cdot 1728^3 \frac{4A^3 \cdot 27^2 B^4}{(4A^3+27B^2)^3} = \left(\frac{2^7 3^5 AB}{4A^3+27B^2}\right)^3 B. \end{aligned}$$

Plugging in $j_0 = 1728 \cdot 4A'^3/(4A'^3 + 27B'^2)$ yields analogous expressions for A'' and B'' in terms of A' and B'. If we let

$$u = \left(\frac{2^7 3^5 AB}{4A^3+27B^2}\right)\left(\frac{4A'^3+27B'^2}{2^7 3^5 A'B'}\right),$$

then $A' = u^2 A$ and $B' = u^3 B$. We now choose $\mu \in K^\times$, where K/k is an extension of degree at most 2, so that we have $\mu^2 = u$. Then $A' = \mu^4 A$ and $B' = \mu^6 B$ and Theorem 13.13 implies that E and E' are isomorphic over K. □

Note that while $j(E) = j(A, B)$ always lies in the minimal field k containing A and B, the converse is not necessarily true. It could be that $j(A, B)$ lies in a proper subfield of k (squares in A can cancel cubes in B, for example). In this case we can construct an elliptic curve E' that is defined over the minimal subfield of k that contains $j(E)$ such that E' is isomorphic to E over $\bar{k}$ (but not necessarily over k).

13.3 Supersingular elliptic curves

Theorem 13.16. *Let E be a supersingular elliptic curve over a field k of characteristic $p > 0$. Then $j(E)$ lies in $\mathbb{F}_{p^2}$ (and possibly in $\mathbb{F}_p$).*

Proof. Since E is supersingular, $\hat{\pi}$ is purely inseparable, so $\hat{\pi} = \hat{\pi}_{\text{sep}}\pi$ with $\deg \hat{\pi}_{\text{sep}} = 1$. We thus have $[p] = \hat{\pi}\pi = \hat{\pi}_{\text{sep}}\pi^2$, so $\hat{\pi}_{\text{sep}}$ is an isomorphism $E^{(p^2)} \to E$. By Theorem 13.13,

$$j(E) = j(E^{(p^2)}) = j(A^{p^2}, B^{p^2}) = j(A, B)^{p^2} = j(E)^{p^2}.$$

Thus $j(E)$ is fixed by the p^2-power Frobenius automorphism $\sigma\colon x \mapsto x^{p^2}$ of k. It follows that $j(E)$ lies in the subfield of k fixed by σ, which is either $\mathbb{F}_{p^2}$ or $\mathbb{F}_p$, depending on whether k contains a quadratic extension of its prime field or not; in either case, $j(E)$ lies in $\mathbb{F}_{p^2}$. □

Remark 13.17. Note that this theorem applies to any field k of characteristic p, not just finite fields. Thus in any field k of positive characteristic, the number of $\bar{k}$-isomorphism classes of supersingular elliptic curves is finite (it certainly cannot exceed $\#\mathbb{F}_{p^2} = p^2$). In fact, there are at most $\lfloor \frac{p}{12} \rfloor + 11$; see [2, Thm. V.4.1].

Theorem 13.18. *Let E be a supersingular elliptic curve over a field k of characteristic p. Then $\mathrm{End}^0(E_{\bar{k}})$ is a quaternion algebra.*

Proof. Without loss of generality we can assume $k = \bar{k}$, so that $\mathrm{End}(E_{\bar{k}}) = \mathrm{End}(E)$. Let us suppose for the sake of contradiction that $\mathrm{End}^0(E)$ is not a quaternion algebra. Then $\mathrm{End}(E)$ is isomorphic to $\mathbb{Z}$ or an order in an imaginary quadratic field $\mathbb{Q}(\sqrt{D})$, where we may assume $D < 0$ is squarefree. We claim there are infinitely many odd primes ℓ that are **not** the degree of any $\phi \in \mathrm{End}(E)$. This is obvious if $\mathrm{End}(E) \simeq \mathbb{Z}$, since $\deg[n] = n^2$ is a square, and if $\mathrm{End}(E)$ is an order in $\mathbb{Q}(\sqrt{D})$ and ℓ is the degree of ϕ then the polynomial $x^2 - (\mathrm{tr}\,\phi)x + \ell$ has a root in $\mathrm{End}^0(E) \simeq \mathbb{Q}(\sqrt{D})$, which implies

$$\mathrm{tr}(\phi)^2 - 4\ell = v^2 D$$

for some integer v, and D must be a square modulo ℓ. There are infinitely many primes $\ell \neq p$ for which this is not true (these are the primes that do not split in the quadratic field $\mathbb{Q}(\sqrt{D})$). So let $\ell_1, \ell_2, \ldots$ be an infinite sequence of odd primes different from p that are not the degree of any $\phi \in \mathrm{End}(E)$.

For each ℓ_i we may construct a separable isogeny $\phi_i \colon E \to E_i$ of degree ℓ_i defined over $\bar{k}$ whose kernel is a cyclic subgroup of order ℓ_i contained in $E[\ell_i]$ using Vélu's formulas (see Theorem 5.15). The elliptic curves E_i are all supersingular, by Theorem 13.2, and Theorem 13.16 implies that only finitely many of them have distinct j-invariants. By Theorem 13.14, over $\bar{k}$ we must have an isomorphism $\iota \colon E_i \xrightarrow{\sim} E_j$ for some distinct i and j. Let us now consider the endomorphism $\phi := \hat{\phi}_j \circ \iota \circ \phi_i \in \mathrm{End}(E)$ of degree $\ell_i \ell_j$. The degree of this endomorphism is not a square, so $\mathrm{End}(E) \not\simeq \mathbb{Z}$ and we have $\mathrm{End}^0(E) \simeq \mathbb{Q}(\sqrt{D})$. As above we must have

$$\mathrm{tr}(\phi)^2 - 4\ell_i\ell_j = v^2 D,$$

for some integer v, which implies that D is a square modulo ℓ_i (and ℓ_j), a contradiction. □

When k is a finite field, the converse of Theorem 13.18 is implied by Theorem 13.6, but in fact the converse holds in general.

Theorem 13.19. *Let E be an elliptic curve over a field k of characteristic p for which $\mathrm{End}^0(E_{\bar{k}})$ is a quaternion algebra. Then E is supersingular.*

Proof. Without loss of generality, we may assume k is algebraically closed, since the property of being supersingular, defined by $E[p] = \{0\}$, is invariant under base change, as is $\mathrm{End}^0(E_{\bar{k}})$. Let $\alpha, \beta \in \mathrm{End}(E)$ be nonzero endomorphisms that satisfy $\alpha\beta = -\beta\alpha$ so that $\alpha\beta + \beta\alpha = 0$ (such α, β exist because $\mathrm{End}(E)$ is a quaternion algebra).

Now suppose E is ordinary. Then $E[p^n] = \langle P \rangle \simeq \mathbb{Z}/p^n\mathbb{Z}$ for some point $P \in E(k)$ of order p^n. We then have $\alpha(P) = aP$ and $\beta(P) = bP$ for some integers a and b. If we choose $n > v_p(\deg\alpha) + v_p(\deg\beta) + 1$ where v_p denotes the p-adic valuation (the exponent of the largest p-power divisor), then $ab + ba = 2ab$ must be nonzero modulo p, since α must send P to a point of order at least $p^{n - v_p(\deg\alpha)}$ and similarly for β (and we handled $p = 2$ by adding 1). But this contradicts $\alpha\beta + \beta\alpha = 0$, so E cannot be ordinary. □

Corollary 13.20. *Let E be an elliptic curve over a finite field $\mathbb{F}_q$ of characteristic p. Either E is supersingular, $\mathrm{tr}\,\pi_E \equiv 0 \bmod p$, and $\mathrm{End}^0(E_{\overline{\mathbb{F}}_q})$ is a quaternion algebra, or E is ordinary, $\mathrm{tr}\,\pi_E \not\equiv 0 \bmod p$, and $\mathrm{End}^0(E_{\overline{\mathbb{F}}_q}) = \mathrm{End}^0(E_{\mathbb{F}_q})$ is an imaginary quadratic field.*

Warning 13.21. If E is a supersingular elliptic curve over $\mathbb{F}_p$ (or any odd degree extension), then $\mathrm{End}^0(E)$ is an imaginary quadratic field (by Corollary 13.7), even though $\mathrm{End}^0(E_{\overline{\mathbb{F}}_p})$ is a quaternion algebra.

References

[1] Gaetan Bisson and Andrew V. Sutherland, *Computing the endomorphism ring of an ordinary elliptic curve over a finite field*, Journal of Number Theory **131** (2011), 815–831.

[2] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer 2009.

[3] Joseph H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Springer, 1994.

14 Elliptic curves over $\mathbb{C}$ (part I)

We now consider elliptic curves over the complex numbers. Our first goal is to prove the Uniformization Theorem, which establishes an explicit correspondence between elliptic curves over $\mathbb{C}$ and tori $\mathbb{C}/L$ defined by lattices L in $\mathbb{C}$:

1. Every lattice L can be used to define an elliptic curve $E/\mathbb{C}$.
2. Every elliptic curve $E/\mathbb{C}$ arises from a lattice L.
3. If $E/\mathbb{C}$ is the elliptic curve corresponding to the lattice L, then there is an isomorphism

$$\mathbb{C}/L \xrightarrow{\Phi} E(\mathbb{C})$$

that is both analytic (as a mapping of complex manifolds) and algebraic: addition of points in $E(\mathbb{C})$ corresponds to addition in $\mathbb{C}$ modulo the lattice L.

To make the correspondence explicit, we need to specify the map Φ. This map is parameterized by *elliptic functions*, specifically the Weierstrass $\wp$-function and its derivative. We will begin by studying general properties of elliptic functions in §14.1 and Eisenstein series in §14.3, then specialize to the Weierstrass $\wp$-function in §14.4 and construct the map Φ in §14.5. Our presentation generally follows that in [2, Ch. 3, §10], but we will fill in some more details for the benefit of those who have not taken a course in complex analysis.

Once we have fleshed out this correspondence, we will have a powerful method to construct elliptic curves with desired properties. The arithmetic properties of lattices over $\mathbb{C}$ are usually easier to understand than those of the corresponding elliptic curve. In particular, by choosing an appropriate lattice, we can construct an elliptic curve with a given endomorphism ring. In the case of elliptic curves over $\mathbb{C}$, the endomorphism ring must either be $\mathbb{Z}$ or an order $\mathcal{O}$ in an imaginary quadratic field (a fact we will prove). The order $\mathcal{O}$ may be viewed as a lattice, and we will see that the elliptic curve corresponding to the torus $\mathbb{C}/\mathcal{O}$ has endomorphism ring $\mathcal{O}$.

This has important implications for elliptic curves over finite fields. If we choose a suitable prime p, we can reduce an elliptic curve $E/\mathbb{C}$ with complex multiplication to an elliptic curve $E_p/\mathbb{F}_p$ with the same endomorphism ring $\mathcal{O}$. The endomorphism ring determines, in particular, the trace of the Frobenius endomorphism π_{E_p} (up to a sign), which in turn determines $\#E_p(\mathbb{F}_p) = p + 1 - \mathrm{tr}(\pi_{E_p})$. This allows us to construct elliptic curves over finite fields that have a prescribed number of rational points, using what is known as the *CM method*. As we will see, this has many practical applications, including cryptography and a faster version of elliptic curve primality proving.

14.1 Elliptic functions

We begin with the definition of a lattice in the complex plane.

Definition 14.1. A *lattice* $L = [\omega_1, \omega_2]$ in $\mathbb{C}$ is an additive subgroup $\omega_1\mathbb{Z} + \omega_2\mathbb{Z}$ of $\mathbb{C}$ generated by complex numbers ω_1 and ω_2 that are linearly independent over $\mathbb{R}$.

Example 14.2. Let τ be the root of a monic quadratic equation $x^2 + bx + c$ with integer coefficients and negative discriminant. Then the lattice $[1, \tau]$ is the additive group of an imaginary quadratic order $\mathcal{O} = \mathbb{Z}[\tau]$. Conversely, if $\mathcal{O}$ is an imaginary quadratic order $\mathbb{Z}[\tau]$, then the additive group of $\mathcal{O}$ is the lattice $[1, \tau]$.

If we take the quotient of the complex plane $\mathbb{C}$ modulo a lattice L, we get a torus $\mathbb{C}/L$. Note that this quotient makes sense not just as a quotient of abelian groups, but also as a quotient of topological spaces (where $\mathbb{C}$ has its usual Euclidean topology and L has the discrete topology); the torus $\mathbb{C}/L$ is a compact topological group.

Definition 14.3. A *fundamental parallelogram* for $L = [\omega_1, \omega_2]$ is any set of the form

$$\mathcal{F}_\alpha = \{\alpha + t_1\omega_1 + t_2\omega_2 : 0 \le t_1, t_2 < 1\}.$$

for some $\alpha \in \mathbb{C}$. We can identify the points in any $\mathcal{F}_\alpha$ with the points of $\mathbb{C}/L$.

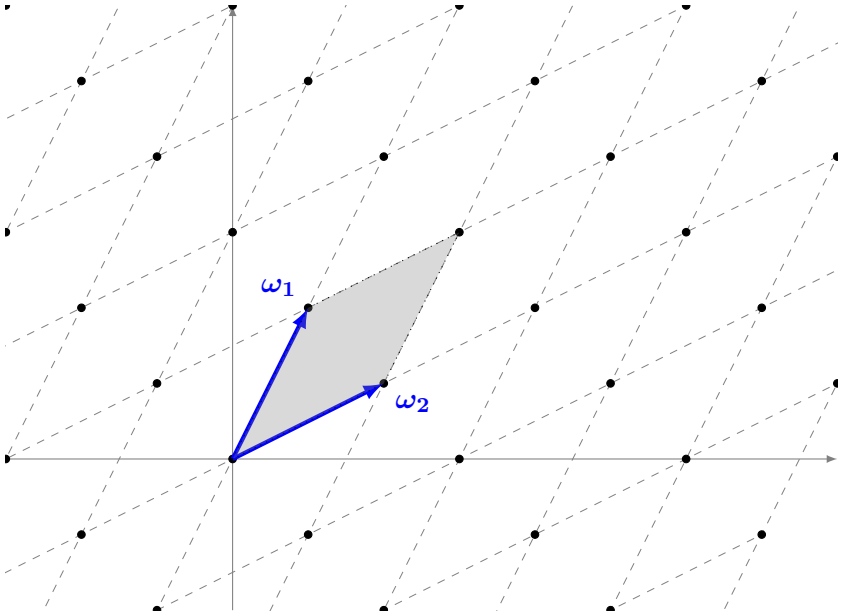


Figure 1: A lattice $[\omega_1, \omega_2]$ with a fundamental parallelogram shaded.

In order to define the correspondence between complex tori and elliptic curves over $\mathbb{C}$, we need to define the notion of an *elliptic function* on $\mathbb{C}$. As complex analysis is not an official prerequisite for this course, we will take a moment to define the terminology we need and recall some elementary results that can be found in standard textbooks such as [1, 4, 6].

Definition 14.4. A function $f\colon \Omega \to \mathbb{C}$ defined on an open neighborhood Ω of a point $z_0 \in \mathbb{C}$ is said to be *holomorphic* at z_0 if the derivative

$$f'(z_0) := \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists.[1] We say that f is holomorphic on an open set Ω if it is holomorphic at every $z_0 \in \Omega$. Functions that are holomorphic on all of $\mathbb{C}$ are simply said to be holomorphic or *entire*.

Examples of holomorphic functions include polynomials and convergent power series. Functions that admit a power series expansion with a positive radius of convergence about a point z_0 are said to be *analytic* at z_0. Remarkably, any function that is holomorphic at z_0 is also analytic at z_0 (see [1, Thm. 5.3] or [6, Thm. 2.4.4]), so the terms analytic and holomorphic may be used interchangeably (modern usage favors holomorphic).

[1]The limit must take the same value no matter how the complex number z approaches z_0; this makes differentiability a much stronger condition on a complex function than it is on a real function.

Definition 14.5. Let k be a positive integer. A complex function $f(z)$ has a *zero of order k* at z_0 if an equation of the form

$$f(z) = (z - z_0)^k g(z)$$

holds in some open neighborhood of z_0 in which $g(z)$ is holomorphic and $g(z_0) \neq 0$. We say that $f(z)$ has a *pole of order k* at z_0 if the function $1/f(z)$ has a zero of order k at z_0. A pole or zero of order 1 is called a *simple pole* or a *simple zero*.

Definition 14.6. A complex function f is *meromorphic* on an open set Ω if it is holomorphic at every point on Ω except for a discrete set of poles.[2]

Definition 14.7. For any nonzero complex function $f(z)$ that is meromorphic on an open neighborhood of a point $z_0 \in \mathbb{C}$ we define

$$\operatorname{ord}_{z_0}(f) := \begin{cases} n & \text{if } f \text{ has a zero of order } n \text{ at } z_0, \\ -n & \text{if } f \text{ has a pole of order } n \text{ at } z_0, \\ 0 & \text{otherwise.} \end{cases}$$

For any open set $\Omega \subseteq \mathbb{C}$, the set of complex functions that are meromorphic on Ω forms a field $\mathbb{C}(\Omega)$ that we view as an extension of $\mathbb{C}$ (the constant functions). For each fixed $z_0 \in \Omega$, we then have a *discrete valuation* $\operatorname{ord}_{z_0} \colon \mathbb{C}(\Omega)^\times \to \mathbb{Z}$, which has the following properties:

1. $\operatorname{ord}_{z_0}(fg) = \operatorname{ord}_{z_0}(f) + \operatorname{ord}_{z_0}(g)$ for all $f, g \in \mathbb{C}(\Omega)^\times$;
2. $\operatorname{ord}_{z_0}(f + g)) \geq \min(\operatorname{ord}_{z_0}(f), \operatorname{ord}_{z_0}(g))$ for all $f, g \in \mathbb{C}(\Omega)^\times$.

We note that the second inequality is in fact an equality whenever $\operatorname{ord}_{z_0}(f) \neq \operatorname{ord}_{z_0}(g)$. It is customary to extend ord_{z_0} to all of $\mathbb{C}(\Omega)$ by defining $\operatorname{ord}_{z_0}(0) := \infty$, with addition and comparisons in $\mathbb{Z} \cup \{\infty\}$ defined in the obvious way.

Definition 14.8. An *elliptic function* for a lattice L is a complex function $f(z)$ such that

1. f is meromorphic on $\mathbb{C}$.
2. f is periodic with respect to L; this means that $f(z + \omega) = f(z)$ for all $\omega \in L$.[3]

The fact that an elliptic function is periodic with respect to L means that it can also be viewed as a function on $\mathbb{C}/L$. Note that if f is an elliptic function for L then it is also an elliptic function for every sub-lattice of L. Sums, differences, products, and quotients of elliptic functions for a lattice L are also elliptic functions for L; thus the set of elliptic functions for a fixed lattice L form a field that we denote $\mathbb{C}(L)$; note that constant functions are elliptic functions for every lattice L.

Definition 14.9. The *order* of an elliptic function is the number of poles it has in any fundamental parallelogram, where each pole is counted with multiplicity equal to its order (this is a finite number because the poles in a fundamental parallelogram are a discrete subset of its closure, which is compact).

As a general rule, whenever we count the poles or zeros of a meromorphic function, we always count them with multiplicity.

Remark 14.10. The elliptic functions of order zero are precisely the constant functions. This follows from Liouville's theorem (see Theorem 14.30 below), since a holomorphic elliptic function is necessarily bounded (as a continuous function it must achieve a maximum value on any compact set, including the closure of a fundamental parallelogram), hence constant.

[2] This means that each pole lies in an open subset of Ω that contains no other poles.

[3] If $L = [\omega_1, \omega_2]$ the function f is also said to be *doubly periodic*, with *periods* ω_1 and ω_2.

14.2 Contour integrals and the residue formula

In order to count poles and zeros of meromorphic functions (and elliptic functions in particular), we need a few standard tools from complex analysis that we briefly recall here. Those who are familiar with this material can skip ahead to Theorem 14.18, which uses Cauchy's argument principle to deduce that an elliptic function has the same number of zeros as poles in any fundamental parallelogram.

Definition 14.11. A *smooth curve* in $\mathbb{C}$ is a continuously differentiable function

$$\gamma\colon [a,b] \to \mathbb{C},$$

where $[a,b]$ is a closed interval in $\mathbb{R}$. A *piecewise smooth curve* $\gamma\colon [a,b] \to \mathbb{C}$ is defined by a finite sequence of n smooth curves $\gamma_i\colon [a_i,b_i] \to \mathbb{C}$ with $a_0 = a$, $a_{i+1} = b_i$, and $b_n = b$. We will simply use the term *curve* to refer to a piecewise smooth curve.[4] A curve is *simple* if its restriction to the open interval (a,b) is injective, and it is *closed* if $\gamma(a) = \gamma(b)$.

For simple closed curves γ the Jordan curve theorem (see [1, §4.2 Ex. 3] or [6, Appendix B, Thm. 2.1]) gives a well-defined notion of interior and exterior, as well as a notion of positive and negative orientation. Loosely speaking, we that a simple closed curve is *positively oriented* if the interior is on the left as we travel along the curve (if γ is a circle, this means counter-clockwise). The notion of orientation can be made completely precise using *winding numbers*, but this is overkill for our purposes here; the simple closed curves we will use (circles and parallelograms) all have obvious interiors and orientation.

Definition 14.12. For a smooth curve $\gamma\colon [a,b] \to \mathbb{C}$ and a complex function $f(z)$ defined on an open set containing γ the *contour integral* of f along γ is defined by

$$\int_\gamma f(z)dz := \int_a^b f(\gamma(t))\gamma'(t)dt.$$

This definition extends to piecewise smooth curves in the obvious way (sum the contour integrals on each smooth piece).

Theorem 14.13. *Let Ω be an open set containing a curve $\gamma\colon [a,b] \to \mathbb{C}$, and let $F(z)$ be a holomorphic function on Ω and let $f(z) = F'(z)$. Then*

$$\int_\gamma f(z)dz = F(\gamma(b)) - F(\gamma(a)).$$

Proof. If γ is smooth then

$$\int_\gamma f(z)dz = \int_a^b F'(\gamma(t))\gamma'(t)dt = \int_a^b \left(\frac{d}{dt}F(\gamma(t))\right)dt = F(\gamma(b)) - F(\gamma(a)).$$

The piecewise smooth case follows by taking summing over smooth pieces. □

It is a non-trivial fact that if $f(z)$ is holomorphic on a simply connected open set Ω then there exists a holomorphic function[5] $F(z)$ for which $f(z) = F'(z)$ (this is obvious locally,

[4] More generally one can define *rectifiable curves* that are defined by continuous (but not necessarily differentiable) functions and have finite length, but we will not need these.

[5] The function $F(z)$ is called a *primitive* of $f(z)$.

since in a neighborhood of each $z_0 \in \Omega$ there is a power series expansion of $f(z)$ about z_0 that we can integrate term by term, but we want a single $F(z)$ that works for all $z_0 \in \Omega$); see [1, §4.1 Thm. 4] or [6, §2 Thm. 2.1] for a proof in the case that Ω is a disc. An important consequence of this fact is Cauchy's theorem.

Theorem 14.14 (Cauchy's theorem)**.** *Let f be a function that is holomorphic on an open set containing a closed curve γ and its interior. Then*

$$\int_\gamma f(z)dz = 0.$$

Proof. See [6, Appendix B Thm. 2.9]. □

A corollary of this theorem is that the countour integral of a holomorphic function depends only on the end points $(\gamma(a), \gamma(b))$ of the curve γ, not the path taken from $\gamma(a)$ to $\gamma(b)$.

We now want to consider countour integrals of functions that are meromorphic but not necessarily holomorphic. Note that a function $f(z)$ that is meromorphic on an open set Ω has a Laurent series expansion

$$f(z) = \sum_{n \geq n_0} a_n (z - z_0)^n$$

about any point $z_0 \in \Omega$. Here $n_0 = \mathrm{ord}_{z_0}(f)$ can be any integer (positive or negative), and we define $a_n = 0$ for all $n < n_0$.

Definition 14.15. The *residue* at z_0 of a function $f(z) = \sum_{n \geq n_0} a_n (z - z_0)^n$ that is meromorphic on an open neighborhood of z_0 is

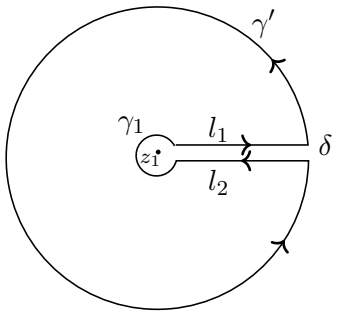
$$\mathrm{res}_{z_0}(f) := a_{-1}.$$

If f is holomorphic at z_0 then $\mathrm{res}_{z_0}(f) = 0$. Even if f has a pole at z_0 it is still possible to have $\mathrm{res}_{z_0}(f) = 0$ when the order of the pole is greater than 1, but if f has a simple pole at z_0 then $\mathrm{res}_{z_0}(f)$ must be nonzero. This definition may look strange at first glance, but it is motivated by the following theorem.

Theorem 14.16 (Residue formula)**.** *Let γ be a simple closed curve with positive orientation and let $f(z)$ be a function that is meromorphic on an open set containing γ and its interior with no poles on γ. Let $z_1, \ldots, z_N$ be the poles of $f(z)$ that lie in the interior of γ. Then*

$$\int_\gamma f(z)dz = 2\pi i \sum_{k=1}^{N} \mathrm{res}_{z_k}(f).$$

Proof. Let us first suppose that γ is a circle and that $f(z)$ has a single pole at z_1 inside γ. We now consider a *keyhole contour* $\tilde{\gamma}$ that approximates γ but whose interior does not contain z_1, as shown below. The function $f(z)$ is holomorphic on an open set that contains $\tilde{\gamma}$ and its interior, but not z_1; thus $\int_{\tilde{\gamma}} f(z)dz = 0$, by Cauchy's theorem.



As the distance δ between the horizontal segments l_1 and l_2 goes to zero, the sum $\int_{l_1} f(z)dz) + \int_{l_2} f(z)dz$ approaches zero while $\int_{\gamma'} f(z)dz$ approaches $\int_\gamma f(z)dz$. In the limit we have

$$\int_{\tilde{\gamma}} f(z)dz = 0 = \int_\gamma f(z)dz - \int_{c_1} f(z)dz,$$

where c_1 is a positively oriented circle with the same radius as the arc γ_1 (which is oriented in the opposite direction; this explains the minus sign in the equation above). Thus

$$\int_\gamma f(z)dz = \int_{c_1} f(z)dz.$$

If $f(z) = \sum_{n\geq n_0} a_n(z-z_1)^n$ is the Laurent series for $f(z)$ about z_1, then

$$\int_{c_1} f(z)dz = \int_{c_1} \left(\sum_{n=n_0}^{-1} a_n(z-z_0)^n + \sum_{n\geq 0} a_n(z-z_0)^n \right) dz.$$

The infinite sum on the right is holomorphic in an open neighborhood of z_0 that we can assume contains c_1, since we can make the radius of c_1 as small as we wish, thus the integral of this sum is zero. It thus suffices to compute the integrals $\int_{c_1}(z-z_0)^n dz$ for negative n. After replacing $z-z_0$ with u and dz by du we can assume c_1 is a circle about 0 parameterized by re^{it}, where r is the radius of c_1. For $n<0$ we then have

$$\int_{c_1} u^n du = \int_0^{2\pi} (re^{it})^n (ire^{it})dt = \int_0^{2\pi} ir^{n+1}e^{(n+1)it}dt = \begin{cases} 0 & \text{if } n<-1, \\ 2\pi i & \text{if } n=-1. \end{cases}$$

Thus

$$\int_\gamma f(z)dz = \int_{c_1} f(z)dz = 2\pi i a_{-1} = 2\pi i \operatorname{res}_{z_1}(f)$$

as desired. The case where $f(z)$ has N poles inside γ is similar; we now approximate γ with a contour $\tilde{\gamma}$ that has N keyholes, one about each z_k, each of which has an inner arc with negative (clockwise) orientation. We then obtain

$$\int_\gamma f(z)dz = 2\pi i \sum_{k=1}^N \operatorname{res}_{z_k}(f).$$

The same argument applies when γ is not a circle, it just requires approximating γ with a more complicated contour $\tilde{\gamma}$. □

We can now use the residue formula to derive a generalization of Cauchy's *argument principle*, which is our main tool for counting the zeros and poles of a meromorphic function.

Theorem 14.17. *Let γ be a simple closed curve with positive orientation, let $f(z)$ be a function that is meromorphic on an open set Ω containing γ and its interior Γ, with no zeros or poles on γ, and let $g(z)$ be a nonzero function that is holomorphic on Ω.*

$$\frac{1}{2\pi i}\int_\gamma g(z)\frac{f'(z)}{f(z)}\,dz = \sum_{w\in\Gamma} g(w)\mathrm{ord}_w(f).$$

When $g(z) = 1$, the RHS is the difference between the number of zeros and poles that $f(z)$ has in Γ (counted with multiplicity), which is the usual argument principle.

Proof. For any $z_0 \in \Gamma$ that is a zero or pole of $f(z)$, we consider the Laurent series expansions

$$f(z) = \sum_{n\ge n_0} a_n(z-z_0)^n, \qquad g(z) = \sum_{n\ge 0} b_n(z-z_0)^n$$

where $n_0 = \mathrm{ord}_{z_0}(f)$ is chosen so that $a_{n_0} \neq 0$ and we note that $g(z_0) = b_0$. Then

$$f'(z) = \sum_{n\ge n_0} n a_n (z-z_0)^{n-1}$$

and we have

$$\frac{f'(z)}{f(z)} = n_0(z-z_0)^{-1} + h_1(z), \qquad g(z)\frac{f'(z)}{f(z)} = b_0 n_0 (z-z_0)^{-1} + h_2(z),$$

where $h_1(z)$ and $h_2(z)$ denote functions that are holomorphic on an open neighborhood of z_0. Thus $g(z)f'(z)/f(z)$ has a simple pole with residue $b_0 n_0 = g(z_0)\mathrm{ord}_{z_0}(f)$ at each zero or pole z_0 of $f(z)$, and no other poles. The theorem follows from the residue formula. □

Applying Theorem 14.17 with $g(z) = 1$ to an elliptic function $f(z)$ yields the following.

Theorem 14.18. *Let $f(z)$ be a nonzero elliptic function for a lattice L. When counted with multiplicity, the number of zeros of $f(z)$ in any fundamental parallelogram $\mathcal{F}_\alpha$ for L is equal to the number of poles of $f(z)$ in $\mathcal{F}_\alpha$.*

Proof. We first note that by the periodicity of $f(z)$, it suffices to prove this for any particular fundamental parallelogram $\mathcal{F}_\alpha$. The zeros and poles of $f(z)$ are discrete (note that $1/f(z)$ is also a meromorphic function), so we can pick an α for which the boundary $\partial\mathcal{F}_\alpha$ of $\mathcal{F}_\alpha$ does not contain any zeros or poles of $f(z)$. We now consider the contour integral

$$\int_{\partial\mathcal{F}_\alpha} \frac{f'(z)}{f(z)}dz,$$

where the simple closed curve $\partial\mathcal{F}_\alpha$ is positively oriented. The fact that $f(z)$ is periodic with respect to L implies that $f'(z)$ is also periodic with respect to L, as is $f'(z)/f(z)$, and it follows that the sum of the integral of $f'(z)/f(z)dz$ along opposite sides of the parallelogram ∂F_α is zero, since $f'(z)/f(z)$ takes on the same values on both sides (because it is periodic) but the oriented curve ∂F_α traverses them in opposite directions. We thus have

$$\frac{1}{2\pi i}\int_{\partial\mathcal{F}_\alpha} \frac{f'(z)}{f(z)}dz = 0,$$

and the theorem then follows from Theorem 14.17. □

14.3 Eisenstein series

Before giving some non-trivial examples of elliptic functions, we first define the Eisenstein series of a lattice.

Definition 14.19. Let L be a lattice in $\mathbb{C}$ and let $k > 2$ be an integer. The *weight-k Eisenstein series* for L is the sum

$$G_k(L) = \sum_{\omega \in L^*} \frac{1}{\omega^k},$$

where $L^* = L - \{0\}$.

Remark 14.20. $G_k(L)$ is a function of the lattice L, so for any fixed lattice, it is a constant. If we consider lattices $L = [1, \tau]$ parameterized by a complex number τ in the *upper half plane* $\mathcal{H} := \{z \in \mathbb{C} : \operatorname{im} z > 0\}$, we can view $G_k(L)$ as a function of τ:

$$G_k(\tau) := G_k([1,\tau]) = \sum_{\substack{m,n \in \mathbb{Z} \\ (m,n) \neq (0,0)}} \frac{1}{(m+n\tau)^k}.$$

Because it comes from a function defined over a lattice, the function $G_k(\tau)$ has some very nice properties. In particular, we have

$$G_k(\tau + 1) = G_k(\tau) \qquad \text{and} \qquad G_k(-1/\tau) = \tau^k G_k(\tau)$$

for all $\tau \in \mathcal{H}$. Eisenstein series are the simplest example of *modular forms*, which we will see later in the course.[6]

Remark 14.21. If k is odd then $G_k(L) = 0$ for any lattice L, since the terms $\frac{1}{\omega^k}$ and $\frac{1}{(-\omega)^k}$ in the sum cancel (note that L is an additive group, so $\omega \in L \implies -\omega \in L$, and in the sum over L^*, each ω is distinct from $-\omega$). Thus the only interesting Eisenstein series are those of even weight.

Lemma 14.22. *For any lattice L, the sum $\sum_{\omega \in L^*} \frac{1}{\omega^k}$ converges absolutely for all $k > 2$.*

Proof. Let δ be the minimum distance between points in L. Consider an annulus A of inner radius r and width $\frac{\delta}{2}$, as depicted in Figure 2.

Any two distinct lattice points in A must be separated by an arc of length at least $\delta/2$ when measured along the inner rim of A. It follows that A contains at most $4\pi r/\delta$ lattice points. If we put $r = n$ and $\delta = 2$ then the number of lattice points in the annulus $\{\omega : n \le |\omega| < n+1\}$ is bounded by $4\pi n/2 = 2\pi n$. We then have

$$\sum_{\omega \in L,\, |\omega| \ge 1} \frac{1}{|\omega|^k} \;\le\; \sum_{n=1}^{\infty} \frac{2\pi n}{n^k} \;=\; 2\pi \sum_{n=1}^{\infty} \frac{1}{n^{k-1}} \;< \infty,$$

since $k > 2$. The finite sum $\sum_{\omega \in L,\, 0<|\omega|<1} \frac{1}{|\omega|^k}$ is clearly bounded, thus

$$\sum_{\omega \in L^*} \frac{1}{|\omega|^k} \;=\; \sum_{\substack{\omega \in L \\ 0<|\omega|<1}} \frac{1}{|\omega|^k} \;+\; \sum_{\substack{\omega \in L \\ |\omega| \ge 1}} \frac{1}{|\omega|^k} \;<\; \infty,$$

so the sum converges absolutely as claimed. □

[6] Many authors use E_k to denote Eisenstein series, rather than G_k, but since we are already using the (often subscripted) symbol E for elliptic curves, we will stick with G_k.

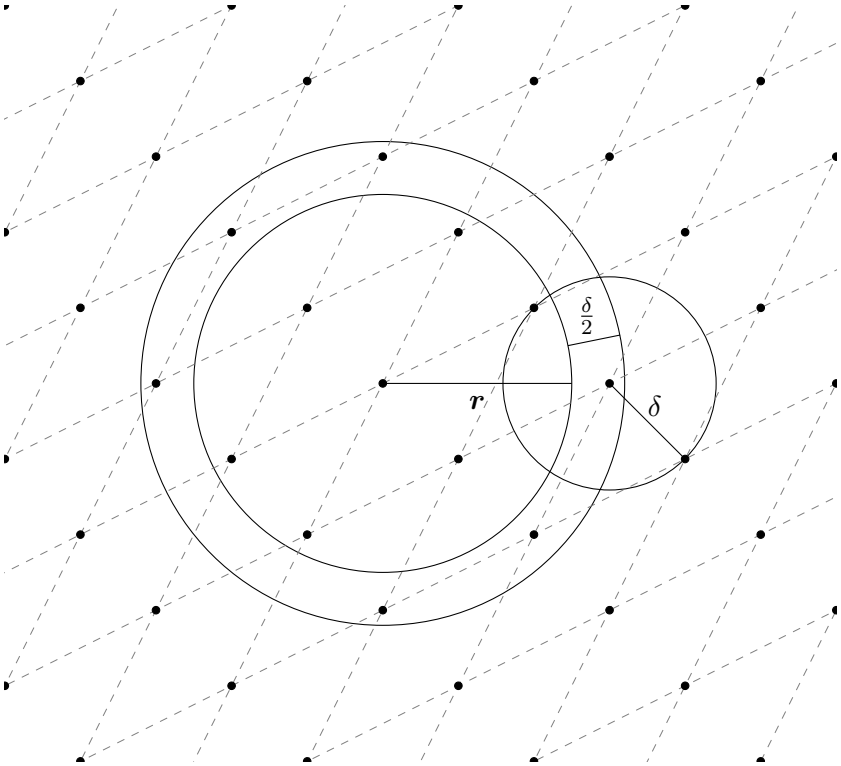


Figure 2: Annulus of radius r and width $\delta/2$.

14.4 The Weierstrass $\wp$-function

We now give our first example of a non-constant elliptic function. It may be regarded as *the* elliptic function in the sense that it can be used to construct every other non-constant elliptic function, a fact we will prove in the next lecture (or see [5, Thm. VI.3.2]).

Definition 14.23. The *Weierstrass $\wp$-function* of a lattice L in $\mathbb{C}$ is defined by

$$\wp(z) := \wp(z; L) := \frac{1}{z^2} + \sum_{\omega \in L^*} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right).$$

When the lattice L is fixed or clear from context we typically just write $\wp(z)$, but we should keep in mind that this function depends on L. It is clear from the definition that $\wp(z)$ has a pole of order 2 at each point in $z \in L$ (including $z = 0$); we will show that it has no other poles and is in fact holomorphic at every point not in L. To do so we rely on the following theorem from complex analysis.

Theorem 14.24. *Suppose $\{f_n\}$ is a sequence of functions holomorphic on an open set Ω, and that $\{f_n\}$ converges to a function f uniformly on every compact subset of Ω. Then f is holomorphic on Ω, and $\{f_n'\}$ converges uniformly to f' on every compact subset of Ω.*

Proof. See [1, §5 Thm. 1] or [6, §2 Thm. 5.2-3]. □

Theorem 14.25. *The function $\wp(z; L)$ is holomorphic at every $z_0 \notin L$.*

Proof. For each positive integer n, we define the function

$$f_n(z) = \frac{1}{z^2} + \sum_{\substack{\omega \in L \\ 0<|\omega|<n}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right).$$

Each $f_n(z)$ is clearly holomorphic at any $z \notin L$, since we can differentiate the finite sum term by term. We will show that the sequence of functions $\{f_n\}$ converges uniformly to $\wp$ on all compact sets S disjoint from L. Theorem 14.24 will then imply that $\wp(z)$ is holomorphic on the open set $\mathbb{C} - L$.

So let S be a compact subset of $\mathbb{C}$ disjoint from L. Then S is bounded and we may fix $r \in \mathbb{R}_{>0}$ such that $|z| \le r$ for all $z \in S$. For all but finitely many $\omega \in L$, we have $|\omega| \ge 2r$. By the triangle inequality, $|\omega - z| + |z| \ge |\omega|$, so $|\omega| \ge 2r$ implies the following inequalities:

$$|\omega - z| \ge |\omega| - |z| \ge \frac{1}{2}|\omega|,$$
$$|2\omega - z| \le |2\omega| + |-z| \le \frac{5}{2}|\omega|.$$

Thus the bound

$$\left| \frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right| = \left| \frac{z(2\omega - z)}{\omega^2 (z-\omega)^2} \right| \le \frac{r\frac{5}{2}|\omega|}{|\omega|^2 (\frac{1}{2}|\omega|)^2} = \frac{10r}{|\omega|^3}$$

holds for all $z \in S$. The series $\sum_{\omega \in L^*} \frac{1}{|\omega|^3}$ converges, by Lemma 14.22, so

$$\sum_{\omega \in L^*} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

converges absolutely for all $z \in S$, and the rate of convergence can be bounded in terms of r and L, independent of z. It follows that $\{f_n\}$ converges uniformly to $\wp$ on S, since for every $\epsilon > 0$ there is an N such that for all $n \ge N$ we have $|\wp(z) - f_n(z)| < \epsilon$ for all $z \in S$. □

With Theorem 14.25 in hand, we can now summarize the key properties of $\wp(z)$.

Theorem 14.26. *The function $\wp(z) = \wp(z; L)$ and its derivative*

$$\wp'(z) = -2 \sum_{\omega \in L} \frac{1}{(z-\omega)^3}$$

satisfy the following:

(i) *$\wp(z)$ is a meromorphic even function whose poles consist of double poles at each $z \in L$.*

(ii) *$\wp'(z)$ is a meromorphic odd function whose poles consist of triple poles at each $z \in L$.*

Proof. We first note that the sequence of functions $\{f_n\}$ defined in the proof of Theorem 14.25 consist of finite partial sums that converge uniformly to $\wp(z)$ on compact subsets of $\mathbb{C} - L$, and Theorem 14.24 implies that the sequence $\{f_n'\}$ converges uniformly to $\wp'(z)$, thus we can therefore differentiate $\wp(z)$ term by term to obtain $\wp'(z)$ (the sum for $\wp'(z)$ includes $\omega = 0$ due to the leading $1/z^2$ term in $\wp(z)$). It is clear that $\wp(z)$ has a double pole at each lattice point, and (i) then follows from Theorem 14.25 and the fact that $\wp(z) = \wp(-z)$. Part (ii) is clear from the formula for $\wp'(z)$ and the fact that the derivative of a function that is holomorphic on an open neighborhood of a point z is also holomorphic on that neighborhood (so $\wp'(z)$ is holomorphic at all $z \notin L$ since $\wp(z)$ is). □

Corollary 14.27. *The function $\wp(z) = \wp(z; L)$ is an elliptic function of order 2 for L, and its derivative $\wp'(z)$ is an elliptic function of order 3 for L.*

Proof. We've just shown that $\wp(z)$ and $\wp'(z)$ are meromorphic. Every fundamental region of L contains exactly one lattice point, so $\wp(z)$ has two poles in each fundamental region, while $\wp'(z)$ has three. It is clear from the formula for $\wp'(z)$ that $\wp'(z)$ is periodic with respect to L, we just need to show that $\wp(z)$ is periodic. Let $L = [\omega_1, \omega_2]$. It suffices to show that

$$\wp(z + \omega_i) = \wp(z), \quad \text{for } i = 1, 2.$$

Now $\wp'(z)$ is periodic, so $\wp'(z + \omega_i) = \wp'(z)$. Integrating then gives

$$\wp(z + \omega_i) - \wp(z) = c_i.$$

for some constant c_i and for all $z \notin L$. To find c_i, plug in $z = -\omega_i/2$. We have

$$\wp(\omega_i/2) - \wp(-\omega_i/2) = c_i,$$

but $\wp(z)$ is an even function, so $c_i = 0$ and $\wp(z + \omega_i) = \wp(z)$ as desired. □

The study of elliptic functions dates back to Gauss, who discovered them as solutions to elliptic integrals (they were later rediscovered by Abel and Jacobi). We will show that $\wp(z)$ satisfies a differential equation of the form $\wp'(z)^2 = f(\wp(z))$, where $f(x)$ is a cubic polynomial over $\mathbb{C}$. Notice that if one views $(\wp(z), \wp'(z))$ as a pair (x, y), this is exactly the equation of an elliptic curve! This explains our interest in $\wp(z)$.

To derive the differential equation satisfied by the Weierstrass $\wp$-function, we first need to compute its Laurent series.

Theorem 14.28. *The Laurent series expansion of $\wp(z) = \wp(z; L)$ at $z = 0$ is given by*

$$\wp(z) = \frac{1}{z^2} + \sum_{n=1}^{\infty} (2n + 1) G_{2n+2}(L) z^{2n},$$

where $G_k(L)$ denotes the Eisenstein series of weight k.

Proof. For all $|x| < 1$ we have the power series expansion

$$\frac{1}{(1-x)^2} = (1 + x + x^2 + \cdots)^2 = \sum_{n=0}^{\infty} (n + 1) x^n.$$

Applying this to $x = \frac{z}{\omega}$ with $|x| < 1$ (which we can assume holds for all $\omega \in L^*$ provided we keep z close to 0),

$$\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} = \frac{1}{\omega^2} \left(\frac{1}{(1-x)^2} - 1 \right) = \frac{1}{\omega^2} \sum_{n=1}^{\infty} (n + 1) x^n = \sum_{n=1}^{\infty} \frac{(n+1) z^n}{\omega^{n+2}}.$$

Summing over ω and changing the order of summation (via absolute convergence) gives

$$\begin{aligned}\wp(z) &= \frac{1}{z^2} + \sum_{\omega\in L^*}\left[\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right]\\ &= \frac{1}{z^2} + \sum_{\omega\in L^*}\sum_{n=1}^{\infty}\frac{(n+1)z^n}{\omega^{n+2}}\\ &= \frac{1}{z^2} + \sum_{n=1}^{\infty}(n+1)z^n\sum_{\omega\in L^*}\frac{1}{\omega^{n+2}}\\ &= \frac{1}{z^2} + \sum_{n=1}^{\infty}(n+1)G_{n+2}(L)z^n\\ &= \frac{1}{z^2} + \sum_{n=1}^{\infty}(2n+1)G_{2n+2}(L)z^{2n}.\end{aligned}$$

In the last step we used the fact that $\wp(z)$ is an even function, so the coefficients of the odd terms are 0 and we can sum over $2n$ rather than n. □

14.5 Lattices define elliptic curves

The key link between $\wp(z)$ and elliptic curves is given by the following differential equation.

Theorem 14.29. *The function $\wp(z) = \wp(z; L)$ satisfies the differential equation*

$$\wp'(z)^2 = 4\wp(z)^3 - g_2(L)\wp(z) - g_3(L), \tag{1}$$

where $g_2(L) \coloneqq 60G_4(L)$ and $g_3(L) \coloneqq 140G_6(L)$.

Proof. We may apply Theorem 14.28 to compute the first few terms of the Laurent series expansions for $\wp(z)$ and $\wp'(z)$ at $z_0 = 0$:

$$\begin{aligned}\wp(z) &= \frac{1}{z^2} + 3G_4(L)z^2 + 5G_6(L)z^4 + \cdots\\ \wp'(z) &= -\frac{2}{z^3} + 6G_4(L)z + 20G_6(L)z^3 + \cdots\\ \wp(z)^3 &= \frac{1}{z^6} + \frac{9G_4(L)}{z^2} + 15G_6(L) + \cdots\\ \wp'(z)^2 &= \frac{4}{z^6} - \frac{24G_4(L)}{z^2} - 80G_6(L) + \cdots\end{aligned}$$

Now let

$$f(z) = \wp'(z)^2 - 4\wp(z)^3 + 60G_4(L)\wp(z) + 140G_6(L).$$

We can compute the Laurent series expansion for $f(z)$ at $z_0 = 0$ as a linear combination of those computed above, and one finds that the non-positive powers of z all cancel; we thus have $f(0) = 0$.

Because $\wp$ and $\wp'$ have poles only at points of L, the function $f(z)$ is holomorphic on the fundamental parallelogram $\mathcal{F}_0$. The function $f(z)$ is periodic with respect to L, since $\wp(z)$ and $\wp'(z)$ are, thus it is holomorphic on the entire complex plane. Note that $f(z)$ is bounded because all values attained by f are attained on the closure of a fundamental parallelogram, which is a compact set. It then follows from Liouville's Theorem (see Theorem 14.30 below) that f is a constant function, hence identically zero. □

Theorem 14.30 (Liouville's Theorem)**.** *The only functions that are bounded and holomorphic on $\mathbb{C}$ are constant functions.*

Proof. See [1, p. 122] or [6, §2 Cor. 4.5]. □

With $y = \wp'(z)$ and $x = \wp(z)$, the differential equation in (1) corresponds to the curve

$$y^2 = 4x^3 - g_2(L)x - g_3(L). \tag{2}$$

This curve can easily be put into Weierstrass form with $g_2(L) = -4A$ and $g_3(L) = -4B$, thus every lattice L gives us an equation we can use to define an elliptic curve over $\mathbb{C}$, provided we can show that the projective curve defined by (2) is not singular. If the partial derivatives of $zy^2 = 4x^3 - g_2(L)xz^2 - g_3(L)z^3$ simultaneously vanish at some point, then there must be a projective solution to the system of equations

$$12x^2 - g_2(L)z^2 = 0, \qquad 2zy = 0, \qquad y^2 + 2g_2(L)xz + 3g_3(L)z^2 = 0.$$

We cannot have $z = 0$, since this would force $x = y = 0$, thus we assume $z = 1$. The second equation then implies $y = 0$ and the third equation forces $x = -3g_3(L)/(2g_2(L))$. Plugging these values into the first equation yields $g_2(L)^3 - 27g_3(L)^2 = 0$. Thus so long as

$$\Delta(L) := g_2(L)^3 - 27g_3(L)^2$$

is nonzero, equation (2) defines an elliptic curve over $\mathbb{C}$.

We will prove that $\Delta(L) \neq 0$, for every lattice L. For this we need the following lemma.

Lemma 14.31. *A point $z \notin L$ is a zero of $\wp'(z; L)$ if and only if $2z \in L$.*

Proof. Suppose $2z \in L$ for some $z \notin L$. Then

$$\wp'(z) = \wp'(z - 2z) = \wp'(-z) = -\wp'(z) = 0,$$

where we have used the fact that $\wp'(z)$ is both periodic with respect to L and an odd function. If $L = [\omega_1, \omega_2]$, then

$$\frac{\omega_1}{2}, \quad \frac{\omega_2}{2}, \quad \frac{\omega_1 + \omega_2}{2}$$

are the only points $z \in \mathcal{F}_0$ that are not in L and also satisfy $2z \in L$. Since $\wp'(z)$ is an elliptic function of order 3, it has only these three zeros in $\mathcal{F}_0$, by Theorem 14.18. Thus for any $z \notin L$ we have $\wp'(z) = 0$ if only if $2z \in L$. □

This lemma is analogous to the fact that the points of order 2 on the elliptic curve (2) are precisely the points $(x, y) = (\wp(z), \wp'(z))$ with $y = \wp'(z) = 0$. The requirement that $z \notin L$ simply means that (x, y) is not the point at infinity.

Remark 14.32. Having shown that the zeros of $\wp'(z)$ correspond to 2-torsion point of $\mathbb{C}/L$ you might wonder about the zeros of $\wp(z)$. As shown by Eichler and Zagier, the zeros of $\wp(z)$ in the fundamental region $\mathcal{F}_0$ for $L = [1, \tau]$ are

$$\frac{1}{2} \pm \left(\frac{\log(5 + 2\sqrt{6})}{2\pi i} + 144\pi i\sqrt{6} \int_\tau^{i\infty} (x - \tau) \frac{\Delta(x)}{G_6(x)^{3/2}} dx \right),$$

a fact that does not appear to have any obvious arithmetic significance.

Lemma 14.33. *For any lattice L, the discriminant $\Delta(L)$ is nonzero.*

Proof. Let $L = [\omega_1, \omega_2]$ and put

$$r_1 := \frac{\omega_1}{2}, \quad r_2 := \frac{\omega_2}{2}, \quad r_3 := \frac{\omega_1 + \omega_2}{2}.$$

Then $r_i \notin L$ and $2r_i \in L$ for $i = 1, 2, 3$. So $\wp'(r_i) = 0$ by Lemma 14.31. From (2) we see that $\wp(r_1), \wp(r_2)$, and $\wp(r_3)$ are the zeros of the cubic $f(x) = 4x^3 - g_2(L)x - g_3(L)$. Now the discriminant $\Delta(f)$ of $f(x)$ is equal to $16\Delta(L)$, thus

$$\Delta(L) = \frac{1}{16} \prod_{i<j} (\wp(r_i) - \wp(r_j))^2,$$

and it suffices to show that the $\wp(r_i)$ are distinct.

Let $g_i(z) = \wp(z) - \wp(r_i)$. Then $g_i(z)$ is an elliptic function of order 2 (its poles are the poles of $\wp(z)$), so it has exactly 2 zeros, by Theorem 14.18. Now r_i is a double zero because $g_i'(r_i) = \wp'(r_i) = 0$, by Lemma 14.31. Thus $g_i(z)$ has no other zeros, and therefore $\wp(r_j) \neq \wp(r_i)$ for $i \neq j$. □

We have shown that every lattice L in $\mathbb{C}$ gives rise to an elliptic curve $E/\mathbb{C}$ defined by $y^2 = 4x^3 - g_2(L)x - g_3(L)$, and that the map

$$\begin{aligned} \Phi \colon \mathbb{C}/L &\longrightarrow E(\mathbb{C}) \\ z &\longrightarrow (\wp(z), \wp'(z)) \end{aligned}$$

sends points on $\mathbb{C}/L$ to points on the elliptic curve. This is the first step in proving the Uniformization Theorem. In the next lecture we will show that Φ is a group isomorphism and that every elliptic curve $E/\mathbb{C}$ arises from some lattice L.

References

[1] Lars V. Ahlfors, *Complex analysis*, third edition, McGraw Hill, 1979.

[2] David A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory, and complex multiplication*, second edition, Wiley, 2013.

[3] Martin Eichler and Don Zagier, *On the zeros of the Weierstrass $\wp$-function*, Math. Annalen **258** (1982), 399-407.

[4] Serge Lang, *Complex analysis*, fourth edition, Springer, 1999.

[5] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer 2009.

[6] Elias M. Stein and Rami Shakarchi, *Complex analysis*, Princeton University Press, 2003.

15 Elliptic curves over $\mathbb{C}$ (part 2)

Last time we showed that every lattice $L \subseteq \mathbb{C}$ gives rise to an elliptic curve

$$E_L \colon y^2 = 4x^3 - g_2(L)x - g_3(L),$$

where

$$g_2(L) := 60G_4(L) := 60\sum_{L^*} \frac{1}{\omega^4}, \qquad g_3(L) := 140G_6(L) := 140\sum_{L^*} \frac{1}{\omega^6},$$

with $L^* := L - \{0\}$, and we defined a map

$$\begin{aligned} \Phi \colon \mathbb{C}/L &\to E_L(\mathbb{C}) \\ z &\mapsto \begin{cases} (\wp(z), \wp'(z)) & z \notin L \\ 0 & z \in L \end{cases} \end{aligned}$$

where

$$\wp(z) = \wp(z; L) = \frac{1}{z^2} + \sum_{\omega \in L^*} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

is the Weierstrass $\wp$-function for the lattice L, and

$$\wp'(z) = -2\sum_{\omega \in L} \frac{1}{(z-\omega)^3}.$$

In this lecture we will prove two theorems. First we will prove that Φ is an isomorphism of additive groups; it is also an isomorphism of complex manifolds [3, Cor. 5.1.1], and of complex Lie groups, but we won't prove this right now.[1] Second, we will prove that every elliptic curve $E/\mathbb{C}$ is isomorphic to E_L for some lattice L; this is the *Uniformization Theorem*.

15.1 The isomorphism from a torus to the corresponding elliptic curve

Theorem 15.1. *Let $L \subseteq \mathbb{C}$ be a lattice and let $E_L \colon y^2 = 4x^3 - g_2(L)x - g_3(L)$ be the corresponding elliptic curve. The map $\Phi \colon \mathbb{C}/L \to E_L(\mathbb{C})$ is a group isomorphism.*

Proof. We first note that $\Phi(0) = 0$, so Φ preserves the identity, and for all $z \notin L$ we have

$$\Phi(-z) = (\wp(-z), \wp'(-z)) = \big(\wp(z), -\wp'(z)\big) = -\Phi(z),$$

since $\wp$ is even and $\wp'$ is odd, so Φ is compatible with taking inverses.

Let $L = [\omega_1, \omega_2]$. There are three points of order 2 in $\mathbb{C}/L$: $\omega_1/2, \omega_2/2$, and $(\omega_1+\omega_2)/2$. By Lemma 14.31, $\wp'$ vanishes at these points, hence Φ maps points of order 2 in $\mathbb{C}/L$ to points of order 2 in $E_L(\mathbb{C})$, since the latter are the points with y-coordinate zero. Moreover, Φ is injective on points of order 2, since $\wp(z)$ maps each point of order 2 in $\mathbb{C}/L$ to a distinct root of $4\wp(z)^3 - g_2(L)\wp(z) - g_3(L)$, as shown in the proof of Lemma 14.33. The restriction of Φ to $(\mathbb{C}/L)[2]$ defines a bijection of $(\mathbb{C}/L)[2] \xrightarrow{\sim} E_L[2] \simeq \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ with $\Phi(0) = 0$, which must be a group isomorphism.

[1] This is not difficult to show, but it would distract us from our immediate goal. We will see an explicit isomorphism of complex manifolds in a few lectures when we study modular curves, and in that case we will take the time to define precisely what this means and to prove it.

To show that Φ is surjective, let $(x_0, y_0) \in E_L(\mathbb{C})$. The elliptic function $f(z) = \wp(z) - x_0$ has order 2, hence it has two zeros in the fundamental parallelogram $\mathcal{F}_0$, by Theorem 14.18. Neither of these zeros occurs at $z = 0$, since f has a pole at 0. So let $z_0 \neq 0$ be a zero of $f(z)$ in $\mathcal{F}_0$. Then $\wp(z_0) = x_0$, which implies $\Phi(z_0) = (x_0, \pm y_0)$ and therefore $(x_0, y_0) = \Phi(\pm z_0)$; thus Φ is surjective.

We now show that Φ is injective. Let $z_1, z_2 \in \mathcal{F}_0$ and suppose that $\Phi(z_1) = \Phi(z_2)$. If $2z_1 \in L$ then z_1 is a 2-torsion element and we have already shown that Φ restricts to a bijection on $(\mathbb{C}/L)[2]$, so we must have $z_1 = z_2$. We now assume $2z_1 \notin L$, which implies $\wp'(z_1) \neq 0$. As argued above, the roots of $f(z) = \wp(z) - \wp(z_1)$ in $\mathcal{F}_0$ are $\pm z_1$, thus $z_2 \equiv \pm z_1 \bmod L$. We also have $\wp'(z_1) = \wp'(z_2)$, and this forces $z_2 \equiv z_1 \bmod L$, since $\wp'(-z_1) = -\wp'(z_1) \neq \wp'(z_1)$ because $\wp'(z_1) \neq 0$.

It remains only to show that $\Phi(z_1 + z_2) = \Phi(z_1) + \Phi(z_2)$. So let $z_1, z_2 \in \mathcal{F}_0$; we may assume that $z_1, z_2, z_1 + z_2 \notin L$ since the case where either z_1 or z_2 lies in L is immediate, and if $z_1 + z_2 \in L$ then z_1 and z_2 are inverses modulo L, a case treated above.

The points $P_1 = \Phi(z_1)$ and $P_2 = \Phi(z_2)$ are affine points in $E_L(\mathbb{C})$, and the line ℓ between them cannot be vertical because P_1 and P_2 are not inverses (since z_1 and z_2 are not). So let $y = mx + b$ be an equation for this line, and let P_3 be the third point where the line intersects the curve E_L. Then $P_1 + P_2 + P_3 = 0$, by the definition of the group law on $E_L(\mathbb{C})$.

Now consider the function $\ell(z) = -\wp'(z) + m\wp(z) + b$. It is an elliptic function of order 3 with a triple pole at 0, so it has three zeros in the fundamental parallelogram $\mathcal{F}_0$, two of which are z_1 and z_2. Let z_3 be the third zero in $\mathcal{F}_0$. The point $\Phi(z_3)$ lies on both the line ℓ and the elliptic curve $E_L(C)$, hence it must lie in $\{P_1, P_2, P_3\}$; moreover, we have a bijection from $\{z_1, z_2, z_3\}$ to $\{\Phi(z_1), \Phi(z_2), \Phi(z_3)\} = \{P_1, P_2, P_3\}$, and this bijection must send z_3 to P_3 if P_3 is distinct from P_1 and P_2. If P_3 coincides with exactly one of P_1 or P_2, say P_1, then $\ell(z)$ has a double zero at z_1 and we must have $z_3 = z_1$; and if $P_1 = P_2 = P_3$ then clearly $z_1 = z_2 = z_3$. Thus in every case we must have $P_3 = \Phi(z_3)$.

We have $P_1 + P_2 + P_3 = 0$, so it suffices to show $z_1 + z_2 + z_3 \in L$, since this implies

$$\Phi(z_1 + z_2) = \Phi(-z_3) = -\Phi(z_3) = -P_3 = P_1 + P_2 = \Phi(z_1) + \Phi(z_2).$$

Let $\mathcal{F}_\alpha$ be a fundamental parallelogram for L whose boundary does not contain any zeros or poles of $\ell(z)$ and replace z_1, z_2, z_3 by equivalent points in $\mathcal{F}_\alpha$ if necessary.

Applying Theorem 14.17 to $g(z) = z$ and $f(z) = \ell(z)$ yields

$$\frac{1}{2\pi i}\int_{\partial\mathcal{F}_\alpha} z\frac{\ell'(z)}{\ell(z)}dz = \sum_{w\in F_\alpha} \operatorname{ord}_w(\ell)w = z_1 + z_2 + z_3 - 3\cdot 0 = z_1 + z_2 + z_3, \tag{1}$$

where the boundary $\partial\mathcal{F}_\alpha$ of $\mathcal{F}_\alpha$ is oriented counter-clockwise.

Let us now evaluate the integral in (1); to ease the notation, define $f(z) := \ell'(z)/\ell(z)$, which we note is an elliptic function (hence periodic with respect to L). We then have

$$\begin{aligned}
\int_{\partial\mathcal{F}_\alpha} zf(z)\,dz &= \int_\alpha^{\alpha+\omega_1} zf(z)dz + \int_{\alpha+\omega_1}^{\alpha+\omega_1+\omega_2} zf(z)dz + \int_{\alpha+\omega_1+\omega_2}^{\alpha+\omega_2} zf(z)dz + \int_{\alpha+\omega_2}^{\alpha} zf(z)dz \\
&= \int_\alpha^{\alpha+\omega_1} zf(z)dz + \int_\alpha^{\alpha+\omega_2} (z+\omega_1)f(z)dz + \int_{\alpha+\omega_1}^{\alpha} (z+\omega_2)f(z)dz + \int_{\alpha+\omega_2}^{\alpha} zf(z)dz \\
&= \omega_1\int_\alpha^{\alpha+\omega_2} f(z)dz + \omega_2\int_{\alpha+\omega_1}^{\alpha} f(z)dz.
\end{aligned} \tag{2}$$

Note that we have used the periodicity of $f(z)$ to replace $f(z+\omega_i)$ by $f(z)$, and to cancel integrals in opposite directions along lines that are equivalent modulo L.

For any closed (not necessarily simple) curve C and a point $z_0 \notin C$, the quantity

$$\frac{1}{2\pi i}\int_C \frac{dz}{z-z_0}$$

is the *winding number* of C about z_0, and it is an integer (it counts the number of times the curve C "winds around" the point z_0); see [1, Lem. 4.2.1] or [4, Lem. B.1.3].

The function $\ell(\alpha+t\omega_2)$ parametrizes a closed curve C_1 from $\ell(\alpha)$ to $\ell(\alpha+\omega_2)=\ell(\alpha)$, as t ranges from 0 to 1. The winding number of C_1 about the point 0 is the integer

$$c_1 := \frac{1}{2\pi i}\int_{C_1}\frac{dz}{z-0} = \frac{1}{2\pi i}\int_0^1 \frac{\ell'(\alpha+t\omega_2)}{\ell(\alpha+t\omega_2)}\omega_2 dt = \frac{1}{2\pi i}\int_\alpha^{\alpha+\omega_2}\frac{\ell'(z)}{\ell(z)}dz = \frac{1}{2\pi i}\int_\alpha^{\alpha+\omega_2} f(z)dz. \quad (3)$$

Similarly, the function $\ell(\alpha+t\omega_1)$ parameterizes a closed curve C_2 from $\ell(\alpha)$ to $\ell(\alpha+\omega_1)$, and we obtain the integer

$$c_2 := \frac{1}{2\pi i}\int_{C_2}\frac{dz}{z-0} = \frac{1}{2\pi i}\int_0^1 \frac{\ell'(\alpha+t\omega_1)}{\ell(\alpha+t\omega_1)}\omega_1 dt = \frac{1}{2\pi i}\int_\alpha^{\alpha+\omega_1}\frac{\ell'(z)}{\ell(z)}dz = \frac{1}{2\pi i}\int_\alpha^{\alpha+\omega_1} f(z)\,dz. \quad (4)$$

Plugging (3), and (4) into (2), and applying (1), we see that

$$z_1+z_2+z_3 = c_1\omega_1 - c_2\omega_2 \in L,$$

as desired. □

15.2 The j-invariant of a lattice

Definition 15.2. The *j-invariant* of a lattice L is defined by

$$j(L) = 1728\frac{g_2(L)^3}{\Delta(L)} = 1728\frac{g_2(L)^3}{g_2(L)^3-27g_3(L)^2}.$$

Recall that $\Delta(L)\neq 0$, by Lemma 14.33, so $j(L)$ is always defined.

The elliptic curve $E_L\colon y^2 = 4x^3 - g_2(L)x - g_3(L)$ is isomorphic to the elliptic curve $y^2 = x^3 + Ax + B$, where $g_2(L) = -4A$ and $g_3(L) = -4B$. Thus

$$j(L) = 1728\frac{g_2(L)^3}{g_2(L)^3-27g_3(L)^2} = 1728\frac{(-4A)^3}{(-4A)^3-27(-4B)^2} = 1728\frac{4A^3}{4A^3+27B^2} = j(E_L).$$

Thus the j-invariant of a lattice L is the same as the j-invariant of the corresponding elliptic curve E_L. We now define the discriminant of an elliptic curve so that it agrees with the discriminant of the corresponding lattice.

Definition 15.3. The *discriminant* of an elliptic curve $E\colon y^2 = x^3 + Ax + B$ is

$$\Delta(E) = -16(4A^3+27B^2).$$

This definition applies to any elliptic curve E/k defined by a short Weierstrass equation, whether $k=\mathbb{C}$ or not, but for the moment we continue to focus on elliptic curves over $\mathbb{C}$.

Recall from Theorem 13.14 that elliptic curves E/k and E'/k are isomorphic over $\bar{k}$ if and only if $j(E) = j(E')$. Thus over an algebraically closed field like $\mathbb{C}$, the j-invariant characterizes elliptic curves up to isomorphism. We now define an analogous notion of isomorphism for lattices.

Definition 15.4. Lattices L and L' are said to be *homothetic* if $L' = \lambda L$ for some $\lambda \in \mathbb{C}^\times$.

Theorem 15.5. *Two lattices L and L' are homothetic if and only if $j(L) = j(L')$.*

Proof. Suppose L and L' are homothetic, with $L' = \lambda L$. Then

$$g_2(L') \;=\; 60 \sum_{\omega \in L'^*} \frac{1}{\omega^4} \;=\; 60 \sum_{\omega \in L^*} \frac{1}{(\lambda\omega)^4} \;=\; \lambda^{-4} g_2(L).$$

Similarly, $g_3(L') = \lambda^{-6} g_3(L)$, and we have

$$j(L') = 1728 \frac{(\lambda^{-4} g_2(L))^3}{(\lambda^{-4} g_2(L))^3 - 27(\lambda^{-6} g_3(L))^2} = 1728 \frac{g_2(L)^3}{g_2(L)^3 - 27 g_3(L)^2} = j(L).$$

To show the converse, let us now assume $j(L) = j(L')$. Let E_L and $E_{L'}$ be the corresponding elliptic curves. Then $j(E_L) = j(E_{L'})$. We may write

$$E_L \colon y^2 = x^3 + Ax + B,$$

with $-4A = g_2(L)$ and $-4B = g_3(L)$, and similarly for $E_{L'}$, with $-4A' = g_2(L')$ and $-4B' = g_3(L')$. By Theorem 13.13, there is a $\mu \in \mathbb{C}^\times$ such that $A' = \mu^4 A$ and $B' = \mu^6 B$, and if we let $\lambda = 1/\mu$, then $g_2(L') = \lambda^{-4} g_2(L) = g_2(\lambda L)$ and $g_3(L') = \lambda^{-6} g_3(L) = g_3(\lambda L)$, as above. We now show that this implies $L' = \lambda L$.

Recall from Theorem 14.29 that the Weierstrass $\wp$-function satisfies

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3.$$

Differentiating both sides yields

$$\begin{aligned} 2\wp'(z)\wp''(z) &= 12\wp(z)^2\wp'(z) - g_2\wp'(z) \\ \wp''(z) &= 6\wp(z)^2 - \frac{g_2}{2}. \end{aligned} \tag{5}$$

By Theorem 14.28, the Laurent series for $\wp(z; L)$ at $z = 0$ is

$$\wp(z) = \frac{1}{z^2} + \sum_{n=1}^{\infty} (2n+1) G_{2n+2} z^{2n} = \frac{1}{z^2} + \sum_{n=1}^{\infty} a_n z^{2n},$$

where $a_1 = g_2/20$ and $a_2 = g_3/28$.

Comparing coefficients for the z^{2n} term in (5), we find that for $n \geq 2$ we have

$$(2n+2)(2n+1)a_{n+1} = 6\left(\sum_{k=1}^{n-1} a_k a_{n-k} + 2a_{n+1}\right),$$

and therefore

$$a_{n+1} = \frac{6}{(2n+2)(2n+1) - 12} \sum_{k=1}^{n-1} a_k a_{n-k}.$$

This allows us to compute a_{n+1} from $a_1, \ldots, a_{n-1}$, for all $n \geq 2$. It follows that $g_2(L)$ and $g_3(L)$ uniquely determine the function $\wp(z) = \wp(z; L)$ (and therefore the lattice L where $\wp(z)$ has poles), since $\wp(z)$ is uniquely determined by its Laurent series expansion about 0.

Now consider L' and λL, where we have $g_2(L') = g_2(\lambda L)$ and $g_3(L') = g_3(\lambda L)$. It follows that $\wp(z; L') = \wp(z; \lambda L)$ and $L' = \lambda L$, as desired. □

Corollary 15.6. *Two lattices L and L' are homothetic if and only if the corresponding elliptic curves E_L and $E_{L'}$ are isomorphic.*

Thus homothety classes of lattices correspond to isomorphism classes of elliptic curves over $\mathbb{C}$, and both are classified by the j-invariant. Recall from Theorem 13.12 that every complex number is the j-invariant of an elliptic curve $E/\mathbb{C}$. To prove the Uniformization Theorem we just need to show that the same is true of lattices.

15.3 The j-function

Every lattice $[\omega_1, \omega_2]$ is homothetic to a lattice of the form $[1, \tau]$, with τ in the upper half plane $\mathcal{H} := \{z \in \mathbb{C} : \operatorname{im} z > 0\}$; we may take $\tau = \pm\omega_2/\omega_1$ with the sign chosen so that $\operatorname{im}\tau > 0$. This leads to the following definition of the j-function.

Definition 15.7. The *j-function* $j\colon \mathcal{H} \to \mathbb{C}$ is defined by $j(\tau) = j([1,\tau])$. We similarly define $g_2(\tau) = g_2([1,\tau])$, $g_3(\tau) = g_3([1,\tau])$, and $\Delta(\tau) = \Delta([1,\tau])$.

Note that for any $\tau \in \mathcal{H}$, both $-1/\tau$ and $\tau + 1$ lie in $\mathcal{H}$ (the maps $\tau \mapsto 1/\tau$ and $\tau \mapsto -\tau$ both swap the upper and lower half-planes; their composition preserves them).

Theorem 15.8. *The j-function is holomorphic on $\mathcal{H}$, and satisfies $j(-1/\tau) = j(\tau)$ and $j(\tau+1) = j(\tau)$.*

Proof. From the definition of $j(\tau) = j([1,\tau])$ we have

$$j(\tau) = 1728\frac{g_2(\tau)^3}{\Delta(\tau)} = 1728\frac{g_2(\tau)^3}{g_2(\tau)^3 - 27g_3(\tau)^2}.$$

The series defining

$$g_2(\tau) = 60 \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0}} \frac{1}{(m+n\tau)^4} \qquad \text{and} \qquad g_3(\tau) = 140 \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}} \frac{1}{(m+n\tau)^6}$$

converge absolutely for any fixed $\tau \in \mathcal{H}$, by Lemma 14.22, and they converge uniformly over τ in any compact subset of $\mathcal{H}$. The proof of this last fact is straight-forward but slightly technical; see [2, Thm. 1.15] for the details. It follows that $g_2(\tau)$ and $g_3(\tau)$ are holomorphic on $\mathcal{H}$, and therefore $\Delta(\tau) = g_2(\tau)^3 - 27g_3(\tau)^2$ is also holomorphic on $\mathcal{H}$. Since $\Delta(\tau)$ is nonzero for all $\tau \in \mathcal{H}$, by Lemma 14.33, the j-function $j(\tau)$ is holomorphic on $\mathcal{H}$ as well.

The lattices $[1,\tau]$ and $[1,-1/\tau] = -1/\tau[1,\tau]$ are homothetic, and the lattices $[1,\tau+1]$ and $[1,\tau]$ are equal; thus $j(-1/\tau) = j(\tau)$ and $j(\tau+1) = j(\tau)$, by Theorem 15.5. □

15.4 The modular group

We now consider the *modular group*

$$\Gamma = \mathrm{SL}_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a,b,c,d \in \mathbb{Z},\ ad - bc = 1 \right\}.$$

As proved in Problem Set 8, the group Γ acts on $\mathcal{H}$ via linear fractional transformations

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tau = \frac{a\tau + b}{c\tau + d},$$

and it is generated by the matrices $S = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. This implies that the j-function is invariant under the action of the modular group; in fact, more is true.

Lemma 15.9. *For any $\tau, \tau' \in \mathcal{H}$ we have $j(\tau) = j(\tau')$ if and only if $\tau' = \gamma\tau$ for some $\gamma \in \Gamma$.*

Proof. We have $j(S\tau) = j(-1/\tau) = j(\tau)$ and $j(T\tau) = j(\tau+1) = j(\tau)$, by Theorem 15.8. It follows that if $\tau' = \gamma\tau$ then $j(\tau') = j(\tau)$, since S and T generate Γ.

To prove the converse, let us suppose that $j(\tau) = j(\tau')$. Then by Theorem 15.5, the lattices $[1,\tau]$ and $[1,\tau']$ are homothetic So $[1,\tau'] = \lambda[1,\tau]$, for some $\lambda \in \mathbb{C}^\times$. There thus exist integers a, b, c, and d such that

$$\begin{aligned}\tau' &= a\lambda\tau + b\lambda\\ 1 &= c\lambda\tau + d\lambda\end{aligned}$$

From the second equation, we see that $\lambda = \frac{1}{c\tau+d}$. Substituting this into the first, we have

$$\tau' = \frac{a\tau+b}{c\tau+d} = \gamma\tau, \qquad \text{where } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbb{Z}^{2\times 2}.$$

Now $[1,\tau'] = \lambda[1,\tau]$ implies $\operatorname{im}\tau' = |\lambda|^2 \operatorname{im}\tau$, since $\tau, \tau' \in \mathcal{H}$ and fundamental parallelograms for $[1,\tau'] = \lambda[1,\tau]$ must have the same area. But we also have

$$\operatorname{im}\tau' = \operatorname{im}(\gamma\tau) = \operatorname{im}\left(\frac{a\tau+b}{c\tau+d}\right) = \frac{\operatorname{im}\big((a\tau+b)(c\bar\tau+d)\big)}{|c\tau+d|^2} = \frac{(ad-bc)\operatorname{im}\tau}{|c\tau+d|^2} = (\det\gamma)|\lambda|^2\operatorname{im}\tau,$$

and therefore $\det\gamma = 1$ and $\gamma \in \mathrm{SL}_2(\mathbb{Z})$. □

Lemma 15.9 implies that when studying the j-function it suffices to study its behavior on Γ-equivalence classes of $\mathcal{H}$, that is, the orbits of $\mathcal{H}$ under the action of Γ. We thus consider the quotient of $\mathcal{H}$ modulo Γ-equivalence, which we denote by $\mathcal{H}/\Gamma$.[2] The actions of γ and $-\gamma$ are identical, so taking the quotient by $\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm 1\}$ yields the same result, but for the sake of clarity we will stick with $\Gamma = \mathrm{SL}_2(\mathbb{Z})$.

We now wish to determine a fundamental domain for $\mathcal{H}/\Gamma$, a set of unique representatives in $\mathcal{H}$ for each Γ-equivalence class. For this purpose we will use the set

$$\mathcal{F} = \{\tau \in \mathcal{H} : \operatorname{re}(\tau) \in [-1/2, 1/2) \text{ and } |\tau| \geq 1, \text{ such that } |\tau| > 1 \text{ if } \operatorname{re}(\tau) > 0\}.$$

Lemma 15.10. *The set $\mathcal{F}$ is a fundamental domain for $\mathcal{H}/\Gamma$.*

Proof. We need to show that for every $\tau \in \mathcal{H}$, there is a unique $\tau' \in \mathcal{F}$ such that $\tau' = \gamma\tau$, for some $\gamma \in \Gamma$. We first prove existence. Let us fix $\tau \in \mathcal{H}$. For any $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ we have

$$\operatorname{im}(\gamma\tau) = \operatorname{im}\left(\frac{a\tau+b}{c\tau+d}\right) = \frac{\operatorname{im}\big((a\tau+b)(c\bar\tau+d)\big)}{|c\tau+d|^2} = \frac{(ad-bc)\operatorname{im}\tau}{|c\tau+d|^2} = \frac{\operatorname{im}\tau}{|c\tau+d|^2} \tag{6}$$

Let $c\tau + d$ be a shortest vector in the lattice $[1,\tau]$. Then c and d must be relatively prime, and we can pick integers a and b so that $ad - bc = 1$. The matrix $\gamma_0 = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ then maximizes the value of $\operatorname{im}(\gamma\tau)$ over $\gamma \in \Gamma$. Let us now choose $\gamma = T^k\gamma_0$, where k is chosen so that $\operatorname{re}(\gamma\tau) \in [1/2, 1/2)$, and note that $\operatorname{im}(\gamma\tau) = \operatorname{im}(\gamma_0\tau)$ remains maximal. We must have $|\gamma\tau| \geq 1$, since otherwise $\operatorname{im}(S\gamma\tau) > \operatorname{im}(\gamma\tau)$, contradicting the maximality of $\operatorname{im}(\gamma\tau)$.

[2] Some authors write this quotient as $\Gamma\backslash\mathcal{H}$ to indicate that the action is on the left.

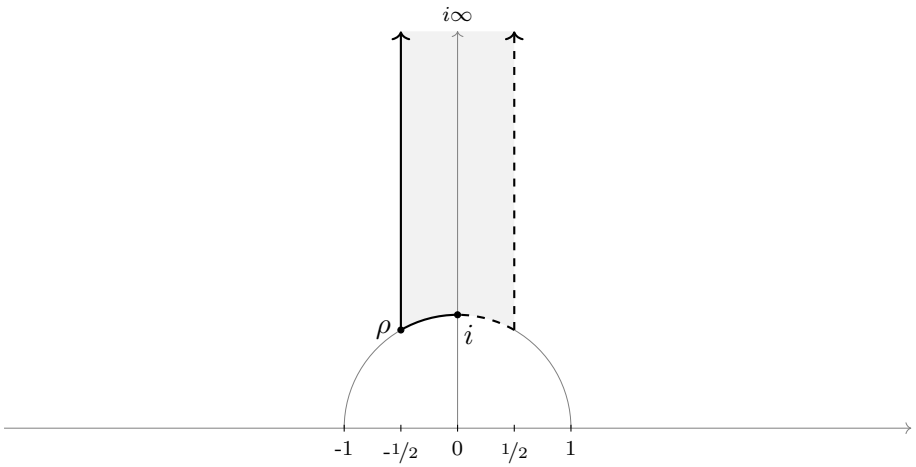


Figure 1: Fundamental domain $\mathcal{F}$ for $\mathcal{H}/\Gamma$, with $i = e^{\pi i/2}$ and $\rho = e^{2\pi i/3}$.

Finally, if $\tau' = \gamma\tau \notin \mathcal{F}$, then we must have $|\gamma\tau| = 1$ and $\mathrm{re}(\gamma\tau) > 0$, in which case we replace γ by $S\gamma$ so that $\tau' = \gamma\tau \in \mathcal{F}$.

It remains to show that τ' is unique. This is equivalent to showing that any two Γ-equivalent points in $\mathcal{F}$ must coincide. So let τ_1 and $\tau_2 = \gamma_1\tau_1$ be two elements of $\mathcal{F}$, with $\gamma_1 = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, and assume $\mathrm{im}\,\tau_1 \le \mathrm{im}\,\tau_2$. By (6), we must have $|c\tau_1 + d|^2 \le 1$, thus

$$1 \ge |c\tau_1 + d|^2 = (c\tau_1 + d)(c\bar{\tau}_1 + d) = c^2|\tau_1|^2 + d^2 + 2cd\,\mathrm{re}\,\tau_1 \ge c^2|\tau_1|^2 + d^2 - |cd| \ge 1,$$

where the last inequality follows from $|\tau_1| \ge 1$ and the fact that c and d cannot both be zero (since $\det\gamma = 1$). Thus $|c\tau_1 + d| = 1$, which implies $\mathrm{im}\,\tau_2 = \mathrm{im}\,\tau_1$. We also have $|c|, |d| \le 1$, and by replacing γ_1 by $-\gamma_1$ if necessary, we may assume that $c \ge 0$. This leaves 3 cases:

1. $c = 0$: then $|d| = 1$ and $a = d$. So $\tau_2 = \tau_1 \pm b$, but $|\,\mathrm{re}\,\tau_2 - \mathrm{re}\,\tau_1| < 1$, so $\tau_2 = \tau_1$.
2. $c = 1, d = 0$: then $b = -1$ and $|\tau_1| = 1$. So τ_1 is on the unit circle and $\tau_2 = a - 1/\tau_1$. Either $a = 0$ and $\tau_2 = \tau_1 = i$, or $a = -1$ and $\tau_2 = \tau_1 = \rho$.
3. $c = 1, |d| = 1$: then $|\tau_1 + d| = 1$, so $\tau_1 = \rho$, and $\mathrm{im}\,\tau_2 = \mathrm{im}\,\tau_1 = \sqrt{3}/2$ implies $\tau_2 = \rho$.

In every case we have $\tau_1 = \tau_2$ as desired. □

Theorem 15.11. *The restriction of the j-function to $\mathcal{F}$ defines a bijection from $\mathcal{F}$ to $\mathbb{C}$.*

Proof. Injectivity follows immediately from Lemmas 15.9 and 15.10. It remains to prove surjectivity. We have

$$g_2(\tau) \;=\; 60 \sum_{\substack{n,m\in\mathbb{Z}\\(m,n)\neq(0,0)}} \frac{1}{(m+n\tau)^4} \;=\; 60\left(2\sum_{m=1}^{\infty}\frac{1}{m^4} + \sum_{\substack{n,m\in\mathbb{Z}\\n\neq 0}}\frac{1}{(m+n\tau)^4}\right).$$

The second sum tends to 0 as $\mathrm{im}\,\tau \to \infty$. Thus we have

$$\lim_{\mathrm{im}\tau\to\infty} g_2(\tau) = 120\sum_{m=1}^{\infty} m^{-4} = 120\,\zeta(4) = 120\,\frac{\pi^4}{90} = \frac{4\pi^4}{3},$$

where $\zeta(s)$ is the Riemann zeta function. Similarly,

$$\lim_{\mathrm{im}\tau\to\infty} g_3(\tau) = 280\,\zeta(6) = 280\,\frac{\pi^6}{945} = \frac{8\pi^6}{27}.$$

Thus

$$\lim_{\mathrm{im}\tau\to\infty} \Delta(\tau) = \left(\frac{4}{3}\pi^4\right)^3 - 27\left(\frac{8}{27}\pi^6\right)^2 = 0.$$

(this explains the coefficients 60 and 140 in the definitions of g_2 and g_3; they are the smallest pair of integers that ensure this limit is 0). Since $\Delta(\tau)$ is the denominator of $j(\tau)$, the quantity $j(\tau) = 1728 g_2(\tau)^3/\Delta(\tau)$ is unbounded as $\mathrm{im}\,\tau \to \infty$.

In particular, the j-function is non-constant, and by Theorem 15.8 it is holomorphic on $\mathcal{H}$. The open mapping theorem implies that $j(\mathcal{H})$ is an open subset of $\mathbb{C}$; see [4, Thm. 3.4.4].

We claim that $j(\mathcal{H})$ is also a closed subset of $\mathbb{C}$. Let $j(\tau_1), j(\tau_2), \ldots$ be an arbitrary convergent sequence in $j(\mathcal{H})$, converging to $w \in \mathbb{C}$. The j-function is Γ-invariant, by Lemma 15.9, so we may assume the τ_n all lie in $\mathcal{F}$. The sequence $\mathrm{im}\,\tau_1, \mathrm{im}\,\tau_2, \ldots$ must be bounded, say by B, since $j(\tau) \to \infty$ as $\mathrm{im}\,\tau \to \infty$, but the sequence $j(\tau_1), j(\tau_2), \ldots$ converges; it follows that the τ_n all lie in the compact set

$$\Omega = \{\tau : \mathrm{re}\,\tau \in [-1/2, 1/2], \mathrm{im}\,\tau \in [1/2, B]\}.$$

There is thus a subsequence of the τ_n that converges to some $\tau \in \Omega \subset \mathcal{H}$. The j-function is holomorphic, hence continuous, so $j(\tau) = w$. It follows that the open set $j(\mathcal{H})$ contains all its limit points and is therefore closed.

The fact that the non-empty set $j(\mathcal{H}) \subseteq \mathbb{C}$ is both open and closed implies that $j(\mathcal{H}) = \mathbb{C}$, since $\mathbb{C}$ is connected. It follows that $j(\mathcal{F}) = \mathbb{C}$, since every element of $\mathcal{H}$ is Γ-equivalent to an element of $\mathcal{F}$ (Lemma 15.10) and the j-function is Γ-invariant (Lemma 15.9). □

Corollary 15.12 (Uniformization Theorem)**.** *For every elliptic curve $E/\mathbb{C}$ there exists a lattice L such that $E = E_L$.*

Proof. Given $E/\mathbb{C}$, pick $\tau \in \mathcal{H}$ so that $j(\tau) = j(E)$ and let $L' = [1, \tau]$. We have

$$j(E) = j(\tau) = j(L') = j(E_{L'}),$$

so E is isomorphic to $E_{L'}$, by Theorem 13.13, where the isomorphism is given by the map $(x, y) \mapsto (\mu^2 x, \mu^3 y)$ for some $\mu \in \mathbb{C}^\times$. If we now let $L = \frac{1}{\mu}L'$, then $E = E_L$. □

References

[1] Lars V. Ahlfors, *Complex analysis*, third edition, McGraw Hill, 1979.

[2] Tom M. Apostol, *Modular functions and Dirichlet series in number theory*, second edition, Springer, 1990.

[3] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer 2009.

[4] Elias M. Stein and Rami Shakarchi, *Complex analysis*, Princeton University Press, 2003.

16 Complex multiplication

Over the course of the last two lectures we established a one-to-one correspondence between lattices $L \subseteq \mathbb{C}$ (up to homothety) and elliptic curves $E/\mathbb{C}$ (up to isomorphism), given by the map that sends each lattice L to the elliptic curve

$$E_L \colon y^2 = 4x^3 - g_2(L)x - g_3(L),$$

together with an explicit isomorphism

$$\begin{aligned} \Phi \colon \mathbb{C}/L &\to E_L(\mathbb{C}) \\ z &\mapsto \begin{cases} (\wp(z), \wp'(z)) & z \notin L; \\ 0 & z \in L, \end{cases} \end{aligned}$$

where $\wp(z)$ is the Weierstrass $\wp$-function for the lattice L.

To complete our understanding of the categorical equivalence of complex tori and elliptic curves, we want to relate morphisms of complex tori to isogenies of elliptic curves. In particular, we want to be able to explicitly understand how to relate the endomorphism ring of a complex torus to the endomorphism ring of the corresponding elliptic curve.

A complex torus $\mathbb{C}/L$ is both a complex manifold and a group in which the group operations are defined by holomorphic maps (this makes it a complex Lie group). A morphism in the category of complex tori must respect both structures: we require morphisms of complex tori to be holomorphic maps that are also group homomorphisms (just as isogenies are morphisms of algebraic varieties that are also homomorphisms of abelian groups).

16.1 Morphisms of complex tori

We have not formally defined what it means to be a holomorphic map of complex manifolds (or even a complex manifold), but for maps $\varphi \colon \mathbb{C}/L_1 \to \mathbb{C}/L_2$ of complex tori it simply means that φ is induced by a holomorphic function $f \colon \mathbb{C} \to \mathbb{C}$ that makes the following diagram commute:

$$\begin{array}{ccc} \mathbb{C} & \xrightarrow{f} & \mathbb{C} \\ \downarrow{\scriptstyle \pi_1} & & \downarrow{\scriptstyle \pi_2} \\ \mathbb{C}/L_1 & \xrightarrow{\varphi} & \mathbb{C}/L_2 \end{array}$$

where π_1 and π_2 are quotient maps.[1]

Each $\alpha \in \mathbb{C}$ determines a holomorphic multiplication-by-α map $z \mapsto \alpha z$ that is an endomorphism of $\mathbb{C}$ (as a group under addition). Whenever $\alpha L_1 \subseteq L_2$ this induces a group homomorphism

$$\begin{aligned} \varphi_\alpha \colon \mathbb{C}/L_1 &\to \mathbb{C}/L_2 \\ z + L_1 &\mapsto \alpha z + L_2 \end{aligned}$$

that is also a holomorphic map of complex manifolds.

Remarkably, every morphism of complex tori arises in this way. In fact, every holomorphic map that fixes zero arises in this way; this is analogous to the fact that every morphism of elliptic curves that fixes zero is automatically a group homomorphism.

[1] We should note that in general holomorphic maps of complex manifolds are defined locally on charts and need not be induced by a single global map; complex tori are a particularly simple special case.

Theorem 16.1. *Let $\varphi\colon \mathbb{C}/L_1 \to \mathbb{C}/L_2$ be a holomorphic map with $\varphi(0) = 0$. There is a unique $\alpha \in \mathbb{C}$ for which $\varphi = \varphi_\alpha$.*

Proof. Let $\pi_i\colon \mathbb{C} \to \mathbb{C}/L_i$ be quotient maps and let $f\colon \mathbb{C} \to \mathbb{C}$ be a holomorphic function for which $\varphi(\pi_1(z)) = \pi_2(f(z))$. For all $z \in \mathbb{C}$ and $\omega \in L_1$ we have

$$\pi_2(f(z+\omega)) = \varphi(\pi_1(z+\omega)) = \varphi(\pi_1(z)) = \pi_2(f(z)),$$

thus $f(z+\omega) - f(z) \in \ker \pi_2 = L_2$. For each $\omega \in L_1$ the function $g_\omega(z) := f(z+\omega) - f(z)$ is a continuous map from the connected set $\mathbb{C}$ to a discrete set L_2; its image must be connected and therefore consists of a single point. It follows that $g_\omega(z)$ is constant and $g'_\omega(z) = 0$, which implies that $f'(z+\omega) = f'(z)$ for all $z \in \mathbb{C}$ and $\omega \in L_1$. Thus $f'(z)$ is periodic with respect to L_1 and is therefore a holomorphic elliptic function, hence constant (see Remark 14.10).

Thus $f(z) = \alpha z + \beta$, for some $\alpha, \beta \in \mathbb{C}$. For all $\omega \in L_1$ we have

$$\pi_2(f(\omega)) = \varphi(\pi_1(\omega)) = \varphi(0) = 0.$$

Taking $\omega = 0$ shows that $\beta = f(0) \in L_2$, and we then have $\alpha L_1 \subseteq L_2$. For all $z \in \mathbb{C}$ we have $\varphi(\pi_1(z)) = \pi_2(f(z)) = \pi_2(\alpha z)$, thus $\varphi = \varphi_\alpha$. The value of α is unique: if $\varphi = \varphi_\gamma$ for some $\gamma \in \mathbb{C}$ then $(\alpha - \gamma)z \in L_2$ for all $z \in \mathbb{C}$, which implies $\alpha - \gamma = ((\alpha-\gamma)z)' = 0$ (as argued above), and therefore $\gamma = \alpha$. □

As noted above, a morphism $\varphi\colon \mathbb{C}/L_1 \to \mathbb{C}/L_2$ of complex tori is a holomorphic map that is also a group homomorphism; in particular, $\varphi(0) = 0$, so Theorem 16.1 applies and we have the following corollary.

Corollary 16.2. *For any two lattices $L_1, L_2 \subseteq \mathbb{C}$ the map*

$$\begin{aligned} \left\{\alpha \in \mathbb{C} : \alpha L_1 \subseteq L_2\right\} &\to \left\{\text{morphisms } \varphi\colon \mathbb{C}/L_1 \to \mathbb{C}/L_2\right\} \\ \alpha &\mapsto \varphi_\alpha \end{aligned}$$

is an isomorphism of additive groups. If $L_1 = L_2$ it is an isomorphism of commutative rings.

The set $\{\alpha \in \mathbb{C} : \alpha L_1 \subseteq L_2\}$ on the LHS contains 0 and is closed under addition and negation and is thus an additive subgroup of $\mathbb{C}$, and if $L_1 = L_2$ it is also closed under multiplication and forms a subring of $\mathbb{C}$. The set of morphisms on the RHS, which we could have written as $\mathrm{Hom}(\mathbb{C}/L_1, \mathbb{C}/L_2)$, is an additive group under pointwise addition, and when $L_1 = L_2$ it is the endomorphism ring $\mathrm{End}(\mathbb{C}/L_1)$ with multiplication given by composition.

Proof. Theorem 16.1 gives us a bijection of sets; we just need to check that it is a group/ring homomorphism. For $i = 1, 2$, let $\pi_i\colon \mathbb{C} \to \mathbb{C}/L_i$ be the projection maps as above. If $\alpha L_1 \subseteq L_2$ and $\beta L_1 \subseteq L_2$ then for all $z \in \mathbb{C}$ we have

$$\varphi_{\alpha+\beta}(\pi_1(z)) = \pi_2((\alpha+\beta)z) = \pi_2(\alpha z) + \pi_2(\beta z) = \varphi_\alpha(\pi_1(z)) + \varphi_\beta(\pi_1(z)) = (\varphi_\alpha + \varphi_\beta)(\pi_1(z)),$$

thus the map $\alpha \mapsto \varphi_\alpha$ defines a homomorphism of additive groups. If $L_1 = L_2$ and we put $\pi = \pi_1 = \pi_2$ then we also have

$$\varphi_{\alpha\beta}(\pi(z)) = \pi(\alpha\beta z) = \varphi_\alpha(\pi(\beta z)) = \varphi_\alpha(\varphi_\beta(\pi(z))) = (\varphi_\alpha\varphi_\beta)(\pi(z)),$$

which shows that $\alpha \mapsto \varphi_\alpha$ is a ring homomorphism. □

We will henceforth identify $\mathrm{Hom}(\mathbb{C}/L_1, \mathbb{C}/L_2)$ with $\{\alpha \in \mathbb{C} : \alpha L_1 \subseteq L_2\}$ and φ_α with α; we thus view any α for which $\alpha L_1 \subseteq L_2$ both as a complex number and a morphism $\mathbb{C}/L_1 \to \mathbb{C}/L_2$. We will also freely use $z \in \mathbb{C}$ to denote its image under the quotient map $\pi_1\colon \mathbb{C} \to \mathbb{C}/L_1$ and use αz to denote $\varphi_\alpha(\pi_1(z)) = \pi_2(\alpha z)$ whenever the context is clear.

16.2 Morphisms of complex tori and isogenies of elliptic curves over $\mathbb{C}$

Let $L_1, L_2 \subseteq \mathbb{C}$ be lattices. In order to complete the proof that complex tori and elliptic curves over $\mathbb{C}$ are equivalent categories, we need to give an explicit isomorphism $\mathrm{Hom}(\mathbb{C}/L_1, \mathbb{C}/L_2) \simeq \mathrm{Hom}(E_{L_1}, E_{L_2})$. To do this we need to first prove a lemma about fields of elliptic functions.

Recall that the set of all elliptic functions for a given lattice L forms a field $\mathbb{C}(L)$ that includes the constant functions $\mathbb{C} \subseteq \mathbb{C}(L)$. We now show that the extension $\mathbb{C}(L)/\mathbb{C}$ is generated by the Weierstrass $\wp$-function and its derivative, and the subfield $\mathbb{C}(L)^{\mathrm{even}}$ of even functions (the $f \in \mathbb{C}(L)$ for which $f(-z) = f(z)$) is generated by the $\wp$-function alone.

Lemma 16.3. *Let $L \subseteq \mathbb{C}$ be a lattice. The following hold:*

(i) $\mathbb{C}(L) = \mathbb{C}(\wp, \wp')$;

(ii) $\mathbb{C}(L)^{\mathrm{even}} = \mathbb{C}(\wp)$;

(iii) *if $f \in \mathbb{C}(L)^{\mathrm{even}}$ is holomorphic on $\mathbb{C} - L$ then $f \in \mathbb{C}[\wp]$.*

Proof. Every $f \in \mathbb{C}(L)$ can be written as the sum of an even function and an odd function:

$$f(z) = \frac{f(z) + f(-z)}{2} + \frac{f(z) - f(-z)}{2}.$$

Any odd function $g \in \mathbb{C}(L)$ can be written as

$$g(z) = \frac{g(z)}{\wp'(z)}\wp'(z),$$

where $g(z)/\wp'(z)$ is an even function; thus (i) follows from (ii).

We now show that (ii) follows from (iii). Let $f \in \mathbb{C}(L)^{\mathrm{even}}$ and let m be the number of poles of f that lie in $\mathcal{F}_0 - \{0\}$, where $\mathcal{F}_0$ is the standard fundamental parallelogram for L. The integer m is nonnegative and bounded by the order of f. If $m > 0$ then $f(z)$ has a pole at some nonzero $w \in \mathcal{F}_0$, say of order n. Now consider the even elliptic function

$$g(z) := (\wp(z) - \wp(w))^n,$$

which is holomorphic on $\mathbb{C} - L$ and has a zero of order at least n at w. The function $gf \in \mathbb{C}(L)^{\mathrm{even}}$ is holomorphic at w, and every pole of gf in $\mathbb{C} - L$ must be a pole of f, so it has strictly fewer than m poles in $\mathcal{F}_0 - \{0\}$. Repeating this process m times yields a polynomial $Q \in \mathbb{C}[x]$ such that $Q(\wp)f \in \mathbb{C}(L)^{\mathrm{even}}$ is holomorphic on $\mathbb{C} - L$; If we assume (iii), then $Q(\wp)f = P(\wp)$ for some $P \in \mathbb{C}[x]$ and $f = P(\wp)/Q(\wp) \in \mathbb{C}(\wp)$, implying (ii).

We now prove (iii). Let $f \in \mathbb{C}(L)^{\mathrm{even}}$ be nonzero and holomorphic on $\mathbb{C} - L$. If the order of f is zero then f is constant (by Liouville's theorem, since an elliptic function is necessarily bounded). Otherwise f must have a pole at 0 and its Laurent series expansion at 0 has the form

$$f(z) = \sum_{k=-n}^{\infty} a_{2k} z^{2k},$$

with $a_{-2n} \neq 0$, where $2n$ is order of f (which must be even). The function

$$f(z) - a_{-2n}\wp^n(z)$$

is an even elliptic function holomorphic on $\mathbb{C} - L$ of order at most $2(n-1)$. Repeating this at most n times yields a polynomial $P \in \mathbb{C}[x]$ such that $f - P(\wp) \in \mathbb{C}$, and (iii) follows. □

Theorem 16.4. *For $i = 1, 2$ let $L_i \subseteq \mathbb{C}$ be a lattice, let $E_i := E_{L_i}$ be the corresponding elliptic curve, define $\wp_i(z) := \wp(z; L_i)$, and let $\Phi_i \colon \mathbb{C}/L_i \to E_i(\mathbb{C})$ be the isomorphism that sends $z \notin L_i$ to $\big(\wp_i(z), \wp_i'(z)\big)$ and $z \in L_i$ to 0. For any $\alpha \in \mathbb{C}^\times$, the following are equivalent:*

(1) $\alpha L_1 \subseteq L_2$;

(2) $\wp_2(\alpha z) = u\big(\wp_1(z)\big)/v\big(\wp_1(z)\big)$ *for some polynomials* $u, v \in \mathbb{C}[x]$;

(3) *There is a unique* $\phi_\alpha \in \mathrm{Hom}(E_1, E_2)$ *such that the following diagram commutes:*

$$\begin{array}{ccccc} \mathbb{C} & \longrightarrow & \mathbb{C}/L_1 & \xrightarrow{\Phi_1} & E_1(\mathbb{C}) \\ \big\downarrow{\scriptstyle \alpha} & & & & \big\downarrow{\scriptstyle \phi_\alpha} \\ \mathbb{C} & \longrightarrow & \mathbb{C}/L_2 & \xrightarrow{\Phi_2} & E_2(\mathbb{C}) \end{array}$$

For every morphism $\phi \in \mathrm{Hom}(E_1, E_2)$ there is a unique $\alpha = \alpha_\phi$ satisfying (1)–(3). *The maps $\phi \mapsto \alpha_\phi$ and $\alpha \mapsto \phi_\alpha$ define inverse isomorphisms of* $\mathrm{Hom}(E_1, E_2)$ *and* $\{\alpha \in \mathbb{C} : \alpha L_1 \subseteq L_2\}$.

Proof. (1) $\Rightarrow$ (2): Let $\omega \in L_1$. We have $\wp_2(\alpha(z + \omega)) = \wp_2(\alpha z + \alpha\omega) = \wp_2(\alpha z)$. Thus $\wp_2(\alpha z)$ is periodic with respect to L_1, and it is meromorphic, so it is an elliptic function for L_1. It is an even function, so it is a rational function $u(\wp_1(z))/v(\wp_1(z))$ of $\wp_1(z)$, by Lemma 16.3.

(2) $\Rightarrow$ (3): Let $\wp_2(\alpha z) = u(\wp_1(z))/v(\wp_1(z))$, let $s := (u'v - v'u)$ and $t := \alpha v^2$, and define

$$\phi_\alpha := \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)} y\right).$$

Then

$$\wp_2'(\alpha z) = \frac{1}{\alpha}\big(\wp_2(\alpha z)\big)' = \frac{1}{\alpha}\left(\frac{u(\wp_1(z))}{v(\wp_1(z))}\right)' = \frac{s(\wp_1(z))}{t(\wp_1(z))}\wp_1'(z),$$

and we have

$$\phi_\alpha(\Phi_1(z)) = \phi_\alpha(\wp_1(z), \wp_1'(z)) = \left(\frac{u(\wp_1(z))}{v(\wp_1(z))}, \frac{s(\wp_1(z))}{t(\wp_1(z))}\wp_1'(z)\right) = (\wp_2(\alpha z), \wp_2'(\alpha z)) = \Phi_2(\alpha z),$$

so the diagram in (3) commutes. If $\phi \in \mathrm{Hom}(E_1, E_2)$ also satisfies $\phi(\Phi_1(z)) = \Phi_2(\alpha z)$ then

$$(\phi - \phi_\alpha)(\Phi_1(z)) = \phi(\Phi_1(z)) - \phi_\alpha(\Phi_1(z)) = \Phi_2(\alpha z) - \Phi_2(\alpha z) = 0,$$

and $\phi = \phi_\alpha$; thus ϕ_α is the only element of $\mathrm{Hom}(E_1, E_2)$ that makes the diagram commute.

(3) $\Rightarrow$ (1): For all $\omega \in L_1$ we have $\Phi_2(\alpha\omega) = \phi_\alpha(\Phi_1(\omega)) = \phi_\alpha(0) = 0$, which implies $\alpha\omega \in L_2$, thus $\alpha L_1 \subseteq L_2$.

For any $\phi \in \mathrm{Hom}(E_1, E_2)$, the map $\Phi_2^{-1} \circ \phi \circ \Phi_1$ is an element of $\mathrm{Hom}(\mathbb{C}/L_1, \mathbb{C}/L_2)$ and is therefore induced by the multiplication-by-α map $z \mapsto \alpha z$ for a unique $\alpha = \alpha_\phi$ satisfying $\alpha L_1 \subseteq L_2$, by Corollary 16.2. The maps $\alpha \mapsto \phi_\alpha$ and $\phi \mapsto \alpha_\phi$ are thus inverse bijections.

We now show that the map $\Psi \colon \mathrm{Hom}(E_1, E_2) \to \{\alpha \in \mathbb{C} : \alpha L_1 \subseteq L_2\}$ defined by $\phi \mapsto \alpha_\phi$ is a group homomorphism. We have $\Psi(0) = 0$, and for all $\phi_1, \phi_2 \in \mathrm{Hom}(E_1, E_2)$

$$\Psi(\phi_1 + \phi_2) \;=\; \Phi_2^{-1} \circ (\phi_1 + \phi_2) \circ \Phi_1 \;=\; \Phi_2^{-1} \circ \phi_1 \circ \Phi_1 + \Phi_2^{-1} \circ \phi_2 \circ \Phi_1 \;=\; \Psi(\phi_1) + \Psi(\phi_2).$$

Thus Ψ is a group homomorphism and therefore an isomorphism, since it is a bijection. □

16.3 Endomorphism rings of complex tori and elliptic curves over $\mathbb{C}$

We now specialize to the case $L = L_2 = L_1$, and put $E = E_L$, in which case the group $\{\alpha \in \mathbb{C} \colon \alpha L \subseteq L\} \simeq \mathrm{Hom}(E, E) = \mathrm{End}(E)$ is also a ring.

Corollary 16.5. *Let $L \subseteq \mathbb{C}$ be a lattice and let $E := E_L$. The maps $\alpha \mapsto \phi_\alpha$ and $\phi \mapsto \alpha_\phi$ are inverse ring isomorphisms between $\{\alpha \in \mathbb{C} : \alpha L \subseteq L\}$ and $\mathrm{End}(E)$, the involution $\phi \mapsto \hat{\phi}$ of $\mathrm{End}(E)$ corresponds to complex conjugation $\alpha \mapsto \bar{\alpha}$ in $\{\alpha \in \mathbb{C} : \alpha L \subseteq L\}$, and we have $\mathrm{T}(\alpha) := \alpha + \bar{\alpha} = \mathrm{tr}\,\phi_\alpha$ and $\mathrm{N}(\alpha) := \alpha\bar{\alpha} = \deg \phi_\alpha = \deg u = \deg v + 1$, where $u, v \in \mathbb{C}[x]$ are as in (2) of Theorem 16.4.*

Proof. Let $\Phi \colon \mathbb{C}/L \to E(\mathbb{C})$ and $\Psi \colon \mathrm{End}(E) \to \{\alpha \in \mathbb{C} : \alpha L \subseteq L\}$ be as in Theorem 16.4 and its proof (so $\Psi(\phi) = \alpha_\phi$); they are both group isomorphisms. For $\phi_1, \phi_2 \in \mathrm{End}(E)$ we have

$$\Psi(\phi_1\phi_2) \;=\; \Phi^{-1} \circ (\phi_1 \circ \phi_2) \circ \Phi \;=\; (\Phi^{-1} \circ \phi_1 \circ \Phi) \circ (\Phi^{-1} \circ \phi_2 \circ \Phi) \;=\; \Psi(\phi_1)\Psi(\phi_2),$$

thus Ψ is a ring homomorphism and therefore a ring isomorphism, since it is a bijection.

For any $\phi \in \mathrm{End}(E)$, the complex number $\alpha := \alpha_\phi$ satisfies the characteristic equation

$$x^2 - (\mathrm{tr}\,\phi)x + \deg \phi = 0,$$

which has integer coefficients and discriminant $\mathrm{tr}(\phi)^2 - 4\deg(\phi) \leq 0$. Thus $\alpha \in \mathbb{Z}$, or α is an algebraic integer in an imaginary quadratic field, and in either case its complex conjugate $\bar{\alpha}$ satisfies the same quadratic equation and we have $\bar{\alpha}\alpha = \deg \phi = \hat{\phi}\phi$, which implies $\bar{\alpha} = \hat{\phi}$ ($\{\alpha \in \mathbb{C} \colon \alpha L \subseteq L\} \simeq \mathrm{End}(E)$ has no zero divisors, so the cancellation law applies), and we have $\mathrm{T}(\alpha) = \alpha + \bar{\alpha} = \phi + \hat{\phi} = \mathrm{tr}\,\phi$ and $\mathrm{N}(\alpha) = \alpha\bar{\alpha} = \phi\hat{\phi} = \deg \phi$.

Finally, for any $\alpha \in \{\alpha \in \mathbb{C} : \alpha L \subseteq L\}$ we can apply (2) in Theorem 16.4 to write $v(\wp(z))\wp(\alpha z) = u(\wp(z))$ for some $u, v \in \mathbb{C}[x]$. The functions $u(\wp(z))$ and $v(\wp(z))$ have poles of order $2\deg u$ and $2\deg v$ at 0, respectively, while $\wp(\alpha z)$ has a pole of order 2 at 0, so we must have $\deg u = \deg v + 1$ and

$$\deg \phi = \max(\deg u, \deg v) = \deg u = \deg v + 1,$$

where $\phi = \phi_\alpha := \left(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y\right)$ is as in the proof of Theorem 16.4. □

Remark 16.6. Theorem 16.4 and Corollary 16.5 explain the origin of the term *complex multiplication* (CM). When $\mathrm{End}(E_L)$ is bigger than $\mathbb{Z}$ the extra endomorphisms in $\mathrm{End}(E_L)$ all correspond to multiplication-by-α maps in $\mathrm{End}(\mathbb{C}/L)$, for some $\alpha \in \mathbb{C} - \mathbb{R}$ that is an algebraic integer in an imaginary quadratic field.

Corollary 16.7. *Let E be an elliptic curve defined over $\mathbb{C}$. Then $\mathrm{End}(E)$ is commutative and therefore isomorphic to either $\mathbb{Z}$ or an order in an imaginary quadratic field.*

Proof. Let L be the lattice corresponding to E. The ring $\mathrm{End}(E) \simeq \{\alpha \in \mathbb{C} : \alpha L \subseteq L\}$ is clearly commutative, and therefore not an order in a quaternion algebra. The result then follows from our classification of endomorphism rings of elliptic curves in Lecture 12, see Theorem 12.17 and Corollary 12.20. □

Remark 16.8. Corollary 16.7 applies to elliptic curves over $\mathbb{Q}$, number fields, or any field that can be embedded in $\mathbb{C}$. It can be extended to all fields of characteristic 0 via the Lefschetz principle; see [1, Thm. VI.6.1].

16.4 Elliptic curves with a given endomorphism ring

We have shown that for any lattice $L \subseteq \mathbb{C}$ we have ring isomorphisms

$$\mathrm{End}(E_L) \simeq \{\alpha \in \mathbb{C} : \alpha L \subseteq L\} \simeq \mathrm{End}(\mathbb{C}/L). \tag{1}$$

As noted above, we have been treating the isomorphism on the left as an equality, and it will be convenient to do the same for the isomorphism on the right. The endomorphism algebra $\mathrm{End}^0(E_L)$ is isomorphic to either $\mathbb{Q}$ or an imaginary quadratic field, so we can always embed $\mathrm{End}^0(E_L)$ in $\mathbb{C}$. Once we have done this, provided that we regard $\mathrm{End}(E_L)$ as a subring of $\mathrm{End}^0(E_L)$ (via the canonical injection $\phi \mapsto \phi \otimes 1$), we actually have an equality $\mathrm{End}(E_L) = \{\alpha \in \mathbb{C} : \alpha L \subseteq L\}$; moreover, when $\mathrm{End}(\mathbb{C}/L)$ is an imaginary quadratic order $\mathcal{O}$, we can choose the embedding of $\mathrm{End}^0(E_L)$ into $\mathbb{C}$ so that each multiplication-by-α endomorphism of $\mathbb{C}/L$ is identified with $\phi_\alpha \in \mathrm{End}(E_L)$ (as opposed to $\hat{\phi}_\alpha$). This is known as the *normalized identification* of $\mathrm{End}(E_L)$ with $\mathrm{End}(\mathbb{C}/L) = \mathcal{O}$, which we henceforth assume.

We now want to focus on the CM case, where $\mathrm{End}(E_L)$ is an order $\mathcal{O}$ in an imaginary quadratic field K. The order $\mathcal{O}$ is a lattice, and we would like to understand how the lattices L and $\mathcal{O}$ are related. In particular, for which lattices L do we have $\mathrm{End}(E_L) = \mathcal{O}$?

An obvious candidate is $L = \mathcal{O}$. If $\alpha \in \mathrm{End}(E_\mathcal{O})$, then $\alpha\mathcal{O} \subseteq \mathcal{O}$ and therefore $\alpha \in \mathcal{O}$, since the ring $\mathcal{O}$ contains 1. Conversely, if $\alpha \in \mathcal{O}$, then $\alpha\mathcal{O} \subseteq \mathcal{O}$, since $\mathcal{O}$ is closed under multiplication, and therefore $\alpha \in \mathrm{End}(E_\mathcal{O})$; thus $\mathrm{End}(E_\mathcal{O}) = \mathcal{O}$.

The same holds for any lattice homothetic to $\mathcal{O}$. Indeed, the set $\{\alpha \in \mathbb{C} \colon \alpha L \subseteq L\}$ does not change if we replace L with $L' = \lambda L$ for any $\lambda \in \mathbb{C}^\times$, so we are really only interested in lattices up to homothety (and elliptic curves up to isomorphism). The question now before us is this: are there any lattices L not homothetic to $\mathcal{O}$ for which we have $\mathrm{End}(E_L) = \mathcal{O}$?

Given that we are only considering lattices up to homothety, we may assume without loss of generality that $L = [1, \tau]$, and we can always write $\mathcal{O} = [1, \omega]$ for some imaginary quadratic integer ω. If $\mathrm{End}(E_L) = \mathcal{O}$, then we must have $\omega \cdot 1 = \omega \in L$, so $\omega = m + n\tau$, for some $m, n \in \mathbb{Z}$ with $n \neq 0$. Thus $nL = [n, n\tau] = [n, \omega - m] \subseteq [1, \omega] = \mathcal{O}$, which means that L is homothetic to a sublattice of $\mathcal{O}$. This sublattice must be closed under multiplication by $\mathcal{O}$, which implies that L is homothetic to an $\mathcal{O}$-ideal (recall that an $\mathcal{O}$-ideal is an additive subgroup of $\mathcal{O}$ closed under multiplication by $\mathcal{O}$, equivalently, any $\mathcal{O}$-submodule of $\mathcal{O}$).

But the situation is a bit more complicated than it appears. While every lattice L for which $\mathrm{End}(E_L) = \mathcal{O}$ is an $\mathcal{O}$-ideal, the converse does not hold (unless $\mathcal{O}$ is the maximal order $\mathcal{O}_K$). If we start with an arbitrary $\mathcal{O}$-ideal L, it is clear that the set

$$\mathcal{O}(L) := \{\alpha \in \mathbb{C} : \alpha L \subseteq L\} = \{\alpha \in K : \alpha L \subseteq L\}$$

is an order in K: note that $\mathcal{O} \subseteq \mathcal{O}(L) = \mathrm{End}(E_L)$, since the $\mathcal{O}$-ideal L is closed under multiplication by $\mathcal{O}$, and this implies that $\mathrm{End}^0(E_L) = K$. But it is not necessarily true that $\mathcal{O}(L)$ is equal to $\mathcal{O}$; if $\mathcal{O} \neq \mathcal{O}_K$ we can always find an $\mathcal{O}$-ideal L for which $\mathcal{O}(L)$ strictly contains $\mathcal{O}$ (see Problem Set 9). This motivates the following definition.

Definition 16.9. Let $\mathcal{O}$ be an order in an imaginary quadratic field K, and let L be an $\mathcal{O}$-ideal. We say that L is a *proper* $\mathcal{O}$-ideal if $\mathcal{O}(L) = \mathcal{O}$.

Given that we are only interested in lattices up to homothety, we shall regard two $\mathcal{O}$-ideals as *equivalent* if they are homothetic as lattices. A homothety $L' = \lambda L$ between lattices that are $\mathcal{O}$-ideals can always be written with $\lambda = \alpha/\beta$ for some $\alpha, \beta \in \mathcal{O}$. To see this, note that if $L = [\omega_1, \omega_2]$ then we can take $\alpha = \lambda\omega_1 \in \mathcal{O}$ and $\beta = \omega_1$. Thus homothetic $\mathcal{O}$-ideals L

and L' always satisfy an equation $\alpha L = \beta L'$ for some $\alpha, \beta \in \mathcal{O}$. This motivates the following definition.

Definition 16.10. Let $\mathcal{O}$ be an order in an imaginary quadratic field K. Two $\mathcal{O}$-ideals $\mathfrak{a}$ and $\mathfrak{b}$ are said to be *equivalent* if they are homothetic as lattices; equivalently, $\alpha\mathfrak{a} = \beta\mathfrak{b}$ for some nonzero $\alpha, \beta \in \mathcal{O}$; we can also write this as $(\alpha)\mathfrak{a} = (\beta)\mathfrak{b}$, where (α) and (β) denote principal ideals and $(\alpha)\mathfrak{a}$ and $(\beta)\mathfrak{b}$ are products of ideals.

Recall that the product of two $\mathcal{O}$-ideals $\mathfrak{a}$ and $\mathfrak{b}$ is the ideal generated by all products ab with $a \in \mathfrak{a}$ and $b \in \mathfrak{b}$, and that ideal multiplication is commutative and associative. It is enough to consider products of generators, so if $\mathfrak{a} = [a_1, a_2]$ and $\mathfrak{b} = [b_1, b_2]$, then $\mathfrak{ab}$ is the ideal generated by the four elements $a_1b_1, a_1b_2, a_2b_1, a_2b_2$. Since $\mathfrak{ab}$ is an additive subgroup of $\mathcal{O}$, it is necessarily a free $\mathbb{Z}$-module of rank 2 and can be written as a lattice $[c_1, c_2]$, where c_1 and c_2 are $\mathcal{O}$-linear combinations of $a_1b_1, a_1b_2, a_2b_1, a_2b_2$. Note that ideal multiplication respects equivalence:

$$\alpha\mathfrak{a} = \beta\mathfrak{b} \text{ and } \gamma\mathfrak{c} = \delta\mathfrak{d} \quad \Longrightarrow \quad \alpha\gamma\mathfrak{ac} = \beta\delta\mathfrak{bd}.$$

Definition 16.11. Let $\mathcal{O}$ be an order in an imaginary quadratic field. The *ideal class group* $\mathrm{cl}(\mathcal{O})$ is the multiplicative group of equivalence classes of proper $\mathcal{O}$-ideals.

We should note that it is not clear *a priori* that $\mathrm{cl}(\mathcal{O})$ is actually a group; it is clearly closed under an associative multiplication and contains an identity element (the class of principal ideals), but it is not obvious that every element has an inverse. We will give an explicit proof of this in the next lecture (see Problem Set 9 for an alternative proof that also shows that $\mathrm{cl}(\mathcal{O})$ is finite). But even without knowing that $\mathrm{cl}(\mathcal{O})$ is actually a group, our discussion above makes the following theorem clear.

Theorem 16.12. *Let $\mathcal{O}$ be an order in an imaginary quadratic field. There is a one-to-one correspondence between elements of the ideal class group $\mathrm{cl}(\mathcal{O})$ and homothety classes of lattices $L \subseteq \mathbb{C}$ for which $\mathrm{End}(E_L) \simeq \mathcal{O}$.*

References

[1] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer 2009.

17 The CM torsor

Over the course of the last three lectures we have established an equivalence of categories between complex tori $\mathbb{C}/L$ and elliptic curves $E/\mathbb{C}$:

$$\begin{aligned}\{\text{lattices } L \subseteq \mathbb{C}\}/_\sim &\xrightarrow{\sim} \{\text{elliptic curves } E/\mathbb{C}\}/_\simeq \\ L &\longmapsto E_L \colon y^2 = 4x^3 - g_2(L)x - g_3(L) \\ j(L) &= j(E_L)\end{aligned}$$

in which homothetic lattices correspond to isomorphic elliptic curves, and we have established ring isomorphisms

$$\mathrm{End}(\mathbb{C}/L) \simeq \mathcal{O}(L) \simeq \mathrm{End}(E_L)$$

where the ring

$$\mathcal{O}(L) := \{\alpha \in \mathbb{C} : \alpha L \subseteq L\}$$

is necessarily equal to $\mathbb{Z}$ or an order $\mathcal{O}$ in an imaginary quadratic field. In the latter case, which we will assume throughout this lecture, the elliptic curve E_L is said to have *complex multiplication* (CM) by $\mathcal{O}$, and the lattice L is necessarily homothetic to an $\mathcal{O}$-ideal.

If we fix the order $\mathcal{O}$, the $\mathcal{O}$-ideals L for which $\mathrm{End}(E_L) \simeq \mathcal{O}$ are precisely those for which $\mathcal{O}(L) = \mathcal{O}$; in the previous lecture we defined such $\mathcal{O}$-ideals to be *proper*. Note that $\mathcal{O} \subseteq \mathcal{O}(L)$ always holds, since L is an $\mathcal{O}$-ideal, but in general $\mathcal{O}(L)$ may be larger than $\mathcal{O}$.

The sets

$$\{L \subseteq \mathbb{C} : \mathcal{O}(L) = \mathcal{O}\}/_\sim \longleftrightarrow \{E/\mathbb{C} : \mathrm{End}(E) = \mathcal{O}\}/_\simeq$$

are both in bijection with the *ideal class group*

$$\mathrm{cl}(\mathcal{O}) := \{\text{proper } \mathcal{O}\text{-ideals } \mathfrak{a}\}/_\sim$$

where the equivalence relation on proper $\mathcal{O}$-ideals is defined by

$$\mathfrak{a} \sim \mathfrak{b} \qquad \Longleftrightarrow \qquad \alpha\mathfrak{a} = \beta\mathfrak{b} \text{ for some nonzero } \alpha, \beta \in \mathcal{O},$$

and the group operation is given by multiplying representative ideals. As noted in the previous lecture it is not immediately obvious that $\mathrm{cl}(\mathcal{O})$ is a group (associativity is clear but the existence of inverses is not); one of our first goals is to prove this.

Remark 17.1. Recall that that an order in a $\mathbb{Q}$-algebra K of dimension r is a subring of K that is also a free $\mathbb{Z}$-module of rank r; see Definition 12.22. When K is an imaginary quadratic field embedded in the complex numbers, every order $\mathcal{O}$ in K is automatically a lattice in $\mathbb{C}$, since in this case $r = \dim K = 2$ and K is not contained in $\mathbb{R}$. Not every lattice in $\mathbb{C}$ is an imaginary quadratic order, but every imaginary quadratic order $\mathcal{O}$ is a lattice in $\mathbb{C}$ (once we fix an embedding of its fraction field), as is every $\mathcal{O}$-ideal (as a free $\mathbb{Z}$-module an $\mathcal{O}$-ideal must have the same rank as $\mathcal{O}$ because it is closed under multiplication by $\mathcal{O}$). Notice that the equivalence relation we have defined on $\mathcal{O}$-ideals coincides with our notion of homothety for lattices.

Recalling that isomorphism classes of elliptic curves over an algebraically closed field are identified by their j-invariants, we now define the set

$$\mathrm{Ell}_\mathcal{O}(\mathbb{C}) = \{j(E) : E \text{ is defined over } \mathbb{C} \text{ and } \mathrm{End}(E) = \mathcal{O}\},$$

and we then have a bijection of sets

$$\begin{aligned} \mathrm{cl}(\mathcal{O}) &\xrightarrow{\sim} \mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) \\ [\mathfrak{a}] &\longmapsto j(E_{\mathfrak{a}}) = j(\mathfrak{a}). \end{aligned}$$

As you will prove in Problem Set 9, the ideal class group $\mathrm{cl}(\mathcal{O})$ is finite, thus the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is finite. Its cardinality is the *class number* $h(\mathcal{O}) = \#\,\mathrm{cl}(\mathcal{O})$. Remarkably, not only are the sets $\mathrm{cl}(\mathcal{O})$ and $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ in bijection, the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ admits a group action by $\mathrm{cl}(\mathcal{O})$. In order to define this action, and to gain a better understanding of what it means for an $\mathcal{O}$-ideal to be proper, we first introduce the notion of a fractional $\mathcal{O}$-ideal.

17.1 Fractional ideals

Definition 17.2. Let $\mathcal{O}$ be an integral domain with fraction field K. For any $\lambda \in K^\times$ and $\mathcal{O}$-ideal $\mathfrak{a}$, the $\mathcal{O}$-module $\mathfrak{b} = \lambda\mathfrak{a} := \{\lambda\alpha : \alpha \in \mathfrak{a}\}$ is called a *fractional $\mathcal{O}$-ideal*.[1] Multiplication of fractional ideals $\mathfrak{b} = \lambda\mathfrak{a}$ and $\mathfrak{b}' = \lambda\mathfrak{a}'$ is defined in the obvious way:

$$\mathfrak{b}\mathfrak{b}' := (\lambda\lambda')\mathfrak{a}\mathfrak{a}',$$

where $\mathfrak{a}\mathfrak{a}'$ is the product of the $\mathcal{O}$-ideals $\mathfrak{a}$ and $\mathfrak{a}'$.[2]

Without loss of generality we can assume $\lambda = 1/\beta$ for some $\beta \in \mathcal{O}$ (if $\lambda = \alpha/\beta$, replace $\mathfrak{a}$ with $\alpha\mathfrak{a}$), and in the case of interest to us, where K is a number field, we can assume $\lambda = 1/b$ for some positive integer b (if $f \in \mathbb{Z}[x]$ is the minimal polynomial of β then $f(\beta) - f(0)$ is divisible by β with $(f(\beta) - f(0))/\beta = -f(0)/\beta \in \mathcal{O}$, and we can take $b = \pm f(0) > 0$).

Fractional $\mathcal{O}$-ideals that lie in $\mathcal{O}$ are $\mathcal{O}$-ideals, and every $\mathcal{O}$-ideal is a fractional $\mathcal{O}$-ideal. Note that $\mathcal{O}$ is itself an $\mathcal{O}$-ideal, hence a fractional $\mathcal{O}$-ideal, and it acts as the multiplicative identity with respect to multiplication of fractional $\mathcal{O}$-ideals. Fractional $\mathcal{O}$-ideals $\mathfrak{b}$ for which there exists a fractional $\mathcal{O}$-ideal $\mathfrak{b}^{-1}$ such that $\mathfrak{b}\mathfrak{b}^{-1} = \mathcal{O}$ are said to be *invertible*. Not every fractional $\mathcal{O}$-ideal is invertible (the zero ideal never is, and in general there may be nonzero fractional $\mathcal{O}$-ideals that are not invertible). The set of invertible fractional $\mathcal{O}$-ideals forms a group under multiplication (this is sometimes called the *ideal group of* $\mathcal{O}$, even though its elements are fractional $\mathcal{O}$-ideals, many of which are not $\mathcal{O}$-ideals).

17.2 Norms

Let $\mathcal{O}$ be an order in an imaginary quadratic field K. We want to define the norm of a fractional $\mathcal{O}$-ideal $\mathfrak{b} = \lambda\mathfrak{a}$, a rational number that is the product of the norms of λ and $\mathfrak{a}$. We first define the norm of a field element $\lambda \in K^\times$, and the norm of an $\mathcal{O}$-ideal $\mathfrak{a}$.

Definition 17.3. Let K/k be a finite extension of fields and let $\lambda \in K^\times$. The multiplication-by-λ map $K \to K$ is an invertible linear transformation $M_\lambda \in \mathrm{GL}(K)$ of K as a k-vector space. The (field) *norm* and *trace* of λ are defined by

$$\mathrm{N}_{K/k}\lambda := \det M_\lambda \in k^\times \qquad \text{and} \qquad \mathrm{T}_{K/k}\lambda := \mathrm{tr}\, M_\lambda \in k.$$

[1] Some authors define fractional $\mathcal{O}$-ideals to be finitely generated $\mathcal{O}$-submodules of K. Every finitely generated $\mathcal{O}$-module in K is a fractional ideal under our definition, and when $\mathcal{O}$ is noetherian (which applies to orders in number fields), the definitions are equivalent.

[2] One can also add fractional $\mathcal{O}$-ideals via $\mathfrak{b} + \mathfrak{b}' := \{b + b' : b \in \mathfrak{b}, b' \in \mathfrak{b}\}$, but we won't need this.

One typically computes the norm and trace by fixing a basis for K as a k vector space and writing M_λ as a matrix using this basis, but the norm and trace of M_λ do not depend on the choice of basis. When K is a number field and $k = \mathbb{Q}$ it is common to simply write $\mathrm{N} := \mathrm{N}_{K/\mathbb{Q}}$ and $\mathrm{T} := \mathrm{T}_{K/\mathbb{Q}}$ when the number field K is clear from context, but note that for $\lambda \in \mathbb{Q}$ we have $\mathrm{N}\lambda = \lambda^{[K:\mathbb{Q}]}$ and $\mathrm{T}\lambda = [K:\mathbb{Q}]\lambda$, which depend on K, not just λ.

When $K \simeq \mathrm{End}^0(E)$ is an imaginary quadratic field, Definition 17.3 coincides with our definition of the (reduced) norm and trace of an element of $\mathrm{End}^0(E)$ (see Definition 12.6). When K is an imaginary quadratic field embedded in $\mathbb{C}$ we have $\mathrm{N}\alpha = \alpha\bar{\alpha}$ and $\mathrm{T}\alpha = \alpha + \bar{\alpha}$, where $\bar{\alpha}$ denotes complex conjugation (equivalently, the action of the unique non-trivial element of $\mathrm{Gal}(K/\mathbb{Q})$). Thus in this setting the complex conjugate

$$\bar{\alpha} = \mathrm{T}\alpha - \alpha = \hat{\alpha}$$

is the dual of $\alpha \in \mathrm{End}^0(E) = K \hookrightarrow \mathbb{C}$.

Definition 17.4. Let $\mathcal{O}$ be an order in a number field K and let $\mathfrak{a}$ be a nonzero $\mathcal{O}$-ideal. The (absolute) *norm* of the ideal $\mathfrak{a}$ is

$$\mathrm{N}\mathfrak{a} := [\mathcal{O} : \mathfrak{a}] = \#(\mathcal{O}/\mathfrak{a}) \in \mathbb{Z}_{>0}.$$

We can also interpret $\mathrm{N}\mathfrak{a}$ as the ratio of the volumes of fundamental parallelepipeds for $\mathfrak{a}$ and $\mathcal{O}$, viewed as lattices in the $\mathbb{Q}$-vector space K.

We now show that our two definitions of norm agree on principal $\mathcal{O}$-ideals.

Lemma 17.5. *Let α be a nonzero element of an order $\mathcal{O}$ in a number field K. Then*

$$\mathrm{N}(\alpha) = |\mathrm{N}\alpha|,$$

where (α) denotes the principal $\mathcal{O}$-ideal generated by α.

Proof. The lemma follows from the fact that the determinant of $M_\alpha \in \mathrm{GL}(K) \simeq \mathrm{GL}_n(\mathbb{Q})$ can be interpreted as the signed volume of the fundamental parallelepiped of the lattice (α) in the $\mathbb{Q}$-vector space $K \simeq \mathbb{Q}^n$, where $n = [K:\mathbb{Q}]$ is the degree of K. Notice that $\mathrm{N}(\alpha) = [\mathcal{O} : (\alpha)] = [\mathcal{O} : \alpha\mathcal{O}] = [\mathcal{O}_K : \alpha\mathcal{O}_K]$ depends only on α and K, not the order $\mathcal{O}$ (N.B. this holds for principal ideals but not in general). □

Warning 17.6. Given that the field norm is multiplicative and that we can view the ideal norm as the absolute value of a determinant, it would be reasonable to expect the ideal norm to be multiplicative. **This is not always true**. As an example, consider the ideal $\mathfrak{a} = [2, 2i]$ in the order $\mathcal{O} = [1, 2i]$, which has norm $\mathrm{N}\mathfrak{a} = [\mathcal{O} : \mathfrak{a}] = 2$. Then $\mathfrak{a}^2 = [4, 4i]$ and

$$\mathrm{N}\mathfrak{a}^2 = 8 \neq 2^2 = (\mathrm{N}\mathfrak{a})^2.$$

However, as we shall see, the ideal norm is multiplicative when $\mathfrak{a}$ and $\mathfrak{b}$ are both proper $\mathcal{O}$-ideals, and when either $\mathfrak{a}$ or $\mathfrak{b}$ is a principal ideal.

Corollary 17.7. *Let $\mathcal{O}$ be an order in a number field, let $\alpha \in \mathcal{O}$ be nonzero, and let $\mathfrak{a}$ be a nonzero $\mathcal{O}$-ideal. Then*

$$\mathrm{N}(\alpha\mathfrak{a}) = \mathrm{N}(\alpha)\mathrm{N}\mathfrak{a}.$$

Proof. $\mathrm{N}(\alpha\mathfrak{a}) = [\mathcal{O} : \alpha\mathfrak{a}] = [\mathcal{O} : \mathfrak{a}][\mathfrak{a} : \alpha\mathfrak{a}] = [\mathcal{O} : \mathfrak{a}][\mathcal{O} : \alpha\mathcal{O}] = \mathrm{N}\mathfrak{a}\mathrm{N}(\alpha) = \mathrm{N}(\alpha)\mathrm{N}\mathfrak{a}$. □

This allows us to make the following definition.

Definition 17.8. Let $\mathfrak{b} = \frac{1}{b}\mathfrak{a}$ be a nonzero fractional ideal in an order $\mathcal{O}$ of a number field, with $b \in \mathbb{Z}_{>0}$ (as above, we can always write $\mathfrak{b}$ this way). The (absolute) *norm* of $\mathfrak{b}$ is

$$\mathrm{N}\mathfrak{b} := \frac{\mathrm{N}\mathfrak{a}}{\mathrm{N}(b)} \in \mathbb{Q}^\times_{>0}.$$

Corollary 17.7 ensures that this does not depend on the choice of b and $\mathfrak{a}$.

When $\mathfrak{b} \subseteq \mathcal{O}$ we can take $b = 1$, in which case this agrees with Definition 17.4.

17.3 Proper and invertible fractional ideals

We now return to our original setting, where $\mathcal{O}$ is an order in an imaginary quadratic field. Extending our terminology for $\mathcal{O}$-ideals, for any fractional $\mathcal{O}$-ideal $\mathfrak{b}$ we define

$$\mathcal{O}(\mathfrak{b}) := \{\alpha : \alpha\mathfrak{b} \subseteq \mathfrak{b}\},$$

and say that $\mathfrak{b}$ is *proper* if $\mathcal{O}(\mathfrak{b}) = \mathcal{O}$. In this section we will show that $\mathfrak{b}$ is proper if and only if $\mathfrak{b}$ is invertible (there is a fractional $\mathcal{O}$-ideal $\mathfrak{b}^{-1}$ for which $\mathfrak{b}\mathfrak{b}^{-1} = \mathcal{O}$). Let us first note that for $\mathfrak{b} = \lambda\mathfrak{a}$, whether $\mathfrak{b}$ is proper or invertible depends only on the $\mathcal{O}$-ideal $\mathfrak{a}$.

Lemma 17.9. *Let $\mathcal{O}$ be an order in an imaginary quadratic field, let $\mathfrak{a}$ be a nonzero $\mathcal{O}$-ideal, and let $\mathfrak{b} = \lambda\mathfrak{a}$ be a fractional $\mathcal{O}$-ideal. Then $\mathfrak{a}$ is proper if and only if $\mathfrak{b}$ is proper, and $\mathfrak{a}$ is invertible if and only if $\mathfrak{b}$ is invertible.*

Proof. For the first statement, note that $\{\alpha : \alpha\mathfrak{b} \subseteq \mathfrak{b}\} = \{\alpha : \alpha\lambda\mathfrak{a} \subseteq \lambda\mathfrak{a}\} = \{\alpha : \alpha\mathfrak{a} \subseteq \mathfrak{a}\}$. For the second, if $\mathfrak{a}$ is invertible then $\mathfrak{b}^{-1} = \lambda^{-1}\mathfrak{a}^{-1}$, and if $\mathfrak{b}$ is invertible then $\mathfrak{a}^{-1} = \lambda\mathfrak{b}^{-1}$, since $\mathfrak{a}\mathfrak{a}^{-1} = \mathfrak{a}\lambda\mathfrak{b}^{-1} = \mathfrak{b}\mathfrak{b}^{-1} = \mathcal{O}$. □

We now prove that the invertible $\mathcal{O}$-ideals are precisely the proper $\mathcal{O}$-ideals and give an explicit formula for the inverse when it exists. Our proof follows the presentation in [1, §7].

Theorem 17.10. *Let $\mathcal{O}$ be an order in an imaginary quadratic field and let $\mathfrak{a} = [\alpha, \beta]$ be an $\mathcal{O}$-ideal. Then $\mathfrak{a}$ is proper if and only if $\mathfrak{a}$ is invertible. Whenever $\mathfrak{a}$ is invertible we have $\mathfrak{a}\bar{\mathfrak{a}} = (\mathrm{N}\mathfrak{a})$, where $\bar{\mathfrak{a}} = [\bar{\alpha}, \bar{\beta}]$ and $(\mathrm{N}\mathfrak{a})$ is the principal $\mathcal{O}$-ideal generated by the integer $\mathrm{N}\mathfrak{a}$; the inverse of $\mathfrak{a}$ is then the fractional $\mathcal{O}$-ideal $\mathfrak{a}^{-1} = \frac{1}{\mathrm{N}\mathfrak{a}}\bar{\mathfrak{a}}$.*

Proof. If $\mathfrak{a}$ is invertible, then for any $\gamma \in \mathbb{C}$ we have

$$\gamma\mathfrak{a} \subseteq \mathfrak{a} \implies \gamma\mathfrak{a}\mathfrak{a}^{-1} \subseteq \mathfrak{a}\mathfrak{a}^{-1} \implies \gamma\mathcal{O} \subseteq \mathcal{O} \implies \gamma \in \mathcal{O},$$

so $\mathcal{O}(\mathfrak{a}) \subseteq \mathcal{O}$, and $\mathfrak{a}$ is a proper $\mathcal{O}$-ideal, since we always have $\mathcal{O} \subseteq \mathcal{O}(\mathfrak{a})$.

We now assume that $\mathfrak{a} = [\alpha, \beta]$ is a proper $\mathcal{O}$-ideal and show that $\mathfrak{a}\bar{\mathfrak{a}} = (\mathrm{N}\mathfrak{a})$, which implies $\mathfrak{a}^{-1} = \frac{1}{\mathrm{N}\mathfrak{a}}\bar{\mathfrak{a}}$. Let $\tau = \beta/\alpha$, so that $\mathfrak{a} = \alpha[1, \tau]$, and let $ax^2 + bx + c \in \mathbb{Z}[x]$ be the minimal polynomial of τ made integral by clearing denominators, with $a > 0$ minimal. The fractional ideal $[1, \tau]$ is homothetic to $\mathfrak{a}$, so $\mathcal{O}([1, \tau]) = \mathcal{O}(\mathfrak{a}) = \mathcal{O}$, since $\mathfrak{a}$ is proper.

Let $\mathcal{O} = [1, \omega]$. Then $\omega \in [1, \tau]$ and $\omega = m + n\tau$ for some $m, n \in \mathbb{Z}$; after replacing ω with $\omega - m$, we may assume $\omega = n\tau$. We also have $\omega\tau \in [1, \tau]$, since $[1, \tau]$ is an $\mathcal{O}$-module, so $n\tau^2 \in [1, \tau]$, which implies that $a|n$, by the minimality of a (Gauss's lemma implies that we must have $\{f \in \mathbb{Z}[x] : f(\tau) = 0\} = (ax^2 + bx + c)$). We also have $a\tau[1, \tau] \subseteq [1, \tau]$ (since

$a\tau$ and $a\tau^2 = -b\tau - c$ lie in $[1,\tau]$), so $a\tau \in \mathcal{O}([1,\tau]) = \mathcal{O} = [1, n\tau]$, and we must have $n = a$ and $\mathcal{O} = [1, a\tau]$. Thus

$$\mathrm{N}(\mathfrak{a}) = [\mathcal{O} : \mathfrak{a}] = \big[[1,a\tau] : \alpha[1,\tau]\big] = \frac{1}{a}\big[[1,a\tau] : \alpha[1,a\tau]\big] = \frac{1}{a}[\mathcal{O} : \alpha\mathcal{O}] = \frac{\mathrm{N}(\alpha)}{a}.$$

We also have

$$\mathfrak{a}\bar{\mathfrak{a}} = \alpha[1,\tau]\bar{\alpha}[1,\bar{\tau}] = \mathrm{N}(\alpha)[1,\tau,\bar{\tau},\tau\bar{\tau}].$$

Using $a\tau^2 + b\tau + c = 0$, we see that $\tau + \bar{\tau} = -b/a$, and $\tau\bar{\tau} = c/a$. We then have

$$\mathfrak{a}\bar{\mathfrak{a}} = \mathrm{N}(\alpha)[1,\tau,\bar{\tau},\tau\bar{\tau}] = \frac{\mathrm{N}(\alpha)}{a}[a, a\tau, -b, c] = \mathrm{N}\mathfrak{a}[1, a\tau] = (\mathrm{N}\mathfrak{a})\mathcal{O} = (\mathrm{N}\mathfrak{a})$$

as claimed, where we have used $\gcd(a,b,c) = 1$ to get $[a, a\tau, -b, c] = [1, a\tau]$, and it follows that $\mathfrak{a}^{-1} = \frac{1}{\mathrm{N}\mathfrak{a}}\bar{\mathfrak{a}}$. □

Corollary 17.11. *The ideal class group* $\mathrm{cl}(\mathcal{O})$ *is the group of invertible fractional* $\mathcal{O}$*-ideals modulo its subgroup of principal fractional* $\mathcal{O}$*-ideals (in particular* $\mathrm{cl}(\mathcal{O})$ *is a group).*

Proof. Recall that $\mathrm{cl}(\mathcal{O}) = \{\text{proper } \mathcal{O}\text{-ideals}\}/\!\sim$, where $\sim$ denotes homothety. Let G be the group of invertible fractional $\mathcal{O}$-ideals and H its subgroup of principal fractional $\mathcal{O}$-ideals.

Every invertible fractional $\mathcal{O}$-ideal $\mathfrak{b} = \frac{1}{b}\mathfrak{a}$ is the product of an invertible principal fractional $\mathcal{O}$-ideal $(\frac{1}{b})$ and an invertible $\mathcal{O}$-ideal $\mathfrak{a}$, by Lemma 17.9. It follows that G/H consists of all cosets $\mathfrak{a}H$, where $\mathfrak{a}$ is any invertible, equivalently, proper $\mathcal{O}$-ideal (by Theorem 17.10). Every nonzero principal fractional $\mathcal{O}$-ideal is invertible, since $(\alpha)^{-1} = (\alpha^{-1})$, so H contains every nonzero principal fractional $\mathcal{O}$-ideal and for any two proper/invertible $\mathcal{O}$-ideals $\mathfrak{a}, \mathfrak{b}$ we have $\mathfrak{a} \sim \mathfrak{b}$ if and only if $\mathfrak{a}H = \mathfrak{b}H$. It follows that $\mathrm{cl}(\mathcal{O}) = G/H$. □

Corollary 17.12. *Let* $\mathcal{O}$ *be an order in an imaginary quadratic field and let* $\mathfrak{a}$ *and* $\mathfrak{b}$ *be invertible (equivalently, proper) fractional* $\mathcal{O}$*-ideals. Then* $\mathrm{N}(\mathfrak{a}\mathfrak{b}) = \mathrm{N}\mathfrak{a}\mathrm{N}\mathfrak{b}$.

Proof. If $\mathfrak{a} = \frac{1}{a}\mathfrak{a}'$ and $\mathfrak{b} = \frac{1}{b}\mathfrak{b}'$ with $a, b \in \mathbb{Z}_{>0}$ and $\mathfrak{a}', \mathfrak{b}' \subseteq \mathcal{O}$ then $\mathrm{N}(\mathfrak{a}\mathfrak{b}) = \frac{\mathrm{N}(\mathfrak{a}'\mathfrak{b}')}{\mathrm{N}a\mathrm{N}b}$, so it is enough to consider the case where $\mathfrak{a}$ and $\mathfrak{b}$ are invertible $\mathcal{O}$-ideals. We have

$$(\mathrm{N}(\mathfrak{a}\mathfrak{b})) = \mathfrak{a}\mathfrak{b}\overline{\mathfrak{a}\mathfrak{b}} = \mathfrak{a}\mathfrak{b}\bar{\mathfrak{a}}\bar{\mathfrak{b}} = \mathfrak{a}\bar{\mathfrak{a}}\mathfrak{b}\bar{\mathfrak{b}} = (\mathrm{N}\mathfrak{a})(\mathrm{N}\mathfrak{b}),$$

and it follows that $\mathrm{N}(\mathfrak{a}\mathfrak{b}) = \mathrm{N}\mathfrak{a}\mathrm{N}\mathfrak{b}$, since $\mathrm{N}\mathfrak{a}, \mathrm{N}\mathfrak{b}, \mathrm{N}(\mathfrak{a}\mathfrak{b}) \in \mathbb{Z}_{>0}$. □

17.4 The action of the ideal class group on CM elliptic curves

Let $\mathcal{O}$ be an order in an imaginary quadratic field. We are ready to define the action of $\mathrm{cl}(\mathcal{O})$ on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) = \{j(E) : E/\mathbb{C} \text{ with } \mathrm{End}(E) = \mathcal{O}\}$, which we will do by defining an action of proper $\mathcal{O}$-ideals on elliptic curves $E/\mathbb{C}$ with CM by $\mathcal{O}$ (up to isomorphism).

Every $E/\mathbb{C}$ with $\mathrm{End}(E) = \mathcal{O}$ is isomorphic to $E_{\mathfrak{b}}$, for some proper $\mathcal{O}$-ideal $\mathfrak{b}$. For any proper $\mathcal{O}$-ideal $\mathfrak{a}$ we define the action of $\mathfrak{a}$ on $E_{\mathfrak{b}}$ via

$$\mathfrak{a}E_{\mathfrak{b}} \coloneqq E_{\mathfrak{a}^{-1}\mathfrak{b}} \tag{1}$$

(we use $E_{\mathfrak{a}^{-1}\mathfrak{b}}$ rather than $E_{\mathfrak{a}\mathfrak{b}}$ because $\mathfrak{a}\mathfrak{b} \subseteq \mathfrak{b}$ but $\mathfrak{b} \subseteq \mathfrak{a}^{-1}\mathfrak{b}$). The action of the equivalence class $[\mathfrak{a}]$ on the isomorphism class $j(E_{\mathfrak{b}})$ is then defined by

$$[\mathfrak{a}]j(E_{\mathfrak{b}}) \coloneqq j(E_{\mathfrak{a}^{-1}\mathfrak{b}}), \tag{2}$$

which we can also write as

$$[\mathfrak{a}]j(\mathfrak{b}) \coloneqq j(\mathfrak{a}^{-1}\mathfrak{b}),$$

which does not depend on the choice of $\mathfrak{a}$ and $\mathfrak{b}$.

If $\mathfrak{a}$ is a nonzero principal $\mathcal{O}$-ideal, then the lattices $\mathfrak{b}$ and $\mathfrak{a}^{-1}\mathfrak{b}$ are homothetic, and we have $\mathfrak{a}E_{\mathfrak{b}} \simeq E_{\mathfrak{b}}$. Thus the identity element of $\mathrm{cl}(\mathcal{O})$ acts trivially on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$. For any proper $\mathcal{O}$-ideals $\mathfrak{a}$,$\mathfrak{b}$, and $\mathfrak{c}$ we have

$$\mathfrak{a}(\mathfrak{b}E_{\mathfrak{c}}) = \mathfrak{a}E_{\mathfrak{b}^{-1}\mathfrak{c}} = E_{\mathfrak{a}^{-1}\mathfrak{b}^{-1}\mathfrak{c}} = E_{(\mathfrak{b}\mathfrak{a})^{-1}\mathfrak{c}} = (\mathfrak{b}\mathfrak{a})E_{\mathfrak{c}} = (\mathfrak{a}\mathfrak{b})E_{\mathfrak{c}}.$$

Thus we have a group action of $\mathrm{cl}(\mathcal{O})$ on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$.

For any proper $\mathcal{O}$-ideals $\mathfrak{a}$ and $\mathfrak{b}$, we have $[\mathfrak{a}]j(\mathfrak{b}) = j(\mathfrak{a}^{-1}\mathfrak{b}) = j(\mathfrak{b})$ if and only if $\mathfrak{b}$ is homothetic to $\mathfrak{a}^{-1}\mathfrak{b}$, by Theorem 15.5, and in this case we have $\mathfrak{a}\mathfrak{b} = \lambda\mathfrak{b}$ for $\lambda \in K^\times$, and then $\mathfrak{a} = \lambda\mathcal{O}$ is principal. This implies that the action of $\mathrm{cl}(\mathcal{O})$ is not only faithful (only the identity fixes every element), it is *free* (every stabilizer is trivial).

The fact that the sets $\mathrm{cl}(\mathcal{O})$ and $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ have the same cardinality implies that the action must also be transitive: if we fix any $j_0 \in \mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ the images $[\mathfrak{a}]j_0$ of j_0 under the action of each $[\mathfrak{a}] \in \mathrm{cl}(\mathcal{O})$ must all be distinct, otherwise the action would not be free; there are only $\#\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) = \#\,\mathrm{cl}(\mathcal{O})$ possibilities, so the $\mathrm{cl}(\mathcal{O})$-orbit of j_0 is all of $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$.

A group action that is both free and transitive is said to be *regular*. Equivalently, the action of a group G on a set X is regular if and only if for all $x, y \in X$ there is a unique $g \in G$ for which $gx = y$. In this situation the set X is said to be a *G-torsor* (or *principal homogeneous space*) for G. We have thus shown that the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is a $\mathrm{cl}(\mathcal{O})$-torsor.

If we fix a particular element x of a G-torsor X, we can then view X as a group that is isomorphic to G under the map that sends $y \in X$ to the unique element $g \in G$ for which $gx = y$. Note that this involves an arbitrary choice of the identity element x; rather than thinking of elements of X as group elements, it is more appropriate to think of the "differences" or "ratios" of elements of X as group elements. In the case of the $\mathrm{cl}(\mathcal{O})$-torsor $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ there is an obvious choice for the identity element: the isomorphism class $j(E_{\mathcal{O}})$. But when we reduce to a finite field $\mathbb{F}_q$ and work with the $\mathrm{cl}(\mathcal{O})$-torsor $\mathrm{Ell}_{\mathcal{O}}(\mathbb{F}_q)$, as we shall soon do, we cannot readily distinguish the element of $\mathrm{Ell}_{\mathcal{O}}(\mathbb{F}_q)$ that corresponds to $j(E_{\mathcal{O}})$, and must make an arbitrary choice.

17.5 The CM action via isogenies

To better understand the $\mathrm{cl}(\mathcal{O})$-action on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ we now want to look at isogenies between elliptic curves with CM by $\mathcal{O}$; but first let us consider the situation more generally.

Let $\phi\colon E_1 \to E_2$ be an isogeny of elliptic curves over $\mathbb{C}$, and let L_1 and L_2 be corresponding lattices, so that $E_1 = E_{L_1}$ and $E_2 = E_{L_2}$. By Theorem 16.4, there is a unique $\alpha = \alpha_\phi$ with $\alpha L_1 \subseteq L_2$ such that the following diagram commutes

$$\begin{array}{ccc} \mathbb{C}/L_1 & \xrightarrow{\ \alpha\ } & \mathbb{C}/L_2 \\ \downarrow{\scriptstyle \Phi_1} & & \downarrow{\scriptstyle \Phi_2} \\ E_1(\mathbb{C}) & \xrightarrow{\ \phi\ } & E_2(\mathbb{C}). \end{array}$$

As we are only interested in lattices up to homothety and elliptic curves up to isomorphism, we can replace L_1 with the homothetic lattice αL_1 and E_1 by an isomorphic elliptic curve

so that $\alpha = 1$ and the isogeny ϕ is induced by the inclusion $L_1 \subseteq L_2$; note that this amounts to composing ϕ with an isomorphism and does not change its degree. Up to an isomorphism of elliptic curves and a homothety of lattices, every isogeny $\phi\colon E_1 \to E_2$ arises from an inclusion of lattices $L_1 \subseteq L_2$. In this situation it is clear what the kernel of ϕ is. By commutativity, since $\alpha = 1$, the kernel of ϕ consists of the images $\Phi_1(z)$ of points $z \in \mathbb{C}$ for which $\Phi_2(z) = 0$; these are precisely the $z \in L_2$ (which includes $L_1 \subseteq L_2$ but may also include $z \in L_2 - L_1$, since L_2 is a finer lattice). We have $\Phi_1(z) = 0$ if and only if $z \in L_1$, and it follows that

$$\#\ker\phi = [L_2 : L_1].$$

We are in characteristic zero, so ϕ is automatically separable and $\deg\phi = \#\ker\phi = [L_2 : L_1]$.

The discussion above applies to any isogeny of elliptic curves over $\mathbb{C}$; up to isomorphism they all arise from lattice inclusions; in particular, the inclusion $nL \subseteq L$ induces the multiplication-by-n endomorphism of E_L.

Let us now specialize to the case where $E_1/\mathbb{C}$ has CM by $\mathcal{O}$. Then L_1 is homothetic to a proper (hence invertible) $\mathcal{O}$-ideal $\mathfrak{b}$, so let us put $L_1 = \mathfrak{b}$ and $E_1 = E_{\mathfrak{b}}$. If $\mathfrak{a}$ is any invertible $\mathcal{O}$-ideal, the inclusion of lattices $\mathfrak{b} \subseteq \mathfrak{a}^{-1}\mathfrak{b}$ (given by $\mathfrak{a}\mathfrak{b} \subseteq \mathfrak{b}$) induces an isogeny

$$\phi_{\mathfrak{a}}\colon E_{\mathfrak{b}} \to E_{\mathfrak{a}^{-1}\mathfrak{b}} = \mathfrak{a}E_{\mathfrak{b}}$$

that corresponds to the action of $\mathfrak{a}$ on $E_{\mathfrak{b}}$ defined in (1). Moreover, if $E_2 = E_{L_2}$ has CM by $\mathcal{O}$, then L_2 is homothetic to an invertible $\mathcal{O}$-ideal $\mathfrak{c}$, and if we replace $\mathfrak{b}$ by the homothetic $\mathcal{O}$-ideal $(\mathrm{N}\mathfrak{c})\mathfrak{b}$, then $\mathfrak{c}$ divides (hence contains) $\mathfrak{b}$, because $\mathrm{N}\mathfrak{c} = \mathfrak{c}\bar{\mathfrak{c}}$, by Theorem 17.10. If we now put $\mathfrak{a} = \mathfrak{b}\mathfrak{c}^{-1}$, then the isogeny $\phi_{\mathfrak{a}}\colon E_{\mathfrak{b}} \to E_{\mathfrak{c}} = \mathfrak{a}E_{\mathfrak{b}}$ induced by the inclusion $\mathfrak{b} \subseteq \mathfrak{c}$ corresponds to the action of $\mathfrak{a}$ on $E_{\mathfrak{b}}$. After rescaling $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$ by integer multiples if necessary, we can assume $\mathfrak{a}$ is an invertible $\mathcal{O}$-ideal.

Thus all elliptic curves over $\mathbb{C}$ with CM by $\mathcal{O}$ are isogenous, and up to isomorphism, every isogeny between elliptic curves over $\mathbb{C}$ with CM by $\mathcal{O}$ is of the form $E_{\mathfrak{b}} \to \mathfrak{a}E_{\mathfrak{b}}$, where $\mathfrak{a}$ and $\mathfrak{b}$ are invertible $\mathcal{O}$-ideals.

Definition 17.13. Let E/k be any elliptic curve with CM by an imaginary quadratic order $\mathcal{O}$, and let $\mathfrak{a}$ be an $\mathcal{O}$-ideal. The $\mathfrak{a}$-*torsion subgroup* of E is defined by

$$E[\mathfrak{a}] := \{P \in E(\bar{k}) : \alpha(P) = 0 \text{ for all } \alpha \in \mathfrak{a}\},$$

where we are viewing each $\alpha \in \mathfrak{a} \subseteq \mathcal{O} \simeq \mathrm{End}(E)$ as an endomorphism.

Theorem 17.14. *Let $\mathcal{O}$ be an imaginary quadratic order, let $E/\mathbb{C}$ be an elliptic curve with endomorphism ring $\mathcal{O}$, let $\mathfrak{a}$ be an invertible $\mathcal{O}$-ideal, and let $\phi_{\mathfrak{a}}$ be the corresponding isogeny from E to $\mathfrak{a}E$. The following hold:*

(i) $\ker\phi_{\mathfrak{a}} = E[\mathfrak{a}]$;

(ii) $\deg\phi_{\mathfrak{a}} = \mathrm{N}\mathfrak{a}$.

Proof. By composing $\phi_{\mathfrak{a}}$ with an isomorphism if necessary, we assume without loss of generality that $E = E_{\mathfrak{b}}$ for some invertible $\mathcal{O}$-ideal $\mathfrak{b}$. Let Φ be the isomorphism from $\mathbb{C}/\mathfrak{b} \to E_{\mathfrak{b}}$

that sends z to $(\wp(z), \wp'(z))$. We have

$$\begin{aligned}
\Phi^{-1}(E[\mathfrak{a}]) &= \{z \in \mathbb{C}/\mathfrak{b} : \alpha z = 0 \text{ for all } \alpha \in \mathfrak{a}\} \\
&= \{z \in \mathbb{C} : \alpha z \in \mathfrak{b} \text{ for all } \alpha \in \mathfrak{a}\}/\mathfrak{b} \\
&= \{z \in \mathbb{C} : z\mathfrak{a} \subseteq \mathfrak{b}\}/\mathfrak{b} \\
&= \{z \in \mathbb{C} : z\mathcal{O} \subseteq \mathfrak{a}^{-1}\mathfrak{b}\}/\mathfrak{b} \\
&= (\mathfrak{a}^{-1}\mathfrak{b})/\mathfrak{b} \\
&= \ker\left(\mathbb{C}/\mathfrak{b} \xrightarrow{z \to z} \mathbb{C}/\mathfrak{a}^{-1}\mathfrak{b}\right) \\
&= \Phi^{-1}(\ker \phi_{\mathfrak{a}}),
\end{aligned}$$

which proves (i). We then note that

$$\#E[\mathfrak{a}] = [\mathfrak{a}^{-1}\mathfrak{b} : \mathfrak{b}] = [\mathfrak{b} : \mathfrak{a}\mathfrak{b}] = [\mathcal{O} : \mathfrak{a}\mathcal{O}] = [\mathcal{O} : \mathfrak{a}] = \mathrm{N}\mathfrak{a},$$

which proves (ii). □

Corollary 17.15. *Let $\mathcal{O}$ be an imaginary quadratic order and let $\mathfrak{a}$ be an invertible $\mathcal{O}$-ideal. For every elliptic curve $E/\mathbb{C}$ with CM by $\mathcal{O}$ the elliptic curves E and $\mathfrak{a}E$ are related by an isogeny $\phi_{\mathfrak{a}} \colon E \to \mathfrak{a}E$ of degree $\mathrm{N}\mathfrak{a}$.*

Proof. This follows immediately from the theorem and discussion above. □

17.6 Discriminants

To streamline our work with imaginary quadratic orders, we define the *discriminant* of $\mathcal{O}$, a negative integer that uniquely determines $\mathcal{O}$. Since $\mathcal{O}$ is a subring of an imaginary quadratic field that has rank 2 as a $\mathbb{Z}$-module, we can always write $\mathcal{O}$ as $[1, \tau]$, where τ is an algebraic integer that does not lie in $\mathbb{Z}$; its minimal polynomial is necessarily of the form $x^2 + bx + c$ with discriminant $b^2 - 4c \in \mathbb{Z}_{<0}$.

Definition 17.16. Let $\mathcal{O} = [1, \tau]$ be an imaginary quadratic order. The *discriminant* of $\mathcal{O}$ is the discriminant of the minimal polynomial of τ, which we can compute as

$$\mathrm{disc}(\mathcal{O}) = (\tau + \bar{\tau})^2 - 4\tau\bar{\tau} = (\tau - \bar{\tau})^2 = \det\begin{pmatrix} 1 & \tau \\ 1 & \bar{\tau} \end{pmatrix}^2.$$

If A is the area of a fundamental parallelogram of $\mathcal{O}$ then

$$\mathrm{disc}(\mathcal{O}) = (\tau - \bar{\tau})^2 = -4|\operatorname{im}\tau|^2 = -4A^2,$$

thus the discriminant does not depend on our choice of τ, it is intrinsic to the lattice $\mathcal{O}$.

Since the discriminant $\mathrm{disc}(\mathcal{O})$ is a negative integer of the form $b^2 - 4c$ with $b, c \in \mathbb{Z}$, it is necessarily a square modulo 4 (hence congruent to 0 or 1 mod 4).

Definition 17.17. A negative integer D that is a square modulo 4 is an (imaginary quadratic) *discriminant*. Discriminants not of the form u^2D' for some integer $u > 1$ and discriminant D' are said to be *fundamental*. Every discriminant can be written uniquely as the product of a square and a fundamental discriminant.

There is a one-to-one relationship between imaginary quadratic discriminants and orders in imaginary quadratic fields; fundamental discriminants correspond to maximal orders.

Theorem 17.18. *Let D be an imaginary quadratic discriminant. There is a unique imaginary quadratic order $\mathcal{O}$ with $\operatorname{disc}(\mathcal{O}) = D = u^2 D_K$, where D_K is the fundamental discriminant of the maximal order $\mathcal{O}_K$ in $K = \mathbb{Q}(\sqrt{D_K})$, and $u = [\mathcal{O}_K : \mathcal{O}]$.*

Proof. Write $D = \operatorname{disc}(\mathcal{O})$ as $D = u^2 D_K$, with $u \in \mathbb{Z}_{>0}$ and D_K a fundamental discriminant. Let $K = \mathbb{Q}(\sqrt{D_K})$, and let $\mathcal{O}_K$ be its ring of integers, the maximal order of K, by Theorem 12.26. Now define

$$\tau := \begin{cases} \frac{\sqrt{D_K}}{2} & \text{if } D_K \equiv 0 \bmod 4; \\ \frac{1+\sqrt{D_K}}{2} & \text{if } D_K \equiv 1 \bmod 4. \end{cases}$$

Then $\operatorname{disc}([1,\tau]) = (\tau - \bar{\tau})^2 = D_K$, and $\tau + \bar{\tau}$ and $\tau\bar{\tau}$ are integers, so $\tau \in O_K$ and $[1,\tau]$ is a suborder of $\mathcal{O}_K$. But $\mathcal{O}_K$ is the maximal order of K, so $\mathcal{O}_K = [1,\tau]$ and $\operatorname{disc}(\mathcal{O}_K) = D_K$. The order $\mathcal{O} = [1, u\tau]$ then has discriminant $(u\tau - \overline{u\tau})^2 = u^2 D_K = D$.

Conversely, if $\mathcal{O} = [1,\omega]$ is any imaginary quadratic order of discriminant D, then ω is the root of a quadratic equation of discriminant D and therefore an algebraic integer in the field $\mathbb{Q}(\sqrt{D}) = \mathbb{Q}(\sqrt{D_K}) = K$. We must have $\mathcal{O} \subseteq \mathcal{O}_K$, since $\mathcal{O}_K$ is the unique maximal order. The ratio of the squares of the areas of the fundamental parallelograms of $\mathcal{O}_K$ and $\mathcal{O}$ must be $D/D_K = u^2$, which implies $[\mathcal{O}_K : \mathcal{O}] = u$. Let $\mathcal{O}_K = [1,\tau]$ with τ defined as above. By Lemma 17.19 below, $u\mathcal{O}_k \subseteq \mathcal{O}$, so $u\tau \in \mathcal{O}$, and the lattice $[1, u\tau] \subseteq \mathcal{O}$ has index u in $\mathcal{O}_K$ and is therefore equal to $\mathcal{O}$. It follows that $[1, u\tau]$ is the unique imaginary quadratic order of discriminant D. □

The index $u = [\mathcal{O}_K : \mathcal{O}]$ is also called the *conductor* of the order $\mathcal{O}$.

Lemma 17.19. *If L' is an index n sublattice of L then nL is an index n sublattice of L'.*

Proof. Without loss of generality, $L = [1,\tau]$ and $L' = [a, b + c\tau]$ (let a be the least positive integer in L'). Comparing areas of fundamental parallelograms yields

$$\begin{aligned} n|\operatorname{im}\tau| &= |a \operatorname{im} c\tau| = |ac||\operatorname{im}\tau| \\ n &= |ac|, \end{aligned}$$

Thus $a|n$, so $n \in L'$, and $a(b+c\tau) - ba = ac\tau = \pm n\tau$, so $n\tau \in L'$; therefore $nL = [n, n\tau] \subseteq L'$. We have $[L : L'] = n$ and $[L : L'][L' : nL] = [nL : L] = n^2$, so $[L' : nL] = n$. □

References

[1] David A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory, and complex multiplication*, second edition, Wiley, 2013.

18 Riemann surfaces and modular curves

Let $\mathcal{O}$ be an order in an imaginary quadratic field and let $\mathrm{cl}(\mathcal{O})$ be its ideal class group (proper $\mathcal{O}$-ideals up to homothety, or equivalently, invertible fractional $\mathcal{O}$-ideals modulo invertible principal $\mathcal{O}$-ideals). In the previous lecture we showed that the set

$$\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) := \{j(E) : E/\mathbb{C} \text{ with } \mathrm{End}(E) = \mathcal{O}\}$$

of isomorphism classes of elliptic curves $E/\mathbb{C}$ with complex multiplication by $\mathcal{O}$ is a torsor for the group $\mathrm{cl}(\mathcal{O})$. If $\mathfrak{a}$ and $\mathfrak{b}$ are proper $\mathcal{O}$-ideals and $E_{\mathfrak{b}}$ is the elliptic curve corresponding to the complex torus $\mathbb{C}/\mathfrak{b}$, then $E_{\mathfrak{b}}$ has CM by $\mathcal{O}$ and the $\mathcal{O}$-ideal $\mathfrak{a}$ acts on $E_{\mathfrak{b}}$ via

$$\mathfrak{a}E_{\mathfrak{b}} = E_{\mathfrak{a}^{-1}\mathfrak{b}}.$$

The isogeny $\phi_{\mathfrak{a}} \colon E_{\mathfrak{b}} \to \mathfrak{a}E_{\mathfrak{b}}$ induced by the lattice inclusion $\mathfrak{b} \subseteq \mathfrak{a}^{-1}\mathfrak{b}$ has kernel

$$\ker \phi_{\mathfrak{a}} = E_{\mathfrak{b}}[\mathfrak{a}] := \{P \in E_{\mathfrak{b}}(\mathbb{C}) : \alpha P = 0 \text{ for all } \alpha \in \mathfrak{a} \subseteq \mathcal{O} \simeq \mathrm{End}(E_{\mathfrak{b}})\},$$
$$\#\ker \phi_{\mathfrak{a}} = \deg \phi_{\mathfrak{a}} = \mathrm{N}\mathfrak{a} := [\mathcal{O} : \mathfrak{a}].$$

To make further progress in our development of the theory of complex multiplication, we need a better understanding of the isogenies $\phi_{\mathfrak{a}}$. The key to doing so, both from a theoretical and practical perspective, is to understand the *modular curves* that "parameterize" isogenies of elliptic curves (in a sense that will be made clear in later lectures).

In this lecture our goal is simply to introduce the notion of a modular curve, beginning with the canonical example $X(1)$. Modular curves, and the *modular functions* that comprise their function fields, are a major topic in their own right, one to which entire courses are devoted; we shall necessarily only scratch the surface of this rich and beautiful subject. Our presentation is adapted from [2, V.1] and [4, I.2].

18.1 The modular curves $X(1)$ and $Y(1)$

Recall from Lecture 15 that the modular group $\Gamma := \mathrm{SL}_2(\mathbb{Z})$ acts on the upper half-plane $\mathcal{H} := \{\tau \in \mathbb{C} : \mathrm{im}\,\tau > 0\}$ via linear fractional transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tau := \frac{a\tau + b}{c\tau + d}.$$

The quotient $\mathcal{H}/\Gamma$ (the Γ-orbits of $\mathcal{H}$) is known as the *modular curve* $Y(1)$, whose points may be identified with points in the fundamental region

$$\mathcal{F} = \{z \in \mathcal{H} : \mathrm{re}(z) \in [-1/2, 1/2) \text{ and } |z| \geq 1, \text{ with } |z| > 1 \text{ if } \mathrm{re}(z) > 0\}.$$

You may be wondering why we call $Y(1)$ a curve. Recall from Theorem 15.11 that the j-function defines a holomorphic bijection from $\mathcal{F}$ to $\mathbb{C}$, and we shall prove that in fact $Y(1)$ is isomorphic, as a complex manifold, to the complex plane $\mathbb{C}$, which we may view as an affine curve: if we put $f(x, y) = y$ then the zero locus of f is $\{(x, 0) : x \in \mathbb{C}\} \simeq \mathbb{C}$.

The fundamental region $\mathcal{F}$ is not a compact subset of $\mathcal{H}$, since it is unbounded along the positive imaginary axis. To remedy this deficiency, we compactify it by adjoining a point at infinity to $\mathcal{H}$ and including it in $\mathcal{F}$. We want $\mathrm{SL}_2(\mathbb{Z})$ to act on our extended upper half-plane, and we want this action to be continuous, as it is on $\mathcal{H}$. Given that

$$\lim_{\mathrm{im}\,\tau \to \infty} \frac{a\tau + b}{c\tau + d} = \frac{a}{c},$$

we should also include the set of rational numbers in our extended upper half-plane. So let

$$\mathcal{H}^* = \mathcal{H} \cup \mathbb{Q} \cup \{\infty\} = \mathcal{H} \cup \mathbb{P}^1(\mathbb{Q}),$$

and let Γ act on $\mathcal{H}^*$ by extending its action on $\mathcal{H}$ to $\mathbb{P}^1(\mathbb{Q})$ via

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} (x : y) = (ax + by : cx + dy).$$

The points in $\mathcal{H}^* - \mathcal{H} = \mathbb{P}^1(\mathbb{Q})$ are called *cusps*; as you proved on Problem Set 8, the cusps are all Γ-equivalent. Thus we may extend our fundamental region $\mathcal{F}$ for $\mathcal{H}$ to a fundamental region $\mathcal{F}^*$ for $\mathcal{H}^*$ by including a single cusp: the point $\infty = (1 : 0) \in \mathbb{P}^1(\mathbb{Q})$, which we may view as a point lying infinitely far up the positive imaginary axis.

We can now define the modular curve $X(1) = \mathcal{H}^*/\Gamma$, which contains all the points in $Y(1)$, plus the cusp at infinity. This is a projective curve, in fact it is the projective closure of $Y(1)$ in $\mathbb{P}^2$. It is also a *Riemann surface*, a connected complex manifold of dimension one. Before stating precisely what this means, our first goal is to prove that $X(1)$ is a compact Hausdorff space.

We extend the topology of $\mathcal{H}$ to a topology on $\mathcal{H}^*$ by taking as a basis of open neighborhoods:

- $\tau \in \mathcal{H}$: all open disks about τ that lie in $\mathcal{H}$;
- $\tau \in \mathbb{Q}$: all sets $\{\tau\} \cup D$, where $D \subseteq \mathcal{H}$ is an open disk tangent to the real line at τ;
- $\tau = \infty$: all sets of the form $\{\tau \in \mathcal{H} : \operatorname{im} \tau > r\}$ for any $r > 0$;

The topology of $\mathcal{H}^*$ is generated by these open neighborhoods under unions and finite intersections; note that the induced subspace topology on $\mathcal{H}$ is just its standard topology.

It is clear that $\mathcal{H}^*$ is a Hausdorff space (any two points can be separated by neighborhoods). It does not immediately follow that $X(1) = \mathcal{H}^*/\Gamma$ is a Hausdorff space; a quotient of a Hausdorff space need not be Hausdorff. To show that $X(1)$ is Hausdorff we first prove two lemmas that will be useful in what follows.

Lemma 18.1. *For any compact sets $A, B \subseteq \mathcal{H}$ the set $S = \{\gamma \in \Gamma : \gamma A \cap B \neq \emptyset\}$ is finite.*

Proof. Recall that for any $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ we have

$$\operatorname{im} \gamma\tau = \operatorname{im} \frac{a\tau + b}{c\tau + d} = \operatorname{im} \frac{(a\tau + b)(ctau + d)}{|c\tau + d|^2} = \frac{(ad - bc) \operatorname{im} \tau}{|c\tau + d|^2} = \frac{\operatorname{im} \tau}{|c\tau + d|^2}.$$

Now define

$$r := \max\{\operatorname{im} \tau_A / \operatorname{im} \tau_B : \tau_A \in A, \tau_B \in B\}.$$

If $\gamma\tau_A = \tau_B$ for some $\tau_A \in A$ and $\tau_B \in B$, then $|c\tau_A + d|^2 = \operatorname{im} \tau_A / \operatorname{im} \tau_B \leq r$, which implies upper bounds on $|c|$ and $|d|$ for any $\gamma \in S$. Thus the number of pairs (c, d) arising among $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in S$ is finite. Let us now fix one such pair and define

$$s = \max\{|\tau_B||c\tau_A + d| : \tau_A \in A, \tau_B \in B\}.$$

For any $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ we have $|\gamma\tau| = |a\tau + b|/|c\tau + d|$. If $\gamma\tau_A = \tau_B$ for some $\tau_A \in A$ and $\tau_B \in B$, then $|a\tau_A + b| = |\tau_B||c\tau_A + d| \leq s$, which gives upper bounds on $|a|$ and $|b|$ as above. The number of pairs (a, b) arising among $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in S$ is thus finite, hence S is finite. □

Lemma 18.2. *For any $\tau_1, \tau_2 \in \mathcal{H}^*$ there exist open neighborhoods U_1, U_2 of τ_1, τ_2 such that*

$$\gamma U_1 \cap U_2 \neq \emptyset \quad \iff \quad \gamma\tau_1 = \tau_2,$$

for all $\gamma \in \Gamma$. In particular, each $\tau \in \mathcal{H}^*$ has an open neighborhood U in which it is the sole representative of its Γ-orbit and $\gamma U \cap \gamma' U = \emptyset$ for all $\gamma, \gamma' \in \Gamma$ such that $\gamma\tau \neq \gamma'\tau$.

Proof. We first note that if $\gamma\tau_1 = \tau_2$, then $\gamma U_1 \cap U_2 \neq \emptyset$ for all open neighborhoods U_1, U_2 of τ_1, τ_2, so we only need to consider γ for which $\gamma\tau_1 \neq \tau_2$.

We first consider $\tau_1, \tau_2 \in \mathcal{H}$. Let $C_1, C_2 \subseteq \mathcal{H}$ be closed disks about τ_1, τ_2 and define

$$S(C_1, C_2) := \{\gamma \in \Gamma : \gamma C_1 \cap C_2 \neq \emptyset \text{ and } \gamma\tau_1 \neq \tau_2\}.$$

If S is nonempty, pick $\gamma \in S$, and let U_3 and U_2' be disjoint open neighborhoods of $\gamma\tau_1$ and τ_2 respectively (they exist because $\mathcal{H}$ is Hausdorff). Then $\gamma^{-1}U_3$ is an open neighborhood of τ_1 (since γ acts continuously), and it contains a closed disk $C_1' \subseteq C_1$ about τ_1, and the open set U_2' similarly contains a closed disk $C_2' \subseteq C_2$ about τ_2. We then have $S(C_1', C_2') \subsetneq S(C_1, C_2)$, since by construction, $\gamma \notin S(C_1', C_2')$. By Lemma 18.1, S is finite, so if we continue in this fashion we will eventually have $S(C_1, C_2) = \emptyset$, at which point we may take U_1, U_2 to be the interiors of C_1, C_2.

We now consider $\tau_1 \in \mathcal{H}$ and $\tau_2 = \infty$. Let U_1 be a neighborhood of τ_1 with $\overline{U}_1 \subseteq \mathcal{H}$. The set $\{|c\tau + d| : \tau \in U_1, c, d \in \mathbb{Z} \text{ not both } 0\}$ is bounded below, and $\{\operatorname{im}\gamma\tau : \gamma \in \Gamma, \tau \in U_1\}$ is bounded above, say by r, since $\operatorname{im}\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)\tau = \operatorname{im}\tau/|c\tau + d|^2$. If we let $U_2 = \{\tau : \operatorname{im}\tau > r\}$ be our neighborhood of $\tau_2 = \infty$, then $\gamma U_1 \cap U_2 = \emptyset$ for all $\gamma \in \Gamma$ and the lemma holds. This argument extends to all the cusps in $\mathcal{H}^*$, since every cusp is Γ-equivalent to ∞, and we can easily reverse the roles of τ_1 and τ_2, since if $\gamma U_1 \cap U_2 = \emptyset$ then $U_1 \cap \gamma^{-1}U_2 = \emptyset$.

Finally, if $\tau_1 = \tau_2 = \infty$ we let $U_1 = U_2 = \{\tau \in \mathcal{H} : \operatorname{im}\tau > 1\} \cup \{\infty\}$: for $\operatorname{im}\tau > 1$ either $\operatorname{im}\gamma\tau = \operatorname{im}\tau$, in which case $\gamma = \left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right)$ fixes ∞, or $\operatorname{im}\gamma\tau = \operatorname{im}\tau/|c\tau + d|^2 < 1$.

To prove the last statement in the lemma, take $\tau_1 = \tau_2 = \tau$ and $U = U_1 \cap U_2$. □

Theorem 18.3. *$X(1)$ is a connected compact Hausdorff space.*

Proof. It is clear that $\mathcal{H}$ is connected, hence its closure $\mathcal{H}^*$ is connected, and the quotient of a connected space is connected, so $X(1)$ is connected.

To show that $X(1)$ is compact, we show that every open cover has a finite subcover. Let $\{U_i\}$ be an open cover of $X(1)$ and let $\pi\colon \mathcal{H}^* \to X(1)$ be the quotient map. Then $\{\pi^{-1}(U_i)\}$ is an open cover of $\mathcal{H}^*$ and it contains an open set V_0 containing the point ∞. Let $\{V_1, \dots, V_n\}$ be a finite subset of $\{\pi^{-1}(U_i)\}$ covering the compact set $\overline{\mathcal{F}} - V_0$ (note that V_0 contains a neighborhood $\{z : \operatorname{im} z > r\}$ of ∞). Then $\{V_0, \dots, V_n\}$ is a finite cover of $\mathcal{F}^*$, and $\{\pi(V_0), \dots, \pi(V_n)\}$ is a finite subcover of $\{U_i\}$.

To show that $X(1)$ is Hausdorff, let $x_1, x_2 \in X(1)$ be distinct, and choose τ_1, τ_2 so that $\pi(\tau_1) = x_1$ and $\pi(\tau_2) = x_2$. Then $\tau_2 \neq \gamma\tau_1$ for all $\gamma \in \Gamma$ (since $x_1 \neq x_2$), so by Lemma 18.2, there are neighborhoods U_1 and U_2 of τ_1 and τ_2 respectively for which $\gamma U_1 \cap U_2 = \emptyset$ for all $\gamma \in \Gamma$. Thus $\pi(U_1)$ and $\pi(U_2)$ are disjoint neighborhoods of x_1 and x_2. □

We note that Lemmas 18.1 and 18.2 and Theorem 18.3 all hold if we replace Γ by any finite index subgroup of Γ; the proofs are essentially the same, the only difference is an additional argument in the proof of Lemma 18.2 to handle inequivalent cusps.

18.2 Riemann surfaces

Definition 18.4. A *complex structure* on a topological space X is an open cover $\{U_i\}$ of X together with a set of compatible homeomorphisms[1] $\psi_i\colon U_i \to \mathbb{C}$ with open images. Homeomorphisms ψ_i and ψ_j are compatible if whenever $U_i \cap U_j \neq \emptyset$ the *transition map*

$$\psi_j \circ \psi_i^{-1}\colon\ \psi_i(U_i \cap U_j) \to \psi_j(U_i \cap U_j)$$

is holomorphic.

The homeomorphisms ψ_i are called *charts* (or *local parameters*), and the collection $\{\psi_i\}$ is called an *atlas*. Each chart ψ_i allows us to view a local neighborhood U_i of X as a region of the complex plane, and the transition maps allow us to move smoothly from one region to another. Note that transition maps are automatically homeomorphisms; the requirement that they be holomorphic is a stronger condition (this is what differentiates complex manifolds from real manifolds).

Definition 18.5. A *Riemann surface* is a connected Hausdorff space with a complex structure (equivalently, it is a connected complex manifold of dimension one).[2]

Example 18.6. The torus $\mathbb{C}/L$ corresponding to an elliptic curve $E/\mathbb{C}$ is a Riemann surface. To give $\mathbb{C}/L$ a complex structure let $\pi\colon \mathbb{C} \to \mathbb{C}/L$ be the quotient map, let $r > 0$ be less than half the length of the shortest vector in L, and for each $z \in \mathbb{C}$ in a fundamental region for L, let $U_z \subseteq \mathbb{C}$ be the open disk of radius r centered at z. The restriction of π to each U_z is injective (by our choice of r) and defines a homeomorphism. We may thus take $\{\pi(U_z)\}$ as our open cover and the inverse maps $\pi^{-1}\colon \pi(U_z) \to U_z$ as our charts. The transition maps are all the identity map, hence holomorphic.

It is clear that $\mathbb{C}/L$ is a connected Hausdorff space, hence a Riemann surface, in fact a compact Riemann surface. We can compute its genus by triangulating a fundamental parallelogram and computing its Euler characteristic. Recall Euler's formula

$$V - E + F = 2 - 2g,$$

where V counts vertices, E counts edges, F counts faces, and g is the genus. If $L = [\omega_1, \omega_2]$, we may triangulate the parallelogram $\overline{\mathcal{F}_0}$ by drawing a diagonal from ω_1 to ω_2. We then have $V = 1$ (every lattice point is equivalent to 0), $E = 3$ (edges on the opposite side of the parallelogram are equivalent, so 2 edges on the border plus the diagonal), and $F = 2$ (two triangles, one on each side of the diagonal). We thus have

$$1 - 3 + 2 = 2 - 2g,$$

and $g = 1$, as expected.

In order to show that $X(1)$ is a Riemann surface, we need to give it a complex structure. The only difficulty that arises when doing so occurs at points in $\mathcal{H}^*$ that possess extra symmetries under the action of Γ. We may restrict our attention to the fundamental region $\mathcal{F}^*$, and in this region there are only three points that we need to worry about, the points $i, \rho := e^{2\pi i/3}$, and ∞. We require the following lemma.

[1] Recall that a homeomorphism is a bicontinuous function, a continuous function with a continuous inverse.

[2] Some texts require Riemann surfaces to be second-countable (admit a countable basis of open sets), but in fact this requirement is automatically satisfied; this is a celebrated theorem of Radó.

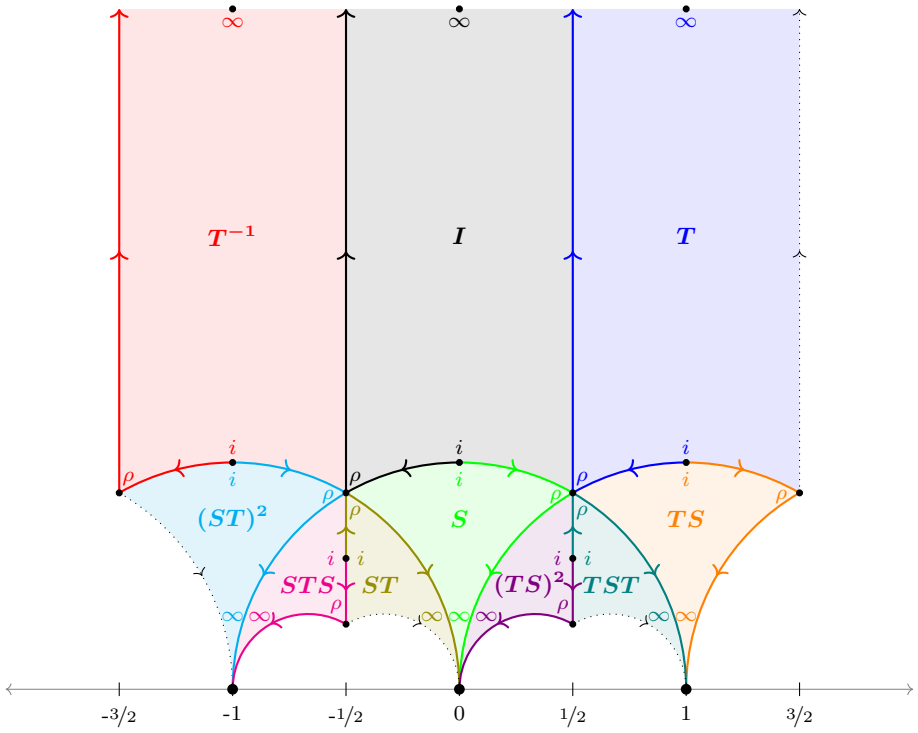


Figure 1: $\mathcal{H}^*/\Gamma$

Lemma 18.7. *For $\tau \in \mathcal{F}^*$, let G_τ denote the stabilizer of τ in $\Gamma = \mathrm{SL}_2(\mathbb{Z})$. Let $S = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Then*

$$G_\tau = \begin{cases} \{\pm I\} \simeq \mathbb{Z}/2\mathbb{Z} & \text{if } \tau \notin \{i, \rho, \infty\}; \\ \langle S \rangle \simeq \mathbb{Z}/4\mathbb{Z} & \text{if } \tau = i; \\ \langle ST \rangle \simeq \mathbb{Z}/6\mathbb{Z} & \text{if } \tau = \rho \\ \langle \pm T \rangle \simeq \mathbb{Z} & \text{if } \tau = \infty. \end{cases}$$

Proof. See Problem Set 8, or stare at Figure 1 and note $-I$ acts trivially and $T\infty = \infty$. □

18.3 The modular curve $X(1)$ as a Riemann surface

We now put a complex structure on $X(1)$. Let $\pi\colon \mathcal{H}^* \to X(1)$ be the quotient map, and for each point $x \in X(1)$ let τ_x be the unique point in the fundamental region $\mathcal{F}^*$ for which $\pi(\tau_x) = x$, and let $G_x = G_{\tau_x}$ be the stabilizer of τ_x. For each $\tau_x \in \mathcal{F}^*$, we can pick a neighborhood U_x such that $\gamma U_x \cap U_x = \emptyset$ for all $\gamma \notin G_x$, by Lemma 18.2. The sets $\pi(U_x)$ form an open cover of $X(1)$. For $x \neq \infty$, we can map U_x to an open subset of the unit disk $\mathcal{D} := \{z \in \mathbb{C} : |z| < 1\}$ via the homeomorphism $\delta_x\colon \mathcal{H} \to \mathcal{D}$ defined by

$$\delta_x(\tau) := \frac{\tau - \tau_x}{\tau - \overline{\tau}_x}. \tag{1}$$

To visualize the map δ_x, note that it sends τ_x to the origin, and if we extend its domain to $\overline{\mathcal{H}} \subseteq \mathbb{C}$, it maps the real line to the unit circle minus the point 1 and sends ∞ to 1. Note that $\operatorname{im}\tau > 0$ and $\operatorname{im}\overline{\tau}_x < 0$, so $\delta_x(\tau)$ is defined and nonzero for all $\tau \in \mathcal{H}$.

To define ψ_x we need to map $\pi(U_x)$ into $\mathcal{D}$. For $\tau_x \neq i, \rho, \infty$ we have $G_x = \{\pm 1\}$, which fixes every point in U_x, not just τ_x. In this case the restriction of π to U_x is injective, we have $U_x/\Gamma = U_x/G_x = U_x$, so we can simply define $\psi_x := \delta_x \circ \pi^{-1}$.

When $|G_x| > 2$, the restriction of π to U_x is no longer injective (it is at τ_x, but not at points near τ_x), so we cannot use $\psi_x = \delta_x \circ \pi^{-1}$. We instead define $\psi_x(z) = \delta_x(\pi^{-1}(z))^n$, where $n = |G_x|/2$ is the size of the Γ-orbits in $U_x - \{\tau_x\}$. Note that when $G_x = \{\pm 1\}$ we have $n = 1$ and this is the same as defining $\psi_x = \delta_x \circ \pi^{-1}$. To prove that this actually works, we will need the following lemma.

Lemma 18.8. *Let $\tau_x \in \mathcal{H}$, with $\delta_x(\tau)$ as in (1), and let $\varphi\colon \mathcal{H} \to \mathcal{H}$ be a holomorphic function fixing τ_x whose n-fold composition with itself is the identity, with n minimal. Then for some primitive nth root of unity ζ, we have $\delta_x(\varphi(\tau)) = \zeta\delta_x(\tau)$ for all $\tau \in \mathcal{H}$.*

Proof. The map $f = \delta_x \circ \varphi \circ \delta_x^{-1}$ is a holomorphic bijection (conformal map) from $\mathcal{D}$ to $\mathcal{D}$ that fixes 0. Every such function is a rotation $f(z) = \zeta z$ with $|\zeta| = 1$, by [5, Cor. 8.2.3]. Since the n-fold composition of f with itself is the identity map, with n minimal, ζ must be a primitive nth root of unity. □

What about $x = \infty$? We have $G_\infty = \langle \pm T \rangle$, so the intersection of the Γ-orbit of any point $\tau \in U_\infty - \{\infty\}$ with U_∞ is the set $\{\tau + m : m \in \mathbb{Z}\}$. We now define

$$\delta_\infty(z) := \begin{cases} e^{2\pi i z} & \text{if } z \neq \infty, \\ 0 & \text{if } z = \infty, \end{cases}$$

and let $\psi_\infty = \delta_\infty \circ \pi^{-1}$. Then $\delta_\infty(\tau + m) = \delta_\infty(\tau)$ for all $\tau \in U_\infty - \{\infty\}$ and $m \in \mathbb{Z}$.

The following commutative diagrams summarize the charts ψ_x:

$$\begin{array}{ccc} U_x & \xrightarrow{\pi} & U_x/G_x \\ \downarrow{\scriptstyle \delta_x} & & \downarrow{\scriptstyle \psi_x} \\ \mathcal{D} & \xrightarrow{z^n} & \mathcal{D} \end{array} \qquad \begin{array}{ccc} U_x & \xrightarrow{\pi} & U_x/G_x \\ & \searrow^{\delta_x} & \downarrow{\scriptstyle \psi_x} \\ & & \mathcal{D} \end{array}$$

$$x \neq \infty,\ \delta_x(\tau) = \tfrac{\tau - \tau_x}{\tau - \overline{\tau}_x} \qquad x = \infty,\ \delta_x(\tau) = e^{2\pi i \tau}$$
$$n = |G_x|/2$$

We are now ready to prove that $X(1)$ is a compact Riemann surface. Theorem 18.3 states that $X(1)$ is a connected compact Hausdorff space, so we just need to prove that we have a complex structure on $X(1)$. This means verifying that the maps $\psi_x\colon \pi(U_x) \to \mathcal{D}$ are well-defined (we must have $\psi(\pi(\gamma\tau)) = \psi(\pi(\tau))$ for all $\tau \in U_x$ and $\gamma \in G_x$), that they are homeomorphisms, and that the transition maps are holomorphic.

Theorem 18.9. *The open cover $\{U_x\}$ and atlas $\{\psi_x\}$ define a complex structure on $X(1)$.*

Proof. As above, let $x = \pi(\tau_x)$ with $\tau_x \in \mathcal{F}^*$. We first verify that the maps ψ_x are well-defined homeomorphisms.

We first consider $x \neq \infty$. By Lemma 18.7, the stabilizer G_x of τ_x is cyclic of order $2n$, and $\gamma^n = \pm 1$ acts trivially for all $\gamma \in G_x$. Applying Lemma 18.8 to the function $\varphi(\tau) = \gamma\tau$, we have $\delta_x(\gamma z) = \zeta\delta_x(z)$ for all $z \in U_x$, where ζ is a primitive nth root of unity. Thus

$$\psi_x(\pi(\gamma z)) = \delta_x(\gamma z)^n = \zeta^n \delta_x(z)^n = \delta_x(z)^n = \psi_x(\pi(z))$$

for all $z \in U_x$. It follows that ψ_x is well defined on U_x/G_x. To show that ψ_x is a homeomorphism, it suffices to show that it is holomorphic and injective, by the open mapping theorem [5, Thm. 5.5.4]. It is clearly holomorphic, since $\delta_x(\tau)$ is a rational function with no poles in U_x. To prove injectivity, assume $\psi_x(\pi(\tau_1)) = \psi_x(\pi(\tau_2))$. Then for some integer k

$$\begin{aligned}\delta_x(\tau_1)^n &= \delta_x(\tau_2)^n\\ \delta_x(\tau_1) &= \zeta^k \delta_x(\tau_2) = \delta_x(\gamma^k \tau_2)\\ \tau_1 &= \gamma^k \tau_2\\ \pi(\tau_1) &= \pi(\tau_2).\end{aligned}$$

Thus ψ_x is injective and therefore a homeomorphism.

For $x = \infty$, the point $\tau = \infty \in \mathcal{H}^*$ is the unique point in U_∞ for which $\pi(\tau) = \infty$, and $\psi_x(\tau) = 0$ if and only if $\tau = \infty$. So ψ_∞ is well defined at ∞. For $\tau \in U_\infty - \{\infty\}$, we have

$$\psi_\infty(\pi(\tau + m)) = \delta_\infty(\tau + m) = e^{2\pi i(\tau+m)} = e^{2\pi i \tau} = \delta_\infty(\tau) = \psi_\infty(\pi(\tau))$$

for all $m \in \mathbb{Z}$, thus ψ_∞ is well defined. The map ψ_∞ is clearly continuous, and it has a continuous inverse

$$\psi_\infty^{-1}(z) = \begin{cases} \pi\left(\frac{1}{2\pi i} \log z\right) & \text{if } z \neq 0,\\ \infty & \text{otherwise,}\end{cases}$$

thus it is a homeomorphism.

We now show that the transition maps are holomorphic. Let us first consider U_x, U_y with $x, y \neq \infty$. For any $z \in \psi_x(\pi(U_x) \cap \pi(U_y)) \subseteq \mathcal{D}$ we have

$$\psi_y \circ \psi_x^{-1}(z) = \psi_y \circ \pi \circ \pi^{-1} \circ \psi_x^{-1}(z) = (\psi_y \circ \pi) \circ (\psi_x \circ \pi)^{-1}(z) = \delta_y^{n_y} \circ \delta_x^{-1}(z^{1/n_x}),$$

where $n_x = |G_x|/2$ and $n_y = |G_y|/2$. The map $\delta_y^{n_y} \circ \delta_x^{-1}$ is holomorphic on $\mathcal{D}$, so it suffices to show that it is a power series in z^{n_x}; this will imply that $\delta_y^{n_y} \circ \delta_x^{-1}(z^{1/n_x})$ is defined by a power series in z, hence holomorphic. Let ζ be an n_xth root of unity such that $\delta_x(\gamma z) = \zeta \delta_x(z)$, where γ generates G_x, as in Lemma 18.8. Note that $\pi \circ \gamma = \pi$ for any $\gamma \in \Gamma$, so we have

$$\delta_y^{n_y} \circ \delta_x^{-1}(\zeta z) = (\psi_y \circ \pi) \circ (\gamma \circ \delta_x^{-1}(z)) = \psi_y \circ \pi \circ \delta_x^{-1}(z) = \delta_y^{n_y} \circ \delta_x^{-1}(z).$$

It follows that $\delta_y^{n_y} \circ \delta_x^{-1}$ is a power series in z^{n_x}, since it maps ζz and z to the same point.

For $x \neq \infty$ and $y = \infty$ we have

$$\begin{aligned}\psi_\infty \circ \psi_x^{-1}(z) &= \psi_y \circ \pi \circ \pi^{-1} \circ \psi_x^{-1}(z) = (\psi_y \circ \pi) \circ (\psi_x \circ \pi)^{-1}(z)\\ &= \delta_\infty \circ \delta_x^{-1}(z^{1/n_x}) = \exp\left(2\pi i \ \delta_x^{-1}(z^{1/n_x})\right),\end{aligned}$$

where $\delta_\infty \circ \delta_x^{-1}$ is holomorphic. By the argument above, it is a power series in z^{n_x}.

For the case $x = \infty$ and $y \neq \infty$, we have

$$\delta_y^{n_y}(z+1) = \psi_y \circ \pi \circ Tz = \psi_y \circ \pi(z) = \delta_y^{n_y}(z),$$

so $\delta_y^{n_y}$ is a holomorphic function in the variable $q = e^{2\pi i z}$ (note $z \in U_\infty \cap U_y$ is bounded). Thus the transition map

$$\psi_y \circ \psi_\infty^{-1}(z) = \delta_y^{n_y}\left(\frac{1}{2\pi i} \log z\right)$$

is holomorphic. The case $x = y = \infty$ is trivial, since $\psi_\infty \circ \psi_\infty^{-1}$ is the identity map. □

Theorem 18.10. *The modular curve $X(1)$ is a compact Riemann surface of genus 0.*

Proof. That $X(1)$ is a compact Riemann surface follows immediately from Theorems 18.3 and 18.9. To show that it has genus 0, we triangulate $X(1)$ by connecting the points i, ρ, and ∞, partitioning the surface into two triangles. Applying Euler's formula

$$V - E + F = 2 - 2g$$

with $V = 3$, $E = 3$, and $F = 2$, we see that $g = 0$. □

Theorem 18.10 implies that $X(1)$ is homeomorphic to the Riemann sphere $S = \mathbb{P}^1(\mathbb{C})$, since up to homeomorphism, S is the unique compact Riemann surface of genus 0. The modular curve $Y(1)$ is also a Riemann surface of genus 0, but it is not compact. As we saw in Lecture 17, $Y(1)$ is homeomorphic to the complex plane $\mathbb{C}$ via the j-function.

18.4 Modular curves

We also wish to consider modular curves defined as quotients $\mathcal{H}^*/\Gamma$ for various finite index subgroups Γ of $\mathrm{SL}_2(\mathbb{Z})$ that have desirable arithmetic properties.

Definition 18.11. The *principal congruence subgroup* $\Gamma(N)$ is defined by

$$\Gamma(N) = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \equiv \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \bmod N \right\}.$$

A *congruence subgroup* (of level N) is any subgroup of $\mathrm{SL}_2(\mathbb{Z})$ that contains $\Gamma(N)$. A *modular curve* is a quotient of $\mathcal{H}^*$ or $\mathcal{H}$ by a congruence subgroup.

Remark 18.12. Every congruence subgroup is a finite index subgroup of $\mathrm{SL}_2(\mathbb{Z})$. The converse does not hold; in fact, most finite index subgroups of $\mathrm{SL}_2(\mathbb{Z})$ are not congruence subgroups, although it is surprisingly difficult to write down explicit examples (you will have the opportunity to explore this question in Problem Set 10).

There are two families of congruence subgroups of particular interest:

$$\Gamma_1(N) := \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \equiv \left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right) \bmod N \right\};$$
$$\Gamma_0(N) := \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \equiv \left(\begin{smallmatrix} * & * \\ 0 & * \end{smallmatrix}\right) \bmod N \right\};$$

Note that $\Gamma(1) = \Gamma_1(1) = \Gamma_0(1) = \mathrm{SL}_2(\mathbb{Z})$. We now define the modular curves

$$X(N) := \mathcal{H}^*/\Gamma(N), \qquad X_1(N) := \mathcal{H}^*/\Gamma_1(N), \qquad X_0(N) := \mathcal{H}^*/\Gamma_0(N),$$

and similarly define

$$Y(N) := \mathcal{H}/\Gamma(N), \qquad Y_1(N) := \mathcal{H}/\Gamma_1(N), \qquad Y_0(N) := \mathcal{H}/\Gamma_0(N).$$

Following the same strategy we used for $X(1)$, one can show that $X(N), X_0(N), X_1(N)$ are all compact Riemann surfaces (the only difference in the proof is that in general a fundamental region may contain multiple cusps, we only had to consider the cusp ∞).

References

[1] Renzo Cavalieri and Eric Miles, *Riemann surfaces and algebraic curves*, Cambridge University Press, 2016.

[2] J. S. Milne, *Elliptic curves*, BookSurge Publishers, 2006.

[3] Rick Miranda, *Algebraic curves and Riemann surfaces*, American Mathematical Society, 1995.

[4] Joseph H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Springer, 1994.

[5] Elias M. Stein and Rami Shakarchi, *Complex analysis*, Princeton University Press, 2003.

19 The modular equation

In the previous lecture we defined modular curves as quotients of the extended upper half plane under the action of a congruence subgroup (a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ that contains a principal congruence subgroup $\Gamma(N) := \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \equiv_N \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)\}$ for some $N \in \mathbb{Z}_{>0}$). Of particular interest is the modular curve $X_0(N) := \mathcal{H}^*/\Gamma_0(N)$, where

$$\Gamma_0(N) = \left\{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}) : c \equiv 0 \bmod N\right\}.$$

This modular curve plays a central role in the theory of elliptic curves. One form of the modularity theorem (a special case of which implies Fermat's last theorem) is that every elliptic curve $E/\mathbb{Q}$ admits a morphism $X_0(N) \to E$ for some $N \in \mathbb{Z}_{\geq 1}$. It is also a key ingredient for algorithms that use isogenies of elliptic curves over finite fields, including the Schoof-Elkies-Atkin algorithm, an improved version of Schoof's algorithm that is the method of choice for counting points on elliptic curves over finite fields of large characteristic. Our immediate interest in the modular curve $X_0(N)$ is that we will use it to prove the first main theorem of complex multiplication; among other things, this theorem implies that the j-invariants of elliptic curves $E/\mathbb{C}$ with complex multiplication are algebraic integers.

There are two properties of $X_0(N)$ that make it so useful. The first, which we will prove in this lecture, is that it has a canonical model over $\mathbb{Q}$ with integer coefficients; this allows us to interpret $X_0(N)$ as a curve over any field, including finite fields. The second is that it parameterizes isogenies between elliptic curves (in a sense that we will make precise in the next lecture). In particular, given the j-invariant of an elliptic curve E and an integer N, we can use our explicit model of $X_0(N)$ to determine the j-invariants of all elliptic curves that are related to E by an isogeny whose kernel is a cyclic group of order N.

In order to better understand modular curves, we need to introduce modular functions.

19.1 Modular functions

Modular functions are meromorphic functions on a modular curve. To make this statement precise, we first need to discuss q-expansions. Let $\mathcal{D} = \{z \in \mathbb{C} : |z| < 1\}$ denote the unit disk. The map $q\colon \mathcal{H} \to \mathcal{D}$ defined by

$$q(\tau) = e^{2\pi i\tau} = e^{-2\pi \operatorname{im}\tau}\left(\cos(2\pi \operatorname{re}\tau) + i\sin(2\pi \operatorname{re}\tau)\right)$$

bijectively maps each vertical strip $\mathcal{H}_n := \{\tau \in \mathcal{H} : n \leq \operatorname{re}\tau < n+1\}$ (for any $n \in \mathbb{Z}$) to the punctured unit disk $\mathcal{D}_0 := \mathcal{D} - \{0\}$. We also note that

$$\lim_{\operatorname{im}\tau\to\infty} q(\tau) = 0.$$

If $f\colon \mathcal{H} \to \mathbb{C}$ is a meromorphic function that satisfies $f(\tau+1) = f(\tau)$ for all $\tau \in \mathcal{H}$, then we can write f in the form $f(\tau) = f^*(q(\tau))$, where $f^*\colon \mathcal{D}_0 \to \mathbb{C}$ is a meromorphic function that we can define by fixing a vertical strip $\mathcal{H}_n$ and putting $f^* := f \circ (q_{|\mathcal{H}_n})^{-1}$.

The *q-expansion* (or *q-series*) of $f(\tau)$ is obtained by composing the Laurent-series expansion of f^* at 0 with the function $q(\tau)$:

$$f(\tau) = f^*(q(\tau)) = \sum_{n=-\infty}^{+\infty} a_n q(\tau)^n = \sum_{n=-\infty}^{+\infty} a_n q^n.$$

As on the RHS above, it is customary to simply write q for $q(\tau) = e^{2\pi i\tau}$, as we shall do henceforth; but keep in mind that the symbol q denotes a function of $\tau \in \mathcal{H}$.

If f^* is meromorphic at 0 (meaning that $z^{-k}f^*(z)$ has an analytic continuation to an open neighborhood of $0 \in \mathcal{D}$ for some $k \in \mathbb{Z}_{\geq 0}$) then the q-expansion of f has only finitely many nonzero a_n with $n < 0$ and we can write

$$f(\tau) = \sum_{n=n_0}^{\infty} a_n q^n,$$

with $a_{n_0} \neq 0$, where n_0 is the order of f^* at 0. We then say that f is *meromorphic at* ∞, and call n_0 the *order of* f *at* ∞.

More generally, if f satisfies $f(\tau + N) = f(\tau)$ for all $\tau \in \mathcal{H}$, then we can write f as

$$f(\tau) = f^*(q(\tau)^{1/N}) = \sum_{n=-\infty}^{\infty} a_n q^{n/N}, \tag{1}$$

and we say that f is meromorphic at ∞ if f^* is meromorphic at 0.

If Γ is a congruence subgroup of level N, then for any Γ-invariant function f we have $f(\tau + N) = f(\tau)$ (for $\gamma = \left(\begin{smallmatrix} 1 & N \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma$ we have $\gamma\tau = \tau + N$), so f can be written as in (1), and it makes sense to say that f is (or is not) meromorphic at ∞.

Definition 19.1. Let $f : \mathcal{H} \to \mathbb{C}$ be a meromorphic function that is Γ-invariant for some congruence subgroup Γ. The function $f(\tau)$ is said to be *meromorphic at the cusps* if for every $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ the function $f(\gamma\tau)$ is meromorphic at ∞.

It follows immediately from the definition that if $f(\tau)$ is meromorphic at the cusps, then for any $\gamma \in \mathrm{SL}_2(\mathbb{Z})$ the function $f(\gamma\tau)$ is also meromorphic at the cusps. In terms of the extended upper half-plane $\mathcal{H}^*$, notice that for any $\gamma \in \mathrm{SL}_2(\mathbb{Z})$,

$$\lim_{\mathrm{im}\,\tau\to\infty} \gamma\tau \in \mathbb{P}^1(\mathbb{Q}),$$

and recall that $\mathbb{P}^1(\mathbb{Q})$ is the $\mathrm{SL}_2(\mathbb{Z})$-orbit of $\infty \in \mathcal{H}^*$, whose elements are called *cusps*. To say that $f(\gamma\tau)$ is meromorphic at ∞ is to say that $f(\tau)$ is meromorphic at $\gamma\infty$. To check whether f is meromorphic at the cusps, it suffices to consider a set of Γ-inequivalent cusp representatives $\gamma_1\infty, \gamma_1\infty, \ldots, \gamma_n\infty$, one for each Γ-orbit of $\mathbb{P}^1(\mathbb{Q})$; this is a finite set because the congruence subgroup Γ has finite index in $\mathrm{SL}_2(\mathbb{Z})$.

If f is a Γ-invariant meromorphic function, then for any $\gamma \in \Gamma$ we must have

$$\lim_{\mathrm{im}\,\tau\to\infty} f(\gamma\tau) = \lim_{\mathrm{im}\,\tau\to\infty} f(\tau)$$

whenever either limit exists, and if f is meromorphic at the cusps it must have the same order at ∞ and $\gamma\infty$ (even when the limits do not exist). Thus if f is meromorphic at the cusps it determines a meromorphic function $g\colon X_\Gamma \to \mathbb{C}$ on the modular curve $X_\Gamma := \mathcal{H}^*/\Gamma$ (as a Riemann surface). Conversely, every meromorphic function $g\colon X_\Gamma \to \mathbb{C}$ determines a Γ-invariant meromorphic function $f\colon \mathcal{H} \to \mathbb{C}$ that is meromorphic at the cusps via $f := g\circ\pi$, where π is the quotient map $\mathcal{H} \to \mathcal{H}/\Gamma$.

Definition 19.2. Let Γ be a congruence subgroup. A *modular function* for Γ is a Γ-invariant meromorphic function $f\colon \mathcal{H} \to \mathbb{C}$ that is meromorphic at the cusps; equivalently, it is a meromorphic function $g\colon X_\Gamma \to \mathbb{C}$ (as explained above).

Sums, products, and quotients of modular functions for Γ are modular functions for Γ, as are constant functions, thus the set of all modular functions for Γ forms a field $\mathbb{C}(\Gamma)$ that we view as a transcendental extension of $\mathbb{C}$. As we will shortly prove for $X_0(N)$, modular curves X_Γ are not only Riemann surfaces, they are algebraic curves over $\mathbb{C}$; the field $\mathbb{C}(\Gamma)$ of modular functions for Γ is isomorphic to the function field $\mathbb{C}(X_\Gamma)$ of $X_\Gamma/\mathbb{C}$.

Remark 19.3. In fact, every compact Riemann surface corresponds to a smooth projective (algebraic) curve over $\mathbb{C}$ that is uniquely determined up to isomorphism. Conversely, if $X/\mathbb{C}$ is a smooth projective curve then the set $X(\mathbb{C})$ can be given a topology and a complex structure that makes it a compact Riemann surface S. The function field of X and the field of meromorphic functions on S are both finite extensions of a purely transcendental extension of $\mathbb{C}$ (of transcendence degree one), and the two fields are isomorphic. We will make this isomorphism completely explicit for $X(1)$ and $X_0(N)$.

Remark 19.4. If f is a modular function for a congruence subgroup Γ, then it is also a modular function for any congruence subgroup $\Gamma' \subseteq \Gamma$, since Γ-invariance obviously implies Γ'-invariance, and the property of being meromorphic at the cusps does not depend on Γ'. Thus for all congruence subgroups Γ and Γ' we have

$$\Gamma' \subseteq \Gamma \Longrightarrow \mathbb{C}(\Gamma) \subseteq \mathbb{C}(\Gamma'),$$

and the corresponding inclusion of function fields $\mathbb{C}(X_\Gamma) \subseteq \mathbb{C}(X_{\Gamma'})$ induces a morphism $X_{\Gamma'} \to X_\Gamma$ of algebraic curves, a fact that has many useful applications.

19.2 Modular Functions for $\Gamma(1)$

We first consider the modular functions for $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$. In Lecture 15 we proved that the j-function is $\Gamma(1)$-invariant and holomorphic (hence meromorphic) on $\mathcal{H}$. To show that the $j(\tau)$ is a modular function for $\Gamma(1)$ we just need to show that it is meromorphic at the cusps. The cusps are all $\Gamma(1)$-equivalent, so it suffices to show that the $j(\tau)$ is meromorphic at ∞, which we do by computing its q-expansion. We first record the following lemma, which was used in Problem Set 8.

Lemma 19.5. *Let $\sigma_k(n) = \sum_{d|n} d^k$, and let $q = e^{2\pi i\tau}$. We have*

$$g_2(\tau) = \frac{4\pi^4}{3}\left(1 + 240\sum_{n=1}^{\infty}\sigma_3(n)q^n\right),$$

$$g_3(\tau) = \frac{8\pi^6}{27}\left(1 - 504\sum_{n=1}^{\infty}\sigma_5(n)q^n\right),$$

$$\Delta(\tau) = g_2(\tau)^3 - 27g_3(\tau)^2 = (2\pi)^{12}q\prod_{n=1}^{\infty}(1-q^n)^{24}.$$

Proof. See Washington [1, pp. 273-274]. □

Corollary 19.6. *With $q = e^{2\pi i\tau}$ we have*

$$j(\tau) = \frac{1}{q} + 744 + \sum_{n=1}^{\infty} a_n q^n,$$

where the a_n are integers.

Proof. Applying Lemma 19.5 yields

$$g_2(\tau)^3 = \frac{64}{27}\pi^{12}(1 + 240q + 2160q^2 + \cdots)^3 = \frac{64}{27}\pi^{12}(1 + 720q + 179280q^2 + \cdots),$$
$$27g_3(\tau)^2 = \frac{64}{27}\pi^{12}(1 - 504q - 16632q^2 - \cdots)^2 = \frac{64}{27}\pi^{12}(1 - 1008q + 220752q^2 + \cdots),$$
$$\Delta(\tau) = \frac{64}{27}\pi^{12}(1728q - 41472q^2 + \cdots) = \frac{64}{27}\pi^{12}1728q(1 - 24q + 252q^2 + \cdots),$$

and we then have

$$j(\tau) = \frac{1728g_2(\tau)^3}{\Delta(\tau)} = \frac{1}{q} + 744 + \sum_{n=1}^{\infty} a_n q^n,$$

with $a_n \in \mathbb{Z}$, since $1 - 24q + 252q^2 + \cdots$ is an element of $1 + \mathbb{Z}[[x]]$, hence invertible. □

Remark 19.7. The proof of Corollary 19.6 explains the factor 1728 that appears in the definition of the j-function: it is the least positive integer that ensures that the q-expansion of $j(\tau)$ has integral coefficients.

The corollary implies that the j-function is a modular function for $\Gamma(1)$, with a simple pole at ∞. We proved in Theorem 18.5 that the j-function defines a holomorphic bijection from $Y(1) = \mathcal{H}/\Gamma(1)$ to $\mathbb{C}$. If we extend the domain of j to $\mathcal{H}^*$ by defining $j(\infty) = \infty$, then the j-function defines an isomorphism from $X(1)$ to the Riemann sphere $\mathcal{S} := \mathbb{P}^1(\mathbb{C})$ that is holomorphic everywhere except for a simple pole at ∞. In fact, if we fix $j(\rho) = 0$, $j(i) = 1728$, and $j(\infty) = \infty$, then the j-function is uniquely determined by this property (as noted above, we put $j(i) = 1728$ to obtain an integral q-expansion). It is for this reason that the j-function is sometimes referred to as *the* modular function. Indeed, every modular function for $\Gamma(1) = \mathrm{SL}_2(\mathbb{Z})$ can be written in terms of the j-function.

Theorem 19.8. *Every modular function for $\Gamma(1)$ is a rational function of $j(\tau)$; in other words $\mathbb{C}(\Gamma(1)) = \mathbb{C}(j)$.*

Proof. As noted above, the j-function is a modular function for $\Gamma(1)$, so $\mathbb{C}(j) \subseteq \mathbb{C}(\Gamma(1))$. If $g\colon X(1) \to \mathbb{C}$ is a modular function for $\Gamma(1)$ then $f := g \circ j^{-1}\colon \mathcal{S} \to \mathbb{C}$ is meromorphic, and Lemma 19.9 below implies that f is a rational function. Thus $g = f \circ j \in \mathbb{C}(j)$. □

Lemma 19.9. *Every meromorphic function $f\colon \mathcal{S} \to \mathbb{C}$ on the Riemann sphere $\mathcal{S} := \mathbb{P}^1(\mathbb{C})$ is a rational function.*

Proof. Let $f\colon \mathcal{S} \to \mathbb{C}$ be a nonzero meromorphic function. We may assume without loss of generality that f has no zeros or poles at $\infty := (1 : 0)$, since we can always apply a linear fractional transformation $\gamma \in \mathrm{SL}_2(\mathbb{C})$ to move a point where f does not have a pole or a zero to ∞ and replace f by $f \circ \gamma$ (note that γ and γ^{-1} are rational functions, and if $f \circ \gamma$ is a rational function, so is $f = f \circ \gamma \circ \gamma^{-1}$).

Let $\{p_i\}$ be the set of poles of $f(z)$, with orders $m_i := -\mathrm{ord}_{p_i}(f)$, and let $\{q_j\}$ be the set of zeros of f, with orders $n_j := \mathrm{ord}_{q_j}(f)$. We claim that

$$\sum_i m_i = \sum_j n_j.$$

To see this, triangulate $\mathcal{S}$ so that all the poles and zeros of $f(z)$ lie in the interior of a triangle. It follows from Cauchy's argument principle (Theorem 14.17) that the contour integral

$$\int_\Delta \frac{f'(z)}{f(z)}\,dz$$

about each triangle (oriented counter clockwise) is the difference between the number of zeros and poles that $f(z)$ has in its interior. The sum of these integrals must be zero, since each edge in the triangulation is traversed twice, once in each direction.

The function $h\colon \mathcal{S} \to \mathbb{C}$ defined by

$$h(z) = f(z) \cdot \frac{\prod_i (z-p_i)^{m_i}}{\prod_j (z-q_j)^{n_j}}$$

has no zeros or poles on $\mathcal{S}$. It follows from Liouville's theorem (Theorem 14.30) that h is a constant function, and therefore $f(z)$ is a rational function of z. □

Corollary 19.10. *Every modular function $f(\tau)$ for $\Gamma(1)$ that is holomorphic on $\mathcal{H}$ is a polynomial in $j(\tau)$.*

Proof. Theorem 19.8 implies that f can be written as a rational function of j, so

$$f(\tau) = c\,\frac{\prod_i (j(\tau)-\alpha_i)}{\prod_k (j(\tau)-\beta_k)},$$

for some $c, \alpha_i, \beta_k \in \mathbb{C}$. Now the restriction of j to any fundamental region for $\Gamma(1)$ is a bijection, so $f(\tau)$ must have a pole at $j^{-1}(\beta_k)$ for each β_k. But $f(\tau)$ is holomorphic and therefore has no poles, so the set $\{\beta_j\}$ is empty and $f(\tau)$ is a polynomial in $j(\tau)$. □

We proved in the previous lecture that the j-function $j\colon X(1) \xrightarrow{\sim} \mathcal{S}$ determines an isomorphism of Riemann surfaces. As an algebraic curve over $\mathbb{C}$, the function field of $X(1) \simeq \mathcal{S} = \mathbb{P}^1(\mathbb{C})$ is the rational function field $\mathbb{C}(t)$, and we have just shown that the field of modular functions for $\Gamma(1)$ is the field $\mathbb{C}(j)$ of rational functions of j. Thus, as claimed in Remark 19.3, the function field $\mathbb{C}(X(1)) = \mathbb{C}(t)$ and the field of modular functions $\mathbb{C}(\Gamma(1)) = \mathbb{C}(j)$ are isomorphic, with the isomorphism given by $t \mapsto j$. More generally, for every congruence subgroup Γ, the field $\mathbb{C}(X_\Gamma) \simeq \mathbb{C}(\Gamma)$ is a finite extension of $\mathbb{C}(t) \simeq \mathbb{C}(j)$.

Theorem 19.11. *Let Γ be a congruence subgroup. The field $\mathbb{C}(\Gamma)$ of modular functions for Γ is a finite extension of $\mathbb{C}(j)$ of degree at most $n := [\Gamma(1) : \Gamma]$.*

Proof. Let γ_1 be the identity in $\Gamma(1)$ and let $\{\gamma_1, \cdots, \gamma_n\} \subseteq \Gamma(1)$ be a set of right coset representatives for Γ as a subgroup of $\Gamma(1)$ (so $\Gamma(1) = \Gamma\gamma_1 \sqcup \cdots \sqcup \Gamma\gamma_n$).

Let $f \in \mathbb{C}(\Gamma)$ and for $1 \le i \le n$ define $f_i(\tau) := f(\gamma_i\tau)$. For any $\gamma_i' \in \Gamma\gamma_i$ the functions $f(\gamma_i'\tau)$ and $f(\gamma_i\tau)$ are the same, since f is Γ-invariant. For any $\gamma \in \Gamma(1)$, the set of functions $\{f(\gamma_i\gamma\tau)\}$ is therefore equal to the set of functions $\{f(\gamma_i\tau)\}$, since multiplication on the right by γ permutes the cosets $\{\Gamma\gamma_i\}$. Any symmetric polynomial in the functions f_i is thus $\Gamma(1)$-invariant, and meromorphic at the cusps (since f, and therefore each f_i, is), hence an element of $\mathbb{C}(j)$, by Theorem 19.8. Now let

$$P(Y) = \prod_{1\le i\le n} (Y - f_i).$$

Then $f = f_1$ is a root of P (since γ_1 is the identity), and the coefficients of $P(Y)$ lie in $\mathbb{C}(j)$, since they are all symmetric polynomials in the f_i.

It follows that every $f \in \mathbb{C}(\Gamma)$ is the root of a monic polynomial in $\mathbb{C}(j)[Y]$ of degree n; this implies that $\mathbb{C}(\Gamma)/\mathbb{C}(j)$ is an algebraic extension, and it is separable, since we are in characteristic zero. We now claim that $\mathbb{C}(\Gamma)$ is finitely generated: if not we could pick functions $g_1, \dots, g_{n+1} \in \mathbb{C}(\Gamma)$ such that

$$\mathbb{C}(j) \subsetneq \mathbb{C}(j)(g_1) \subsetneq \mathbb{C}(j)(g_1, g_2) \subsetneq \cdots \subsetneq \mathbb{C}(j)(g_1, \dots, g_{n+1}).$$

But then $\mathbb{C}(j)(g_1, \dots, g_{n+1})$ is a finite separable extension of $\mathbb{C}(j)$ of degree at least $n+1$, and the primitive element theorem implies it is generated by some $g \in \mathbb{C}(\Gamma)$ whose minimal polynomial must have degree greater than n, which is a contradiction. The same argument then shows that $[\mathbb{C}(\Gamma) : \mathbb{C}(j)] \le n$. □

Remark 19.12. If $-I \in \Gamma$ then in fact $[\mathbb{C}(\Gamma) : \mathbb{C}(\Gamma(1))] = [\Gamma(1) : \Gamma]$; we will prove this for $\Gamma = \Gamma_0(N)$ in the next section. In general $[\mathbb{C}(\Gamma) : \mathbb{C}(\Gamma(1))] = [\Gamma(1) : \overline{\Gamma}]$, where $\overline{\Gamma}$ denotes the image of Γ in $\mathrm{PSL}_2(\mathbb{Z}) := \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$.

19.2.1 Modular functions for $\Gamma_0(N)$

We now consider modular functions for the congruence subgroup $\Gamma_0(N)$.

Theorem 19.13. *The function $j_N(\tau) := j(N\tau)$ is a modular function for $\Gamma_0(N)$.*

Proof. The function $j_N(\tau)$ is obviously meromorphic (in fact holomorphic) on $\mathcal{H}$, since $j(\tau)$ is, and it is meromorphic at the cusps for the same reason (note that τ is a cusp if and only if $N\tau$ is). We just need to show that $j_N(\tau)$ is $\Gamma_0(N)$-invariant.

Let $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$. Then $c \equiv 0 \bmod N$ and

$$j_N(\gamma\tau) = j(N\gamma\tau) = j\left(\frac{N(a\tau+b)}{c\tau+d}\right) = j\left(\frac{aN\tau+bN}{\frac{c}{N}N\tau+d}\right) = j(\gamma' N\tau) = j(N\tau) = j_N(\tau),$$

where

$$\gamma' = \begin{pmatrix} a & bN \\ c/N & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}),$$

since c/N is an integer and $\det(\gamma') = \det(\gamma) = 1$. Thus $j_N(\tau)$ is $\Gamma_0(N)$-invariant. □

Theorem 19.14. *The field of modular functions for $\Gamma_0(N)$ is an extension of $\mathbb{C}(j)$ of degree $n := [\Gamma(1) : \Gamma_0(N)]$ generated by $j_N(\tau)$.*

Proof. By the previous theorem, we have $j_N \in \mathbb{C}(\Gamma_0(N))$, and from Theorem 19.11 we know that $\mathbb{C}(\Gamma_0(N))$ is a finite extension of $\mathbb{C}(j)$ of degree at most n, so it suffices to show that the minimal polynomial of j_N over $\mathbb{C}(j)$ has degree at least n.

As in the proof of Theorem 19.11, let us fix right coset representatives $\{\gamma_1, \cdots, \gamma_n\}$ for $\Gamma_0(N) \subseteq \Gamma(1)$, and let $P \in \mathbb{C}(j)[Y]$ be the minimal polynomial of j_N over $\mathbb{C}(j)$. We may view $P(j(\tau), j_N(\tau))$ as a function of τ, which must be the zero function. If we replace τ by $\gamma_i\tau$ then for each γ_i we have

$$0 = P(j(\gamma_i\tau), j_N(\gamma_i\tau)) = P(j(\tau), j_N(\gamma_i\tau)),$$

so the function $j_N(\gamma_i\tau)$ is also a root of $P(Y)$.

To prove that $\deg P \geq n$ it suffices to show that the n functions $j_N(\gamma_i\tau)$ are distinct. Suppose not. Then $j(N\gamma_i\tau) = j(N\gamma_k\tau)$ for some $i \neq k$ and $\tau \in \mathcal{H}$ that we can choose to have stabilizer $\pm I$. Fix a fundamental region $\mathcal{F}$ for $\mathcal{H}/\Gamma(1)$ and pick $\alpha, \beta \in \Gamma(1)$ so that $\alpha N\gamma_i\tau$ and $\beta N\gamma_k\tau$ lie in $\mathcal{F}$. The j-function is injective on $\mathcal{F}$, so

$$j(\alpha N\gamma_i\tau) = j(\beta N\gamma_k\tau) \qquad \iff \qquad \alpha N\gamma_i\tau = \pm\beta N\gamma_k\tau \qquad \iff \qquad \alpha N\gamma_i = \pm\beta N\gamma_k,$$

where we may view N as the matrix $\left(\begin{smallmatrix} N & 0 \\ 0 & 1 \end{smallmatrix}\right)$, since $N\tau = \frac{N\tau+0}{0\tau+1}$.

Now let $\gamma = \alpha^{-1}\beta = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. We have

$$\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_i \;=\; \pm \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_k,$$

and therefore

$$\gamma_i\gamma_k^{-1} \;=\; \pm \begin{pmatrix} 1/N & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \;=\; \pm \begin{pmatrix} a & b/N \\ cN & d \end{pmatrix}.$$

We have $\gamma_i\gamma_k^{-1} \in \mathrm{SL}_2(\mathbb{Z})$, so b/N is an integer, and $cN \equiv 0 \bmod N$, so $\gamma_i\gamma_k^{-1} \in \Gamma_0(N)$. But then γ_i and γ_k lie in the same right coset of $\Gamma_0(N)$, which is a contradiction. □

19.3 The modular polynomial

Definition 19.15. The *modular polynomial* Φ_N is the minimal polynomial of j_N over $\mathbb{C}(j)$.

It follows from the proof of Theorem 19.14 that we may write $\Phi_N \in \mathbb{C}(j)[Y]$ as

$$\Phi_N(Y) = \prod_{i=1}^{n} (Y - j_N(\gamma_i\tau)),$$

where $\{\gamma_1, \ldots \gamma_n\}$ is a set of right coset representatives for $\Gamma_0(N)$. The coefficients of $\Phi_N(Y)$ are symmetric polynomials in $j_N(\gamma_i\tau)$, so as in the proof of Theorem 19.11 they are $\Gamma(1)$-invariant. They are holomorphic on $\mathcal{H}$, so they are polynomials in j, by Corollary 19.10. Thus $\Phi_N \in \mathbb{C}[j, Y]$. If we replace every occurrence of j in Φ_N with a new variable X we obtain a polynomial in $\mathbb{C}[X, Y]$ that we write as $\Phi_N(X, Y)$.

Our next task is to prove that the coefficients of $\Phi_N(X, Y)$ are actually integers, not just complex numbers. To simplify the presentation, we will only prove for prime N, which is all that is needed in most practical applications (such as the SEA algorithm), and suffices to prove the main theorem of complex multiplication. The proof for composite N is essentially the same, but explicitly writing down a set of right coset representatives γ_i and computing the q-expansions of the functions $j_N(\gamma_i\tau)$ is more complicated.

We begin by fixing a specific set of right coset representatives for $\Gamma_0(N)$.

Lemma 19.16. *For prime N we can write the right cosets of $\Gamma_0(N)$ in $\Gamma(1)$ as*

$$\Big\{\Gamma_0(N)\Big\} \cup \Big\{\Gamma_0(N)ST^k : 0 \leq k < N\Big\},$$

where $S = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

Proof. We first show that these cosets cover $\Gamma(1)$. Let $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma(1)$. If $c \equiv 0 \bmod N$, then $\gamma \in \Gamma_0(N)$ lies in the first coset. Otherwise, pick $k \in [0, N-1]$ so that $kc \equiv d \bmod N$ (c is nonzero modulo the prime N, so this is possible), and let

$$\gamma_0 := \begin{pmatrix} ka - b & a \\ kc - d & c \end{pmatrix} \in \Gamma_0(N).$$

Then

$$\gamma_0 S T^k = \begin{pmatrix} ka - b & a \\ kc - d & c \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & k \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \gamma,$$

lies in $\Gamma_0(N) S T^k$.

We now show the cosets are distinct. Suppose not. Then there must exist $\gamma_1, \gamma_2 \in \Gamma_0(N)$ such that either (a) $\gamma_1 = \gamma_2 S T^k$ for some $0 \le k < N$, or (b) $\gamma_1 S T^j = \gamma_2 S T^k$ with $0 \le j < k < N$. Let $\gamma_2 = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. In case (a) we have

$$\gamma_1 = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & k \end{pmatrix} = \begin{pmatrix} b & bk - a \\ d & dk - c \end{pmatrix} \in \Gamma_0(N),$$

which implies $d \equiv 0 \bmod N$ and $\det \gamma_2 = ad - bc \equiv 0 \bmod N$, a contradiction. In case (b), with $m = k - j$ we have

$$\gamma_1 = \gamma_2 S T^m S^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & m \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} a - bm & b \\ c - dm & d \end{pmatrix} \in \Gamma_0(N).$$

Thus $c - dm \equiv 0 \bmod N$, and since $c \equiv 0 \bmod N$ and $m \not\equiv 0 \bmod N$, we must have $d \equiv 0 \bmod N$, which again implies $\det \gamma_2 = ad - bc \equiv 0 \bmod N$, a contradiction. □

Theorem 19.17. $\Phi_N \in \mathbb{Z}[X, Y]$.

Proof (for N prime). Let $\gamma_k := S T^k$. By Lemma 19.16 we have

$$\Phi_N(Y) = \big(Y - j_N(\tau)\big) \prod_{k=0}^{N-1} \big(Y - j_N(\gamma_k \tau)\big).$$

Let $f(\tau)$ be a coefficient of $\Phi_N(Y)$. Then $f(\tau)$ is a holomorphic function on $\mathcal{H}$, since $j(\tau)$ is, and $f(\tau)$ is $\Gamma(1)$-invariant, since it is a symmetric polynomial in $j_N(\tau)$ and the functions $j_N(\gamma_k \tau)$, corresponding to a complete set of right coset representatives for $\Gamma_0(N)$; and $f(\tau)$ is meromorphic at the cusps, since it is a polynomial in functions that are meromorphic at the cusps. Thus $f(\tau)$ is a modular function for $\Gamma(1)$ holomorphic on $\mathcal{H}$ and therefore a polynomial in $j(\tau)$, by Corollary 19.10. By Lemma 19.18 below, if we can show that the q-expansion of $f(\tau)$ has integer coefficients, then it will follow that $f(\tau)$ is an integer polynomial in $j(\tau)$ and therefore $\Phi_N \in \mathbb{Z}[X, Y]$.

We first show that the q-expansion of $f(\tau)$ has rational coefficients. We have

$$j_N(\tau) = j(N\tau) = \frac{1}{q^N} + 744 + \sum_{n=1}^{\infty} a_n q^{nN},$$

where the a_n are integers, thus $j_N \in \mathbb{Z}((q))$. For $j_N(\gamma_k \tau)$, we have

$$\begin{aligned} j_N(\gamma_k \tau) = j(N \gamma_k \tau) &= j\left(\left(\begin{smallmatrix} N & 0 \\ 0 & 1 \end{smallmatrix}\right) S T^k \tau\right) \\ &= j\left(S \left(\begin{smallmatrix} 1 & 0 \\ 0 & N \end{smallmatrix}\right) \left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right) \tau\right) = j\left(\left(\begin{smallmatrix} 1 & 0 \\ 0 & N \end{smallmatrix}\right) \left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right) \tau\right) = j\left(\frac{\tau + k}{N}\right), \end{aligned}$$

where we are able to drop the S because $j(\tau)$ is Γ-invariant. If we let $\zeta_N = e^{\frac{2\pi i}{N}}$, then

$$q^{((\tau+k)/N)} = e^{2\pi i\left(\frac{\tau+k}{N}\right)} = e^{2\pi i\frac{k}{N}} q^{1/N} = \zeta_N^k q^{1/N},$$

and

$$j_N(\gamma_k\tau) = \frac{\zeta_N^{-k}}{q^{1/N}} + \sum_{n=0}^{\infty} a_n \zeta_N^{kn} q^{n/N},$$

thus $j_N(\gamma_k\tau) \in \mathbb{Q}(\zeta_N)((q^{1/N}))$. The action of the Galois group $\mathrm{Gal}(\mathbb{Q}(\zeta_N)/\mathbb{Q})$ on the coefficients of the q-expansions of each $j_N(\gamma_k\tau)$ induces a permutation of the set $\{j_N(\gamma_k\tau)\}$ and fixes $j_N(\tau)$. It follows that the coefficients of the q-expansion of f are fixed by $\mathrm{Gal}(\mathbb{Q}(\zeta_N)/\mathbb{Q})$ and must lie in $\mathbb{Q}$. Thus $f \in \mathbb{Q}((q^{1/N}))$, and $f(\tau)$ is a polynomial in $j(\tau)$, so its q-expansion contains only integral powers of q and $f \in \mathbb{Q}((q))$.

We now note that the coefficients of the q-expansion of $f(\tau)$ are algebraic integers, since the coefficients of the q-expansions of $j_N(\tau)$ and the $j_N(\gamma_k\tau)$ are algebraic integers, as is any polynomial combination of them. This implies $f(\tau) \in \mathbb{Z}((q))$. □

Lemma 19.18 (Hasse q-expansion principle)**.** *Let $f(\tau)$ be a modular function for $\Gamma(1)$ that is holomorphic on $\mathcal{H}$ and whose q-expansion has coefficients that lie in an additive subgroup A of $\mathbb{C}$. Then $f(\tau) = P(j(\tau))$, for some polynomial $P \in A[X]$.*

Proof. By Corollary 19.10, we know that $f(\tau) = P(j(\tau))$ for some $P \in \mathbb{C}[X]$, we just need to show that $P \in A[X]$. We proceed by induction on $d = \deg P$. The lemma clearly holds for $d = 0$, so assume $d > 0$. The q-expansion of the j-function begins with q^{-1}, so the q-expansion of $f(\tau)$ must have the form $\sum_{n=-d}^{\infty} a_n q^n$, with $a_n \in A$ and $a_{-d} \neq 0$. Let $P_1(X) = P(X) - a_{-d}X^d$, and let $f_1(\tau) = P_1(j(\tau)) = f(\tau) - a_{-d}j(\tau)^d$. The q-expansion of the function $f_1(\tau)$ has coefficients in A, and by the inductive hypothesis, so does $P_1(X)$, and therefore $P(X) = P_1(X) + a_{-d}X^d$ also has coefficients in A. □

References

[1] Lawrence C. Washington, *Elliptic curves: Number theory and cryptography*, second edition, Chapman & Hall/CRC, 2008.

20 The Hilbert class polynomial

In the previous lecture we proved that the field of modular functions for $\Gamma_0(N)$ is generated by the functions $j(\tau)$ and $j_N(\tau) := j(N\tau)$, that is, $\mathbb{C}(\Gamma_0(N)) = \mathbb{C}(j, j_N)$, and we showed that $\mathbb{C}(j, j_N)$ is a finite extension of $\mathbb{C}(j)$ of degree $[\Gamma(1) : \Gamma_0(N)]$. We then defined the modular polynomial $\Phi_N(Y)$ as the minimal polynomial of j_N over $\mathbb{C}(j)$ and proved that its coefficients lie in $\mathbb{Z}[j] \subseteq \mathbb{C}(j)$. Replacing j with a formal variable X, we obtain a polynomial $\Phi_N \in \mathbb{Z}[X, Y]$ that gives a canonical defining equation for the modular curve $X_0(N)$.[1]

In this lecture we will use Φ_N to prove that the *Hilbert class polynomial*[2]

$$H_D(X) := H_{\mathcal{O}}(X) := \prod_{j(E)\in \mathrm{Ell}_{\mathcal{O}}(\mathbb{C})} \Big(X - j(E)\Big)$$

also has integer coefficients; here $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) := \{j(E) : \mathrm{End}(E) \simeq \mathcal{O}\}$ is the set of j-invariants of elliptic curves $E/\mathbb{C}$ with complex multiplication (CM) by the imaginary quadratic order $\mathcal{O}$ with discriminant $D = \mathrm{disc}(\mathcal{O})$. Recall that D uniquely determines $\mathcal{O}$ (and vice versa), by Theorem 17.18, so the notation H_D is unambiguous (both H_D and $H_{\mathcal{O}}$ appear in the literature, we will use the former).

The fact that $H_D \in \mathbb{Z}[x]$ implies that the j-invariant of any elliptic curve $E/\mathbb{C}$ with complex multiplication must be an algebraic integer, meaning that E can actually be defined over a number field (a finite extension of $\mathbb{Q}$). This is a remarkable result. It implies that of the uncountably many isomorphism classes of elliptic curves over $\mathbb{C}$, only countably many have complex multiplication. In order to prove this we will exploit the interpretation of $X_0(N)$ as the "moduli space" of cyclic N-isogenies of elliptic curves; our first task is to explain what this means.

20.1 Isogenies

Recall from §17.5 in Lecture 17 that if $L_1 \subseteq L_2$ are lattices in $\mathbb{C}$, and E_1 and E_2 are the elliptic curves corresponding to the complex tori $\mathbb{C}/L_1$ and $\mathbb{C}/L_2$, then the inclusion $L_1 \subseteq L_2$ induces an isogeny $\phi\colon E_1 \to E_2$ whose kernel is isomorphic to the finite abelian group L_2/L_1. Indeed, we have the commutative diagram

$$\begin{array}{ccc} \mathbb{C}/L_1 & \xrightarrow{\ \iota\ } & \mathbb{C}/L_2 \\ \downarrow\wr & & \downarrow\wr \\ E_1(\mathbb{C}) & \xrightarrow{\ \phi\ } & E_2(\mathbb{C}) \end{array}$$

where the top map ι is induced by the inclusion $L_1 \subseteq L_2$ (lift from $\mathbb{C}/L_1$ to $\mathbb{C}$ then project to $\mathbb{C}/L_2$). If we replace L_2 by the homothetic lattice NL_2, where $N = [L_2 : L_1] = \deg\phi$, the inclusion $NL_2 \subseteq L_1$ induces an isogeny in the reverse direction which, after composing with the isomorphism corresponding to the homothety $L_2 \sim NL_2$, is the dual isogeny $\hat{\phi}\colon E_2 \to E_1$. The composition $\phi \circ \hat{\phi}$ is the multiplication-by-N map on E_2, corresponding to the lattice inclusion $NL_2 \subseteq L_2$, with kernel isomorphic to $L_2/NL_2 \simeq \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$.

[1] The curve $\Phi_N(X, Y) = 0$ is a singular affine curve with the same function field as $X_0(N)$; the desingularization of its projective closure is a smooth projective curve isomorphic to $X_0(N)$.

[2] Some authors use the term *Hilbert class polynomial* only when $\mathcal{O}$ is a maximal order (they then use the term *ring class polynomial* for the general case); we won't make this distinction.

Definition 20.1. If L_1 is a sublattice of L_2 for which the group L_2/L_1 is cyclic, then we say that L_1 is a *cyclic sublattice* of L_2. Similarly, an isogeny $\phi\colon E_1 \to E_2$ is said to be *cyclic* if its kernel is a cyclic group. If ϕ is induced by the lattice inclusion $L_1 \subseteq L_2$ then ϕ is cyclic if and only if L_1 is a cyclic sublattice of L_2.

As we proved in Corollary 5.12, up to isomorphism, every isogeny is a composition of isogenies of prime degree, which are necessarily cyclic. So we may as well restrict our attention to cyclic isogenies ϕ, which we will show correspond to points on the modular curve $X_0(N)$, with $N = \deg \phi$.

In our proofs we will often restrict to the case where N is prime. We can always decompose ϕ into a composition of isogenies of prime degree, and in fact the prime degree case will suffice for everything we want to prove. It is thus enough for us to understand cyclic sublattices of prime index.

Lemma 20.2. *Let $L = [1, \tau]$ be a lattice with $\tau \in \mathcal{H}$ and let N be prime. The cyclic sublattices of L of index N are the lattice $[1, N\tau]$ and the lattices $[N, \tau + k]$, for $0 \le k < N$.*

Proof. The lattices $[1, N\tau]$ and $[N, \tau + k]$ are clearly index N sublattices of L, and they must be cyclic sublattices, since N is prime. Conversely, any sublattice $L' \subseteq L$ can be written as $[d, a\tau + k]$, where d is the least positive integer in L' and the index of L' in L is $ad = N$. Since N is prime, either $d = 1$ and $a = N$, in which case $L' = [1, N\tau]$, or $d = N$ and $a = 1$, in which case $L' = [N, \tau + k]$, and we may assume $0 \le k < N$. □

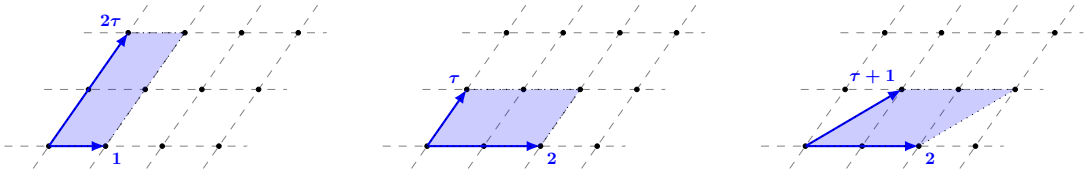


Figure 1: The three cyclic sublattices of $[1, \tau]$ of index 2.

Theorem 20.3. *For all $j_1, j_2 \in \mathbb{C}$, we have $\Phi_N(j_1, j_2) = 0$ if and only if j_1 and j_2 are the j-invariants of elliptic curves over $\mathbb{C}$ that are related by a cyclic isogeny of degree N.*

Proof for N prime. We will prove the equivalent statement that $\Phi_N\big(j(L_1), j(L_2)\big) = 0$ if and only if L_1 is homothetic to a cyclic sublattice of L_2 of index N, equivalently, L_2 is homothetic to a cyclic sublattice of L_1. We may assume without loss of generality that $L_1 = [1, \tau_1]$ and $L_2 = [1, \tau_2]$, where $\tau_1, \tau_2 \in \mathcal{H}$. As in the proof of Theorem 19.17 we have

$$\Phi_N\big(j(\tau), Y\big) = \big(Y - j(N\tau)\big) \prod_{k=0}^{N-1} \big(Y - j(N\gamma_k\tau)\big), \tag{1}$$

where $\gamma_k := ST^k$, and

$$j\Big(N\gamma_k\tau\Big) = j\Big(\left(\begin{smallmatrix} N & 0 \\ 0 & 1 \end{smallmatrix}\right) ST^k\tau\Big) = j\Big(S\left(\begin{smallmatrix} 1 & k \\ 0 & N \end{smallmatrix}\right)\tau\Big) = j\Big(\left(\begin{smallmatrix} 1 & k \\ 0 & N \end{smallmatrix}\right)\tau\Big) = j\Big(\frac{\tau + k}{N}\Big).$$

Thus

$$\Phi_N\big(j(L_1), j(L_2)\big) = \Phi_N\big(j([1, \tau_1]), j([1, \tau_2])\big) = \Phi_N\big(j(\tau_1), j(\tau_2)\big)$$

is zero if and only if τ_2 is $\mathrm{SL}_2(\mathbb{Z})$-equivalent to $N\tau_1$ or $(\tau_1 + k)/N$, with $0 \le k < N$, hence if and only if L_2 is homothetic to a cyclic sublattice of L_1 of index N, by Lemma 20.2. □

Theorem 20.3 applies more generally to any field that can be embedded in $\mathbb{C}$, including all number fields. It can be extended via the Lefschetz principle [6, Thm. VI.6.1] to any field of characteristic zero, and as shown by Igusa [4], to fields of positive characteristic $p \nmid N$. We state the more general version of Theorem 20.3 for future reference.

Theorem 20.4. *Let $N > 1$ be an integer and let k be a field of characteristic not dividing N. For all $j_1, j_2 \in k$ we have $\Phi_N(j_1, j_2) = 0$ if and only if j_1 and j_2 are the j-invariants of elliptic curves over k that are related by a cyclic isogeny of degree N defined over k.*

Remark 20.5. In Theorem 20.3 we could have written $\Phi_N(j(E_1), j(E_2)) = 0$ if and only if E_1 and E_2 are related by a cyclic isogeny of degree N, because over $\mathbb{C}$ the j-invariant characterizes elliptic curves up to isomorphism; but this is not true in the more general context of Theorem 20.4. Over fields k that are not algebraically closed it is not necessarily true that $\Phi_N(j(E_1), j(E_2)) = 0$ implies the existence of a cyclic N-isogeny $E_1 \to E_2$; one might need to replace E_1 or E_2 by a twist (a curve with the same j-invariant that is isomorphic over an extension of k but not necessarily over k).

Remark 20.6. We should note that if $\phi\colon E_1 \to E_2$ is a cyclic N-isogeny, the pair of j-invariants $(j(E_1), j(E_2))$ does *not* uniquely determine ϕ, not even up to isomorphism. For example, suppose $\mathrm{End}(E_1) \simeq \mathcal{O}$ and $\mathfrak{p} \neq \bar{\mathfrak{p}}$ is a proper $\mathcal{O}$-ideal of prime norm p such that $[\mathfrak{p}]$ has order 2 in the class group $\mathrm{cl}(\mathcal{O})$. Then $\mathfrak{p}E_1 \simeq \bar{\mathfrak{p}}E_1$, and the isogenies $\phi_{\mathfrak{p}} : E_1 \to \mathfrak{p}E_1$ and $\phi_{\bar{\mathfrak{p}}} : E_1 \to \bar{\mathfrak{p}}E_1$ have distinct kernels but isomorphic images. These isogenies are not isomorphic (there is no automorphism we can compose with one to get the other, their kernels are distinct). In this situation $\Phi_p(j(E_1), Y)$ will have $j(E_2)$ as a double root.

The existence of the dual isogeny implies that $\Phi_N(j_1, j_2) = 0$ if and only if $\Phi_N(j_2, j_1) = 0$. In fact $\Phi_N(X, Y) = \Phi_N(Y, X)$ is symmetric in the variables X and Y.

Theorem 20.7. $\Phi_N(X, Y) = \Phi_N(Y, X)$ *for all* $N > 1$.

Proof. As in the proof of Theorem 20.3, the function $j(N\gamma_0\tau) = j(\tau/N)$ is a root of $\Phi_N(j, Y) \in \mathbb{C}(j)[Y]$ (this is true whether or not N is prime). We also have the identity $\Phi_N(j(\tau), j(N\tau)) = 0$, which implies $\Phi_N(j(\tau/N), j(\tau)) = 0$, so $j(\tau/N)$ is also a root of $\Phi_N(Y, j) \in \mathbb{C}(j)[Y]$. But $\Phi_N(j, Y)$ is irreducible in $\mathbb{C}(j)[Y]$, since it is the minimal polynomial of j_N over $\mathbb{C}(j)$, so $\Phi_N(j, Y)$ must divide $\Phi_N(Y, j)$ in $\mathbb{C}(j)[Y]$ (otherwise their GCD would properly divide $\Phi_N(j, Y)$). It follows from Theorem 20.3 that $\Phi_N(j, Y)$ and $\Phi_N(Y, j)$ have the same degree, since in both cases, for any lattice $L \subseteq \mathbb{C}$, the number of roots of $\Phi_N(j(L), Y)$ and $\Phi_N(Y, j(L))$ when counted with multiplicity is the number of cyclic sublattices of index N in $L \simeq \mathbb{Z} \times \mathbb{Z}$, which is the same for every lattice L.[3] It follows that $\Phi_N(Y, j) = f(j)\Phi_N(j, Y)$ for some $f \in \mathbb{C}(j)$, and plugging in $Y = j$ shows that $f(j) = 1$ ($\Phi_N(j, j) \neq 0$ since $j(\tau)$ is not a root of the minimal polynomial of $j(N\tau)$ for $N > 1$). □

It follows that for prime N the polynomial $\Phi_N(X, Y)$ has degree $N + 1$ in X and Y.

Example 20.8. For $N = 2$ we have

$$\Phi_2(X, Y) = X^3 + Y^3 - X^2Y^2 + 1488(X^2Y + XY^2) - 162000(X^2 + Y^2) + 40773375XY + 8748000000(X + Y) - 157464000000000.$$

[3] Note that, per Remark 20.6, we cannot assume the j-invariants are distinct, but the cyclic sublattices are distinct; some may have the same j-invariant because distinct sublattices may be homothetic.

As can be seen in this example, the integer coefficients of Φ_N are already large when $N = 2$, and they grow rapidly as N increases. For N prime it is known that the logarithm of the absolute value of the largest coefficient of Φ_N is on the order of $6N \log N + O(N)$, see [2], and it has $O(N^2)$ coefficients. Thus the total number of bits required to write down Φ_N is quasi-cubic in N; in practical terms, Φ_{1009} is about 4 GB, and Φ_{10007} is about 5 TB. This makes it quite challenging to compute these polynomials; you will explore an efficient method for doing so on Problem Set 12.

20.2 Modular curves as moduli spaces

In the same way that the j-function defines a bijection from $Y(1) = \mathcal{H}/\Gamma(1)$ to $\mathbb{C}$ (which we may regard as an affine curve in $\mathbb{C}^2$), the functions $j(\tau)$ and $j_N(\tau)$ define a bijection from $Y_0(N) = \mathcal{H}/\Gamma_0(N)$ to the affine curve $\Phi_N(X, Y) = 0$ via the map

$$\tau \mapsto \big(j(\tau), j_N(\tau)\big).$$

If $\{\gamma_k\}$ is a set of right coset representatives for $\Gamma_0(N)$ then for each γ_k we have

$$\gamma_k\tau \mapsto \big(j(\gamma_k\tau), j_N(\gamma_k\tau)\big) = \big(j(\tau), j_N(\gamma_k\tau)\big),$$

and as in the proof of Theorem 20.3, each of these points corresponds to a cyclic N-isogeny $E \to E'$ with $j(E) = j(\tau)$ and $j(E') = j_N(\gamma_k\tau)$. We can thus view the modular curve $Y_0(N)$, equivalently, the non-cuspidal points on $X_0(N)$, as parameterizing cyclic N-isogenies.

As noted above such an isogeny is not always uniquely determined by a pair of j-invariants (these correspond to singular points on the curve $\Phi_N(X, Y) = 0$), but a cyclic N-isogeny $\phi\colon E \to E'$ is uniquely determined by the pair $(E, \langle P\rangle)$, where P is any generator for $\ker\phi$ (so P is a point of order N). Recall from Theorem 5.11 that every finite subgroup of points on an elliptic curve determines a separable isogeny that is unique up to isomorphism. Every pair $(E, \langle P\rangle)$ thus corresponds to a non-cuspidal point of $X_0(N)$; two pairs $(E, \langle P\rangle)$ and $(E', \langle P'\rangle)$ correspond to the same point if and only if there exists an isomorphism $\varphi\colon E \xrightarrow{\sim} E'$ such that $\varphi(\langle P\rangle) = \langle P'\rangle$.

With this interpretation the modular curve $X_0(N)$ can be viewed as the "moduli space" of cyclic N-isogenies of elliptic curves, each identified by a pair $(E, \langle P\rangle)$, up to the isomorphism defined above. We won't formally define the notion of a moduli space in this course, but this can be done, and it provides an alternative definition of $X_0(N)$. The key point from our perspective is that this moduli interpretation is valid over any field, not just $\mathbb{C}$. The modular curves $X_0(N)$ play a key role in many algorithms that work with elliptic curves over finite fields, including the Schoof-Elkies-Atkin (SEA) point-counting algorithm (a faster version of Schoof's algorithm), and fast algorithms to compute Hilbert class polynomials, which are the key to the CM method that we will discuss in the next lecture.

Other modular curves also have characterizations as moduli spaces. We have already seen that the modular curve $X(1)$ is the moduli space of isomorphism classes of elliptic curves, and for $N > 1$ the modular curve $X(N)$ is the moduli space of triples (E, P_1, P_2), where $\{P_1, P_2\}$ is a basis for the N-torsion subgroup of E, and the modular curve $X_1(N)$ is the moduli space of pairs (E, P), where P is a point of order N on E. Note that in each case one considers triples or pairs only up to a suitable isomorphism, as with $X_0(N)$ above.

20.3 The Hilbert class polynomial

We now turn our attention to the Hilbert class polynomial. Recall that for each imaginary quadratic order $\mathcal{O}$, we have the set

$$\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) := \{j(E) \in \mathbb{C} : \mathrm{End}(E) \simeq \mathcal{O}\}$$

of isomorphism classes of elliptic curves with complex multiplication (CM) by $\mathcal{O}$, and the ideal class group $\mathrm{cl}(\mathcal{O})$ acts on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ via isogenies, as we now recall. Every elliptic curve $E/\mathbb{C}$ with CM by $\mathcal{O}$ is of the form $E_{\mathfrak{b}}$ corresponding to the torus $\mathbb{C}/\mathfrak{b}$, where $\mathfrak{b}$ is a proper $\mathcal{O}$-ideal for which $j(\mathfrak{b}) = j(E)$ (note that $j(\mathfrak{b}) = j(E)$ depends only on the class $[\mathfrak{b}]$ in $\mathrm{cl}(\mathcal{O})$). If $[\mathfrak{a}]$ is an element of $\mathrm{cl}(\mathcal{O})$, then $\mathfrak{a}$ acts on $E_{\mathfrak{b}}$ by the isogeny

$$\phi_{\mathfrak{a}} \colon E_{\mathfrak{b}} \to E_{\mathfrak{a}^{-1}\mathfrak{b}}$$

of degree $\mathrm{N}\mathfrak{a}$ induced by the lattice inclusion $\mathfrak{b} \subseteq \mathfrak{a}^{-1}\mathfrak{b}$. As with $E_{\mathfrak{b}}$, the isomorphism class of $E_{\mathfrak{a}^{-1}\mathfrak{b}}$ depends only on the class $[\mathfrak{a}^{-1}\mathfrak{b}]$ in $\mathrm{cl}(\mathcal{O})$, and we proved that this action is free and transitive, meaning that $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is a $\mathrm{cl}(\mathcal{O})$-torsor. This implies that the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is finite, with cardinality equal to the class number $h(\mathcal{O}) := \#\,\mathrm{cl}(\mathcal{O})$.

We may uniquely identify $\mathcal{O}$ by its discriminant D (by Theorem 17.18), and the Hilbert class polynomial

$$H_D(X) = \prod_{j(E) \in \mathrm{Ell}_{\mathcal{O}}(\mathbb{C})} \big(X - j(E)\big)$$

is the monic polynomial whose roots are the distinct j-invariants of all elliptic curves with CM by $\mathcal{O}$. We now want to use the fact that $\Phi_N \in \mathbb{Z}[X,Y]$ to prove that $H_D \in \mathbb{Z}[X]$. To do this we need the following lemma.

Lemma 20.9. *If N is prime then the leading term of $\Phi_N(X,X) \in \mathbb{Z}[X]$ is $-X^{2N}$.*

Proof. Replacing Y with $j(\tau)$ in equation (1) for $\Phi_N(Y)$ yields

$$\Phi_N\big(j(\tau), j(\tau)\big) = \Big(j(\tau) - j(N\tau)\Big) \prod_{k=0}^{N-1} \Big(j(\tau) - j\Big(\frac{\tau+k}{N}\Big)\Big).$$

Recall from the proof of Theorem 19.17 that we have the q-expansions

$$j(N\tau) = \frac{1}{q^N} + \cdots,$$
$$j\Big(\frac{\tau+k}{N}\Big) = \frac{\zeta_N^{-k}}{q^{1/N}} + \cdots,$$

where $q := e^{2\pi i \tau}$, $\zeta_N := e^{2\pi i/N}$, and the ellipsis denote terms involving larger powers of q. Thus

$$j(\tau) - j(N\tau) = -\frac{1}{q^N} + \frac{1}{q} + \cdots,$$
$$j(\tau) - j\Big(\frac{\tau+k}{N}\Big) = \frac{1}{q} - \frac{\zeta_N^{-k}}{q^{1/N}} + \cdots,$$

which implies that the q-expansion of $f(\tau) = \Phi_N\big(j(\tau), j(\tau)\big)$ begins $-\frac{1}{q^{2N}} + \cdots$. Since $f(\tau)$ is a polynomial in $j(\tau) = \frac{1}{q} + \cdots$, the leading term of $\Phi_N(X,X)$ must be $-X^{2N}$. □

Remark 20.10. Lemma 20.9 does not hold in general; in particular, when N is square $\Phi_N(X,X)$ is not even primitive (its coefficients have a non-trivial common divisor).

Before proving $H_D \in \mathbb{Z}[X]$, we record the following classical result, which was proved for maximal orders by Dirichlet and later generalized by Weber; see [3, p. 190]. Today this is typically cited as a consequence of the Chebotarev[4] density theorem, but since the proof of the Chebotarev density theorem actually uses class field theory, a small part of which we are about to prove, we should note that the result we need was proved earlier.

Theorem 20.11. *Let $\mathcal{O}$ be an imaginary quadratic order. Every ideal class in $\mathrm{cl}(\mathcal{O})$ contains infinitely many ideals of prime norm.*

Proof. This follows from Theorems 7.7 and 9.12 in [3]. □

Theorem 20.12. *The coefficients of the Hilbert class polynomial $H_D(X)$ are integers.*

Proof. Let $\mathcal{O}$ be the imaginary quadratic order of discriminant D, let $E/\mathbb{C}$ be an elliptic curve with CM by $\mathcal{O}$, and let $\mathfrak{p}$ be a principal $\mathcal{O}$-ideal of prime norm p (by Theorem 20.11 there are infinitely many choices for $\mathfrak{p}$). Then $[\mathfrak{p}]$ is the identity element of $\mathrm{cl}(\mathcal{O})$, so $\mathfrak{p}$ acts trivially on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$. Thus $\mathfrak{p}E \simeq E$, which implies that, after composing with an isomorphism if necessary, we have a p-isogeny from E to itself, equivalently, an endomorphism of degree p. Such an isogeny is necessarily cyclic, since it has prime degree, so we must have $\Phi_p\big(j(E),j(E)\big) = 0$. Thus $j(E)$ is the root of the polynomial $-\Phi_p(X,X)$, which is monic, by Lemma 20.9, and has integer coefficients, by Theorem 19.17. The j-invariant $j(E)$ is thus an algebraic integer, and the elliptic curve E can be defined by a Weierstrass equation $y^2 = x^3 + Ax + B$ whose coefficients lie in the number field $\mathbb{Q}(j(E))$, by Theorem 13.12.

The absolute Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ acts on the set of elliptic curves defined over number fields via its action on the Weierstrass coefficients A and B: for each field automorphism $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ the curve E^σ is defined by the equation $y^2 = x^3 + \sigma(A)x + \sigma(B)$. Similarly, σ acts on isogenies via its action on the coefficients of the rational map defining the isogeny. If $\phi\colon E \to E$ is an endomorphism, then so is $\phi^\sigma\colon E^\sigma \to E^\sigma$, and for any $\phi,\psi \in \mathrm{End}(E)$ we have $(\phi+\psi)^\sigma = \phi^\sigma + \psi^\sigma$ and $(\phi\circ\psi)^\sigma = \phi^\sigma \circ \psi^\sigma$. Thus each $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ induces a ring homomorphism

$$\mathrm{End}(E) \xrightarrow{\sigma} \mathrm{End}(E^\sigma).$$

Applying σ^{-1} to E^σ induces an inverse homomorphism, we thus have a ring isomorphism $\mathrm{End}(E) \simeq \mathrm{End}(E^\sigma)$, which implies that E^σ also has CM by $\mathcal{O}$.

The j-invariant of E is a rational function $1728\cdot 4A^3/(4A^3+27B^2)$ of A and B, so $j(E^\sigma) = j(E)^\sigma$, and we have shown that $j(E^\sigma) \in \mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$. It follows that $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ acts on the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$, which are the roots of $H_D(X)$. The coefficients of $H_D(X)$ are symmetric polynomials in its roots, hence they are fixed by $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and lie in the fixed field $\mathbb{Q}$; moreover, they are algebraic integers (since the roots are), so they lie in $\overline{\mathbb{Z}}\cap\mathbb{Q} = \mathbb{Z}$. □

Corollary 20.13. *Let $E/\mathbb{C}$ be an elliptic curve with complex multiplication. Then $j(E)$ is an algebraic integer.*

From the proof of Theorem 20.12, we now have two groups acting on the roots of $H_D(X)$: the class group $\mathrm{cl}(\mathcal{O})$ and the Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. In the latter case there is no need

[4]Many different transliterations of Chebotarev's Russian name appear in the literature, including Chebotaryov, Čebotarev, Chebotarëv, Čhebotarëv, Tchebotarev, and Tschebotarëw; none is universally accepted.

to consider the entire Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, we can always restrict our attention to any Galois extension of $\mathbb{Q}$ that contains the splitting field L of $H_D(X)$, since the action of any $\sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on the roots of $H_D(X)$ is determined by its restriction to $\mathrm{Gal}(L/\mathbb{Q})$. We then have two finite group actions, and it is reasonable to ask whether they are in some sense compatible.

In order to obtain compatible actions we do not want to work with the splitting field L of $H_D(X)$ over $\mathbb{Q}$, since $\mathrm{Gal}(L/\mathbb{Q})$, may contain automorphisms that don't fix the order $\mathcal{O}$. But if we instead let L be the splitting field of $H_D(X)$ over $K := \mathbb{Q}(\sqrt{D})$, the Galois group $\mathrm{Gal}(L/K)$ fixes $\mathcal{O}$, and we will show that its action on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is compatible with that of the class group $\mathrm{cl}(\mathcal{O})$. In fact, $\mathrm{Gal}(L/K) \simeq \mathrm{cl}(\mathcal{O})$. This isomorphism is part of the *First Main Theorem of Complex Multiplication*, and our next goal is to prove it.

So let $\mathcal{O}$ be the imaginary quadratic order of discriminant D, and let us fix an elliptic curve E_1 with CM by $\mathcal{O}$. Each $\sigma \in \mathrm{Gal}(L/K)$ can be viewed as the restriction to L of an element of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ that fixes K, thus as in the proof of Theorem 20.12, the elliptic curve E_1^σ also has CM by $\mathcal{O}$. Therefore $E_1^\sigma \simeq \mathfrak{a}E_1$ for some proper $\mathcal{O}$-ideal $\mathfrak{a}$, since $\mathrm{cl}(\mathcal{O})$ acts transitively on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$. If $E_2 \simeq \mathfrak{b}E_1$ is any other elliptic curve with CM by $\mathcal{O}$, we then have

$$E_2^\sigma \simeq (\mathfrak{b}E_1)^\sigma = \mathfrak{b}^\sigma E_1^\sigma = \mathfrak{b}E_1^\sigma \simeq \mathfrak{b}\mathfrak{a}E_1 = \mathfrak{a}\mathfrak{b}E_1 \simeq \mathfrak{a}E_2. \tag{2}$$

The innocent looking identity $(\mathfrak{b}E_1)^\sigma = \mathfrak{b}^\sigma E_1^\sigma$ used in (2) is not immediate, it requires a somewhat lengthy argument involving a diagram chase that we omit; see [7, Prop. II.2.5] for a proof. The second identity *is* immediate, because $\mathfrak{b} \subset K$ and $\sigma \in \mathrm{Gal}(L/K)$ fixes K; but note that this would not be true if we had instead used $\sigma \in \mathrm{Gal}(L/\mathbb{Q})$.

Since our choice of E_2 was arbitrary, it follows from (2) that the action of σ on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is the same as the action of $\mathfrak{a}$ on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$. Because $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is a $\mathrm{cl}(\mathcal{O})$-torsor, the map that sends each $\sigma \in \mathrm{Gal}(L/K)$ to the unique class $[\mathfrak{a}] \in \mathrm{cl}(\mathcal{O})$ for which $E_1^\sigma = \mathfrak{a}E_1$ defines a group homomorphism

$$\Psi\colon \mathrm{Gal}(L/K) \to \mathrm{cl}(\mathcal{O}).$$

This homomorphism is injective because, by definition of the splitting field, the only element of $\mathrm{Gal}(L/K)$ that acts trivially on the roots of $H_D(X)$ is the identity element, and the same is true of $\mathrm{cl}(\mathcal{O})$. We summarize this discussion with the following theorem.

Theorem 20.14. *Let $\mathcal{O}$ be an imaginary quadratic order of discriminant D and let L be the splitting field of $H_D(X)$ over $K := \mathbb{Q}(\sqrt{D})$. The map $\Psi : \mathrm{Gal}(L/K) \to \mathrm{cl}(\mathcal{O})$ that sends each $\sigma \in \mathrm{Gal}(L/K)$ to the unique $\alpha_\sigma \in \mathrm{cl}(\mathcal{O})$ for which $j(E)^\sigma = \alpha_\sigma j(E)$ for all $j(E) \in \mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is an injective group homomorphism.*

We thus have an embedding of $\mathrm{Gal}(L/K)$ in $\mathrm{cl}(\mathcal{O})$ that is compatible with the actions of both groups on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$. It remains only to prove that Ψ is surjective, which is equivalent to proving that $H_D(X)$ is irreducible over K.

References

[1] Michael Artin, *Algebra*, second edition, Pearson, 2011.

[2] Paula Cohen, *On the coefficients of the transformation polynomials for the elliptic modular function*, Mathematical Proceedings of the Cambridge Philosophical Society **95** (1984), 389–402.

[3] David A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory, and complex multiplication*, second edition, Wiley, 2013.

[4] Jun-Ichi Igusa, *Kroneckerian Model of Fields of Elliptic Modular Functions*, American Journal of Mathematics **81** (1959), 561–577.

[5] J. S. Milne, *Elliptic curves*, BookSurge Publishers, 2006.

[6] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer, 2009.

[7] Joseph H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Springer, 1994.

[8] Lawrence C. Washington, *Elliptic curves: number theory and cryptography*, second edition, Chapman & Hall/CRC, 2008.

21 Ring class fields and the CM method

Let $\mathcal{O}$ be an imaginary quadratic order with discriminant D, and let

$$\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) := \{j(E) \in \mathbb{C} : \mathrm{End}(E) = \mathcal{O}\}.$$

In the previous lecture we proved that the Hilbert class polynomial

$$H_D(X) := H_{\mathcal{O}}(X) := \prod_{j(E)\in \mathrm{Ell}_{\mathcal{O}}(\mathbb{C})} \Big(X - j(E)\Big)$$

has integer coefficients. We then defined L to be the splitting field of $H_D(X)$ over the field $K = \mathbb{Q}(\sqrt{D})$, and showed that there is an injective group homomorphism

$$\Psi\colon \mathrm{Gal}(L/K) \hookrightarrow \mathrm{cl}(\mathcal{O})$$

that commutes with the group actions of $\mathrm{Gal}(L/K)$ and $\mathrm{cl}(\mathcal{O})$ on the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) = \mathrm{Ell}_{\mathcal{O}}(L)$ of roots of $H_D(X)$. To complete the proof of the First Main Theorem of Complex Multiplication, which asserts that Ψ is an isomorphism, we just need to show that Ψ is surjective, equivalently, that $H_D(X)$ is irreducible over K.

To do this we need to introduce the Artin map (named after Emil Artin), which allows us to associate to each $\mathcal{O}$-ideal $\mathfrak{p}$ of prime norm satisfying certain constraints an automorphism $\sigma_{\mathfrak{p}} \in \mathrm{Gal}(L/K)$ whose action on $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ corresponds to the action of $[\mathfrak{p}]$. In order to define the Artin map we need to briefly delve into a bit of algebraic number theory. We will restrict our attention to the absolute minimum that we need. Those who would like to know more may wish to consult one of [7, 8] or these 18.785 lecture notes; those who do not may treat the Artin map as a black box.

21.1 The Artin map

Let L be a finite Galois extension of a number field K. Nonzero prime ideals $\mathfrak{p}$ of the ring of integers $\mathcal{O}_K$ are called "primes of K".[1] The $\mathcal{O}_L$-ideal $\mathfrak{p}\mathcal{O}_L$ is typically not a prime ideal, but it can be uniquely factored as

$$\mathfrak{p}\mathcal{O}_L = \mathfrak{q}_1 \cdots \mathfrak{q}_n$$

where the $\mathfrak{q}_i$ are not-necessarily-distinct primes of L (prime ideals of $\mathcal{O}_L$) that are characterized by the property $\mathfrak{q}_i \cap \mathcal{O}_K = \mathfrak{p}$. The primes $\mathfrak{q}_i$ are said to "lie above" the prime $\mathfrak{p}$, and it is standard to write $\mathfrak{q}_i|\mathfrak{p}$ as shorthand for $\mathfrak{q}_i|\mathfrak{p}\mathcal{O}_L$ and use $\{\mathfrak{q}|\mathfrak{p}\}$ to denote the set $\{\mathfrak{q}_1, \ldots, \mathfrak{q}_n\}$.

We should note that the ring $\mathcal{O}_L$ is typically *not* a unique factorization domain, but it is a *Dedekind domain*, and this implies unique factorization of ideals.[2]

When the $\mathfrak{q}_i$ are distinct, we say that $\mathfrak{p}$ is *unramified* in L, which is true for all but finitely many primes $\mathfrak{p}$. If we apply an automorphism $\sigma \in \mathrm{Gal}(L/K)$ to both sides of the equation above, the LHS must remain the same: σ fixes every element of $\mathfrak{p} \subseteq K$, and it maps algebraic integers to algebraic integers, so it preserves the set $\mathcal{O}_L$. For the RHS, it is

[1] This is an abuse of terminology: as a ring, K does not have any nonzero prime ideals (it is a field).

[2] There are several equivalent definitions of Dedekind domains: it is an integral domain with unique factorization of ideals, and it is also an integral domain in which every nonzero fractional ideal is invertible. We have seen that the latter applies to rings of integers in number fields (at least for imaginary quadratic fields), so the former must as well (this equivalence is a standard result from commutative algebra).

clear that σ must map $\mathcal{O}_L$-ideals to $\mathcal{O}_L$-ideals, and since the $\mathfrak{q}_i$ are all prime ideals, σ must permute them. Thus the Galois group $\mathrm{Gal}(L/K)$ acts on the set $\{\mathfrak{q}_1,\dots,\mathfrak{q}_n\} = \{\mathfrak{q}|\mathfrak{p}\}$; one can show that this action is transitive, but it is typically not faithful.

For each $\mathfrak{q}|\mathfrak{p}$, the stabilizer of $\mathfrak{q}$ under this action is a subgroup

$$D_\mathfrak{q} := \{\sigma \in \mathrm{Gal}(L/K) : \mathfrak{q}^\sigma = \mathfrak{q}\} \subseteq \mathrm{Gal}(L/K)$$

known as the *decomposition group* of $\mathfrak{q}$. Each $\sigma \in D_\mathfrak{q}$ fixes $\mathfrak{q}$ and therefore induces an automorphism $\bar\sigma$ of the quotient $\mathbb{F}_\mathfrak{q} := \mathcal{O}_L/\mathfrak{q}$ defined by $\bar\sigma(\bar x) = \overline{\sigma(x)}$, where $x \mapsto \bar x$ is the quotient map $\mathcal{O}_L \to \mathcal{O}_L/\mathfrak{q}$. The quotient $\mathcal{O}_L/\mathfrak{q}$ is a field (in a Dedekind domain every nonzero prime ideal is maximal), and $\mathfrak{q}$ has finite index $\mathrm{N}\mathfrak{q} := [\mathcal{O}_L : \mathfrak{q}]$ in $\mathcal{O}_L$, so it is a finite field of cardinality $\mathrm{N}\mathfrak{q}$ (which must be a prime power). The image of $\mathcal{O}_K$ under the quotient map $\mathcal{O}_L \to \mathcal{O}_L/\mathfrak{q} = \mathbb{F}_\mathfrak{q}$ is $\mathcal{O}_K/(\mathfrak{q}\cap\mathcal{O}_K) = \mathcal{O}_K/\mathfrak{p} = \mathbb{F}_\mathfrak{p}$, thus the finite field $\mathbb{F}_\mathfrak{p}$ is a subfield of $\mathbb{F}_\mathfrak{q}$ (and necessarily has the same characteristic). It follows that $\bar\sigma \in \mathrm{Gal}(\mathbb{F}_\mathfrak{q}/\mathbb{F}_\mathfrak{p})$, and we have a group homomorphism

$$\begin{aligned} D_\mathfrak{q} &\to \mathrm{Gal}(\mathbb{F}_\mathfrak{q}/\mathbb{F}_\mathfrak{p}) \\ \sigma &\mapsto \bar\sigma. \end{aligned}$$

This homomorphism is surjective [8, Prop. I.9.4], and when $\mathfrak{p}$ is unramified it is also injective [8, Prop. I.9.5], and therefore an isomorphism, which we now assume.

The group $\mathrm{Gal}(\mathbb{F}_\mathfrak{q}/\mathbb{F}_\mathfrak{p})$ is cyclic, generated by the Frobenius automorphism $x \to x^{\mathrm{N}\mathfrak{p}}$, where $\mathrm{N}\mathfrak{p} = [\mathcal{O}_K : \mathfrak{p}] = \#\mathbb{F}_\mathfrak{p}$. The unique $\sigma_\mathfrak{q} \in D_\mathfrak{q}$ for which $\bar\sigma_\mathfrak{q}$ is the Frobenius automorphism is called the *Frobenius element* of $\mathrm{Gal}(L/K)$ at $\mathfrak{q}$. In general the Frobenius element $\sigma_\mathfrak{q}$ depends on our choice of $\mathfrak{q}$, but the $\sigma_\mathfrak{q}$ for $\mathfrak{q}|\mathfrak{p}$ are all conjugate, since if $\tau(\mathfrak{q}_i) = \mathfrak{q}_j$ then we must have $\sigma_{\mathfrak{q}_j} = \tau^{-1}\sigma_{\mathfrak{q}_i}\tau$. This implies that the $\bar\sigma_\mathfrak{q}$ all have the same order, hence the extensions $\mathbb{F}_\mathfrak{q}/\mathbb{F}_\mathfrak{p}$ all have the same degree and are thus isomorphic.

In the case we are interested in, $\mathrm{Gal}(L/K) \hookrightarrow \mathrm{cl}(\mathcal{O})$ is abelian, so conjugacy implies equality, and the $\sigma_\mathfrak{q}$ are all the same. Thus when $\mathrm{Gal}(L/K)$ is abelian, each prime $\mathfrak{p}$ of K determines a unique Frobenius element that we denote $\sigma_\mathfrak{p}$. The map

$$\mathfrak{p} \mapsto \sigma_\mathfrak{p}$$

is known as the *Artin map* (it extends multiplicatively to all $\mathcal{O}_K$-ideals that are products of unramified prime ideals, but this is not relevant to us). The automorphism $\sigma_\mathfrak{p}$ is uniquely characterized by the fact that

$$\sigma_\mathfrak{p}(x) \equiv x^{\mathrm{N}\mathfrak{p}} \bmod \mathfrak{q}, \tag{1}$$

for all $x \in \mathcal{O}_L$ and primes $\mathfrak{q}|\mathfrak{p}$.

If $E/\mathbb{C}$ has CM by $\mathcal{O}$ then $j(E) \in L$, and this implies that (up to isomorphism) E can be defined by a Weierstrass equation $y^2 = x^3 + Ax + B$ with $A, B \in \mathcal{O}_L$. Indeed, as in the proof of Theorem 13.12, for $j(E) \neq 0, 1728$ we can take $A = 3j(E)(1728 - j(E))$ and $B = 2j(E)(1728 - j(E))^2$.

For each prime $\mathfrak{q}$ of L, so long as the discriminant $\Delta(E) := -16(4A^3 + 27B^2)$ does not lie in $\mathfrak{q}$, equivalently, the image of $\Delta(E)$ under the quotient map $\mathcal{O}_L \to \mathcal{O}_L/\mathfrak{q} = \mathbb{F}_q$ is nonzero, reducing modulo $\mathfrak{q}$ yields an elliptic curve $\bar E/\mathbb{F}_\mathfrak{q}$ defined by $y^2 = x^3 + \bar A x + \bar B$. We then say that E has good reduction modulo $\mathfrak{q}$. This holds for all but finitely many primes $\mathfrak{q}$ of L, since the principal ideal $(\Delta(E))$ is divisible by only finitely many prime ideals.

21.2 The First Main Theorem of Complex Multiplication

With the Artin map in hand, we can now complete our proof of the First Main Theorem of Complex Multiplication.

Theorem 21.1. *Let $\mathcal{O}$ be an imaginary quadratic order of discriminant D and let L be the splitting field of $H_D(X)$ over $K := \mathbb{Q}(\sqrt{D})$. The map $\Psi\colon \mathrm{Gal}(L/K) \to \mathrm{cl}(\mathcal{O})$ that sends each $\sigma \in \mathrm{Gal}(L/K)$ to the unique $\alpha_\sigma \in \mathrm{cl}(\mathcal{O})$ such that $j(E)^\sigma = \alpha_\sigma j(E)$ for all $j(E) \in \mathrm{Ell}_{\mathcal{O}}(L)$ is a group isomorphism compatible with the actions of $\mathrm{Gal}(L/K)$ and $\mathrm{cl}(\mathcal{O})$ on $\mathrm{Ell}_{\mathcal{O}}(L)$.*

Proof. In the previous lecture we showed that Ψ is well-defined, injective, and commutes with the group actions of $\mathrm{Gal}(L/K)$ and $\mathrm{cl}(\mathcal{O})$; see Theorem 20.14 and the discussion preceding it. It remains only to show that Ψ is surjective.

Fix $\alpha \in \mathrm{cl}(\mathcal{O})$, and let $\mathfrak{p}$ be a prime of K such that the following hold:

(i) $\mathfrak{p} \cap \mathcal{O}$ is a proper $\mathcal{O}$-ideal of prime norm p such that $[\mathfrak{p}] = \alpha$;

(ii) p is unramified in K and $\mathfrak{p}$ is unramified in L;

(iii) Each $j(E) \in \mathrm{Ell}_{\mathcal{O}}(L)$ is the j-invariant of an elliptic curve E/L that has good reduction modulo every prime $\mathfrak{q}|\mathfrak{p}$ (prime ideals $\mathfrak{q}$ of $\mathcal{O}_L$ dividing $\mathfrak{p}\mathcal{O}_L$).

(iv) The $j(E) \in \mathrm{Ell}_{\mathcal{O}}(L)$ are distinct modulo every prime $\mathfrak{q}|\mathfrak{p}$.

By Theorem 20.11, there are infinitely many $\mathfrak{p}$ for which (i) holds, and conditions (ii)-(iv) prohibit only finitely many primes, so such a $\mathfrak{p}$ exists. To ease the notation, we will also use $\mathfrak{p}$ to denote the $\mathcal{O}$-ideal $\mathfrak{p} \cap \mathcal{O}$; it will be clear from context whether we are viewing $\mathfrak{p}$ as an $\mathcal{O}_K$-ideal or as an $\mathcal{O}$-ideal (in particular, anytime we write $[\mathfrak{p}]$ we must mean $[\mathfrak{p} \cap \mathcal{O}]$, since we are using $[\,\cdot\,]$ to denote equivalence classes of $\mathcal{O}$-ideals).

Let us now consider a particular prime $\mathfrak{q}|\mathfrak{p}$ and curve E/L with CM by $\mathcal{O}$ that has good reduction modulo $\mathfrak{q}$, defined by $E\colon y^2 = x^3 + Ax + B$ with $A, B \in \mathcal{O}_L$ and $\mathfrak{q} \nmid \Delta(E)$. Put $\mathbb{F}_\mathfrak{q} := \mathcal{O}_L/\mathfrak{q}$, and let $\overline{E}/\mathbb{F}_\mathfrak{q}$ be the reduction of E modulo $\mathfrak{q}$, defined by $\overline{E}\colon y^2 = x^3 + \bar{A}x + \bar{B}$. The Frobenius element $\sigma_\mathfrak{p}$ induces the p-power Frobenius automorphism $\overline{\sigma}_\mathfrak{p} \in \mathrm{Gal}(\mathbb{F}_\mathfrak{q}/\mathbb{F}_p)$, since $\mathrm{N}\mathfrak{p} = p$, and we have a corresponding isogeny

$$\pi\colon \overline{E} \to \overline{E^{\sigma_\mathfrak{p}}} = \overline{E}^{\bar{\sigma}_p} = \overline{E}^{(p)}$$

defined by $(x, y) \mapsto (x^p, y^p)$, where $\overline{E}^p$ is the curve $y^2 = x^3 + \bar{A}^p x + \bar{B}^p$. The isogeny π is purely inseparable of degree p.

The CM action of the proper $\mathcal{O}$-ideal $\mathfrak{p} \cap \mathcal{O}$ corresponds to an isogeny $\phi_\mathfrak{p}\colon E \to \mathfrak{p}E$ of degree $\mathrm{N}\mathfrak{p} = p$, with $\mathfrak{p}E$ of good reduction modulo $\mathfrak{q}$, by (iii), which we can assume is defined by a rational map $(\frac{u(x)}{v(x)}, \frac{s(x)}{t(x)}y)$ where $u, v, s, t \in \mathcal{O}_L[x]$, with u monic and v nonzero modulo $\mathfrak{q}$. The isogeny $\phi\colon \overline{E} \to \overline{\mathfrak{p}E}$ obtained by reducing the coefficients of u, v, s, t modulo $\mathfrak{q}$ has the same degree p as the isogeny π (we can assume $\deg v < \deg u$ and u is monic so its degree doesn't change when it is reduced). The composition of ϕ with its dual $\hat{\phi}$ is the multiplication-by-p map on $\overline{E}$, which is inseparable since $\mathbb{F}_\mathfrak{q}$ has characteristic p. This implies that at least one of ϕ and $\hat{\phi}$ is inseparable. Without loss of generality we may assume ϕ is inseparable: if not, we can replace E by $\mathfrak{p}E$ and $\mathfrak{p}$ by its complex conjugate $\bar{\mathfrak{p}}$, which also satisfies (i)-(iv) and induces the dual isogeny $\hat{\phi}_\mathfrak{p}\colon \mathfrak{p}E \to E$ (up to an isomorphism), since the ideal $\bar{\mathfrak{p}}\mathfrak{p} = (\mathrm{N}\mathfrak{p}) = (p)$ induces the multiplication-by-p map on E, and reducing the rational maps defining $\hat{\phi}_\mathfrak{p}$ yields the dual isogeny $\hat{\phi}\colon \overline{\mathfrak{p}E} \to \overline{E}$.

By Corollary 5.4, we can decompose the inseparable isogeny ϕ of degree p as $\phi = \phi_{\text{sep}} \circ \pi$, where ϕ_{sep} has degree 1 and must be an isomorphism. Thus $\overline{\mathfrak{p}E} \simeq \overline{E^{\sigma_\mathfrak{p}}}$ and therefore $j(\overline{\mathfrak{p}E}) = j(\overline{E^{\sigma_\mathfrak{p}}})$, and (iv) implies $j(\mathfrak{p}E) = j(E^{\sigma_\mathfrak{p}})$. It follows that $\Psi(\sigma_\mathfrak{p}) = [\mathfrak{p}] = \alpha$, since each element of $\mathrm{cl}(\mathcal{O})$ is determined by its action on any element of the $\mathrm{cl}(\mathcal{O})$-torsor $\mathrm{Ell}_\mathcal{O}(L)$. □

Corollary 21.2. *Let $\mathcal{O}$ be an imaginary quadratic order with discriminant D. The Hilbert class polynomial $H_D(x)$ is irreducible over $K = \mathbb{Q}(\sqrt{D})$ and for any elliptic curve $E/\mathbb{C}$ with CM by $\mathcal{O}$ the field $K(j(E))$ is a finite abelian extension of K with $\mathrm{Gal}(K(j(E))/K) \simeq \mathrm{cl}(\mathcal{O})$.*

Proof. Let L be the splitting field of $H_D(X)$ over K. The class group $\mathrm{cl}(\mathcal{O})$ acts transitively on the roots of $H_D(X)$ (the set $\mathrm{Ell}_\mathcal{O}(\mathbb{C})$), hence by Theorem 21.1, the Galois group $\mathrm{Gal}(L/K)$ also acts transitively on the roots of $H_D(X)$, which implies that $H_D(X)$ is irreducible over K and is the minimal polynomial of each of its roots. The degree of H_D is equal to the class number $h(D) = \#\,\mathrm{cl}(\mathcal{O}) = \#\mathrm{Gal}(L/K) = [L : K]$, so we have $L = K(j(E))$ for every root $j(E)$ of $H_D(X)$, equivalently, every $j(E) \in \mathrm{Ell}_\mathcal{O}(\mathbb{C}) = \{j(E) : \mathrm{End}(E) = \mathcal{O}\}$. We have $\mathrm{Gal}(L/K) \simeq \mathrm{cl}(\mathcal{O})$ by Theorem 21.1, which is abelian. □

21.3 The ring class field of an imaginary quadratic order

Definition 21.3. Let $\mathcal{O}$ be an imaginary quadratic order with discriminant D. The splitting field of the Hilbert class polynomial $H_D(X)$ over $K = \mathbb{Q}(\sqrt{D})$, equivalently, the extension of K generated by the j-invariant of any elliptic curve $E/\mathbb{C}$ with CM by $\mathcal{O}$, is known as the *ring class field* of the imaginary quadratic order $\mathcal{O}$ with discriminant D.

We say that an integer prime p is *unramified* in a number field L if the ideal $p\mathcal{O}_L$ factors into distinct prime ideals $\mathfrak{q}$ in $\mathcal{O}_L$, and we say that p *splits completely* in L if the prime ideals $\mathfrak{q}|p$ are distinct and have minimal norm $\mathrm{N}\mathfrak{q} = p$.

For an imaginary quadratic field K of discriminant D there are three possibilities for the factorization of the ideal $p\mathcal{O}_K$ in $\mathcal{O}_K$: it either *splits* (completely into two distinct prime ideals), *ramifies* (is the square of a prime ideal), or remains *inert* (the ideal $p\mathcal{O}_K$ is already prime). These are distinguished by the *Kronecker symbol* $\left(\frac{D}{p}\right)$, which is 1, 0, -1, respectively, in these three cases (as proved in Lemma 21.8 below).

Definition 21.4. Let p be a prime and D an integer. For $p > 2$ the *Kronecker symbol* is

$$\left(\frac{D}{p}\right) := \#\{x \in \mathbb{F}_p : x^2 = D\} - 1.$$

For $p = 2$, we define $\left(\frac{D}{p}\right)$ to be 1 for $D \equiv \pm 1 \bmod 8$, zero if $p|D$, and -1 for $D \equiv \pm 3 \bmod 8$.

Theorem 21.5. *Let $\mathcal{O}$ be an imaginary quadratic order with discriminant D and ring class field L. Let $p \nmid D$ be an odd prime unramified in L.*[3] *The following are equivalent:*

(i) *p is the norm of a principal $\mathcal{O}$-ideal;*

(ii) *$\left(\frac{D}{p}\right) = 1$ and $H_D(X)$ splits into linear factors in $\mathbb{F}_p[X]$;*

(iii) *p splits completely in L;*

(iv) *$4p = t^2 - v^2 D$ for some integers t and v with $t \not\equiv 0 \bmod p$.*

[3] If p does not divide D then it must be unramified in L, but we have not proved this yet, so we include it as a hypothesis which will be removed in Corollary 21.10.

Proof. Let $K := \mathbb{Q}(\sqrt{D})$, let $\mathcal{O}_K = [1, \omega]$ be the ring of integers of K. By Theorem 17.18, we may write $D = u^2 D_K$, where $u = [\mathcal{O}_K : \mathcal{O}]$ and $D_K = \operatorname{disc} \mathcal{O}_K$ is a fundamental discriminant, and we then have $\mathcal{O} = [1, u\omega]$.

(i)⇒(iv): Let (λ) be a principal $\mathcal{O}$-ideal of norm p. Then $[1, \lambda]$ is a suborder of $\mathcal{O}$ with discriminant $v^2 u^2 D_K = v^2 D$, where $v = \big[\mathcal{O} : [1, \lambda]\big]$. Let $t := \lambda + \bar{\lambda}$ so that $x^2 - tx + p$ is the minimal polynomial of λ, with discriminant $\operatorname{disc}[1, \lambda] = t^2 - 4p = v^2 D$. Then (iv) holds with $t \not\equiv 0 \bmod p$ because $p \nmid D$ (if $p | t$ then $p | v$ and $p^2 | 4p$, a contradiction for $p \neq 2$).

(iv)⇒(i): If $4p = t^2 - v^2 D$ then the polynomial $x^2 - tx + p$ with discriminant $v^2 D$ has a root $\lambda \in \mathcal{O}_K$; the order $[1, \lambda]$ has discriminant $v^2 D$ and therefore lies in $\mathcal{O}$, by Theorem 17.18, so $\lambda \in \mathcal{O}$, and (λ) is a principal $\mathcal{O}$-ideal of norm $\lambda\bar{\lambda} = p$.

(i)⇒(ii): Since (i)⇒(iv) we have $4p = t^2 - v^2 D$ for some $t, v \in \mathbb{Z}$ with $t \not\equiv 0 \bmod p$, and

$$\left(\frac{D}{p}\right) = \left(\frac{v^2 D}{p}\right) = \left(\frac{t^2 - 4p}{p}\right) = 1,$$

since $t^2 \not\equiv 0 \bmod p$. If $\mathfrak{p}$ is a principal $\mathcal{O}$-ideal of norm p, then $\mathfrak{p}$ is unramified in L (since $p = \mathfrak{p}\bar{\mathfrak{p}}$ is unramified in L), and $\mathfrak{p}$ is principal, so $[\mathfrak{p}]$ and therefore $\sigma_{\mathfrak{p}}$ acts trivially on the roots of $H_D(X)$, by Theorem 21.1. The roots of $H_D(X) \bmod p$ must therefore lie in $\mathbb{F}_{\mathfrak{p}} = \mathbb{F}_p$ and $H_D(X)$ splits into linear factors in $\mathbb{F}_p[X]$.

(ii)⇒(iii): If $\left(\frac{D}{p}\right) = 1$, then $p\mathcal{O}_K = \mathfrak{p}\bar{\mathfrak{p}}$ splits into distinct primes of norm p in K, by Lemma 21.8, and if $H_D(X)$ splits into linear factors in $\mathbb{F}_p[x]$, then its roots are all fixed by $\sigma_{\mathfrak{p}}$. This implies $[\mathbb{F}_{\mathfrak{q}} : \mathbb{F}_{\mathfrak{p}}] = 1$, and therefore $\mathrm{N}\mathfrak{q} = [\mathcal{O}_L : \mathfrak{q}] = [\mathcal{O}_K : \mathfrak{p}] = p$ for every prime $\mathfrak{q} | \mathfrak{p}$, so $\mathfrak{p}$ splits completely in L (it must be unramified, since p is). If $\mathfrak{p}\mathcal{O}_L = \mathfrak{q}_1 \cdots \mathfrak{q}_n$, then $\bar{\mathfrak{p}}\mathcal{O}_L = \bar{\mathfrak{q}}_1 \cdots \bar{\mathfrak{q}}_n$ (note that $\bar{\mathcal{O}}_L = \mathcal{O}_L$), and $p\mathcal{O}_L = \mathfrak{p}\bar{\mathfrak{p}}\mathcal{O}_L = \mathfrak{q}_1 \cdots \mathfrak{q}_n \bar{\mathfrak{q}}_1 \cdots \bar{\mathfrak{q}}_n$ splits completely in L (the $\mathfrak{q}_i$ and $\bar{\mathfrak{q}}_i$ must all be distinct since p is unramified in L).

(iii)⇒(i): If $p\mathcal{O}_L = \mathfrak{q}_1 \cdots \mathfrak{q}_n$ with the $\mathrm{N}\mathfrak{q}_1 = \cdots \mathrm{N}\mathfrak{q}_n = p$ then $\mathbb{F}_{\mathfrak{q}} := \mathcal{O}_L/\mathfrak{q} = \mathbb{F}_p$ for all primes $\mathfrak{q}$ dividing $p\mathcal{O}_L$. If $\mathfrak{p}$ is a prime of K dividing $p\mathcal{O}_K$, then $\mathfrak{p}\mathcal{O}_L$ divides $p\mathcal{O}_L$ and must be divisible by some prime ideal $\mathfrak{q}$ dividing $p\mathcal{O}_L$. The inclusions $p\mathbb{Z} \subseteq \mathfrak{p} \subseteq \mathfrak{q}$ imply the inclusions $\mathbb{F}_p \subseteq \mathbb{F}_{\mathfrak{p}} \subseteq \mathbb{F}_{\mathfrak{q}} = \mathbb{F}_p$, where $\mathbb{F}_{\mathfrak{p}} := \mathcal{O}_K/\mathfrak{p}$, so $\mathbb{F}_{\mathfrak{p}} = \mathbb{F}_p$, and $\mathfrak{p}$ has norm p. The extension $\mathbb{F}_{\mathfrak{q}}/\mathbb{F}_{\mathfrak{p}}$ is trivial, so the Frobenius element $\sigma_{\mathfrak{p}} \in \mathrm{Gal}(L/K)$ is the identity, and so is $[\mathfrak{p} \cap O] \in \mathrm{cl}(\mathcal{O})$, by Theorem 21.1 (note: $\mathfrak{p} \cap \mathcal{O}$ is a proper $\mathcal{O}$-ideal because $\mathrm{N}\mathfrak{p} = p$ does not divide $D = u^2 D_K$). Thus $\mathfrak{p} \cap \mathcal{O}$ is a principal $\mathcal{O}$-ideal of norm $[\mathcal{O} : \mathfrak{p} \cap \mathcal{O}] = [\mathcal{O}_K : \mathfrak{p}] = p$. □

Corollary 21.6. *Let $\mathcal{O} \subseteq \mathcal{O}'$ be imaginary quadratic orders and let L and L' be their respective ring class fields. Then $L' \subseteq L$.*

Proof. We first note that $L' \subseteq L$ if and only if every prime p that splits completely in L also splits completely in L'. This follows from Theorem 21.18 in the lecture notes for 18.785 (note that $L/\mathbb{Q}$ is a Galois extension of $\mathbb{Q}$ because it is the compositum of the splitting field of $H_D(X)$ over $\mathbb{Q}$ and the quadratic field $K = \mathbb{Q}(\sqrt{D})$, both of which are Galois, and similarly for $L'/\mathbb{Q}$). Let D and D' be the discriminants of $\mathcal{O}$ and $\mathcal{O}'$, respectively. Then $D = u^2 D'$ with $u = [\mathcal{O}' : \mathcal{O}]$ and every prime p that splits completely in L also splits completely in L', since $4p = t^2 - v^2 D$ implies $4p = t^2 - (v')^2 D'$ for $v' = uv$, thus $L' \subseteq L$. □

Remark 21.7. The converse of Corollary 21.6 does not hold because we can have $L \subseteq L'$ even when L and L' are ring class fields of orders in different imaginary quadratic fields. This happens when $L' = \mathbb{Q}(\sqrt{-15})(\sqrt{5})$ is the ring class field of the order $\mathcal{O}'$ of discriminant -15, since L' contains the ring class field $L = \mathbb{Q}(\sqrt{-3})$ of the order O of discriminant -3, which is not a suborder of $\mathcal{O}'$ (note that $\mathcal{O}'$ does not contain $\zeta_3 \in \mathcal{O}$, for example).

Lemma 21.8. *Let K be an imaginary quadratic field of discriminant D with ring of integers $\mathcal{O}_K = [1, \omega]$ and let p be prime. Every $\mathcal{O}_K$-ideal of norm p is of the form $\mathfrak{p} = [p, \omega - r]$, where $r \in \mathbb{Z}$ is a root of the minimal polynomial of ω modulo p. The number of such ideals $\mathfrak{p}$ is $1 + \left(\frac{D}{p}\right) \in \{0, 1, 2\}$ and the factorization of the principal $\mathcal{O}_K$-ideal into prime ideals is*

$$(p) = \begin{cases} \mathfrak{p}\overline{\mathfrak{p}} & \text{if } \left(\frac{D}{p}\right) = 1, \\ \mathfrak{p}^2 & \text{if } \left(\frac{D}{p}\right) = 0, \\ (p) & \text{if } \left(\frac{D}{p}\right) = -1. \end{cases}$$

with $\mathfrak{p} \neq \overline{\mathfrak{p}}$ when $\left(\frac{D}{p}\right) = 1$.

Proof. Let $f(x) = x^2 - (\omega + \overline{\omega})x + \omega\overline{\omega} \in \mathbb{Z}[x]$ be the minimal polynomial of ω and let $\mathfrak{p}$ be an $\mathcal{O}_K$-ideal of norm p. Every nonzero $\mathcal{O}_K$-ideal is invertible, so by Theorem 17.10 we have $\mathfrak{p}\overline{\mathfrak{p}} = (\mathrm{N}\mathfrak{p}) = (p)$. Thus $p \in \mathfrak{p}$, and every integer $n \in \mathfrak{p}$ must be a multiple of p because otherwise $\gcd(n, p) = 1 \in \mathfrak{p}$ would imply $\mathfrak{p} = \mathcal{O}_K$ has norm $1 \neq p$. Therefore $\mathfrak{p} \cap \mathbb{Z} = p\mathbb{Z}$.

We can thus write $\mathfrak{p} = [p, a\omega - r]$ for some $a, r \in \mathbb{Z}$, and $[\mathcal{O}_K : \mathfrak{p}] = p$ then implies $a = 1$. The ideal $\mathfrak{p}$ is closed under multiplication by $\mathcal{O}_K$, so in particular it must contain

$$(\overline{\omega} - r)(\omega - r) = \overline{\omega}\omega - (\overline{\omega} + \omega)r + r^2 = f(r),$$

which is both an integer and an element of $\mathfrak{p}$, hence a multiple of p. Thus r must be a root of $f(x) \bmod p$. Conversely, if r is any root of $f(x) \bmod p$, then $[p, \omega - r]$ is an $\mathcal{O}_K$-ideal of norm p, and if $f(x) \bmod p$ has roots r and s that are distinct modulo p, then the $\mathcal{O}_K$-ideals $[p, \omega - r]$ and $[p, \omega - s]$ are clearly distinct.

It follows that the number of $\mathcal{O}_K$-ideals of prime norm p is equal to the number of distinct roots of $f(x) \bmod p$. The discriminant of $f(x)$ is

$$(\omega + \overline{\omega})^2 - 4\omega\overline{\omega} = (\omega - \overline{\omega})^2 = \operatorname{disc} \mathcal{O}_K = D, \tag{2}$$

and when p is odd it follows from the quadratic equation that the number of distinct roots of $f(x) \bmod p$ is $1 + \left(\frac{D}{p}\right)$, since this is the number of distinct square-roots of D modulo p.

For $p = 2$, we first note that if $D \equiv 0 \bmod 4$ then (2) implies that $\omega + \overline{\omega}$ is even, so $f(x) \equiv x^2 \bmod 2$ has $1 = 1 + \left(\frac{D}{2}\right)$ distinct roots. If $D \equiv 1 \bmod 4$ then $\omega + \overline{\omega}$ must be odd. If $D \equiv 1 \bmod 8$ then (2) implies that $\omega\overline{\omega}$ must be even (since $(\omega + \overline{\omega})^2 \equiv 1 \bmod 8$), and then $f(x) \equiv x^2 + x \bmod 2$ has $2 = 1 + \left(\frac{D}{2}\right)$ distinct roots. If $D \equiv 5 \bmod 8$ then $\omega\overline{\omega}$ must be odd, and then $f(x) \equiv x^2 + x + 1 \bmod 2$ has $0 = 1 + \left(\frac{D}{2}\right)$ distinct roots. □

Corollary 21.9. *Let $\mathcal{O}$ be an order of discriminant D in an imaginary quadratic field K, and let p be a prime. When p divides the conductor $[\mathcal{O}_K : \mathcal{O}]$ there are no proper $\mathcal{O}$-ideals of norm p and otherwise there are $1 + \left(\frac{D}{p}\right) = 0, 1, 2$, depending on whether p is inert, ramified, or split in K, respectively.*

21.4 Class field theory

The theory of complex multiplication was originally motivated not by the study of elliptic curves, but as a way to construct abelian extensions of imaginary quadratic fields. A celebrated theorem of Kronecker and Weber states that every finite abelian extension of $\mathbb{Q}$ lies in a cyclotomic field (a field of the form $\mathbb{Q}(\zeta_n)$, for some nth root of unity ζ_n). The effort to generalize this result led to the development of *class field theory*, a branch of algebraic number theory that was one of the major advances of early 20th century number theory.

In 1898 Hilbert conjectured that every number field K has a unique maximal abelian extension L/K that is unramified at every prime[4] of K, for which $\mathrm{Gal}(L/K) \simeq \mathrm{cl}(\mathcal{O}_K)$. This conjecture was proved shortly thereafter by Furtwängler, and the field L is now known as the *Hilbert class field* of K. While its existence was quickly proved, the problem of explicitly constructing L, say by specifying a generator for L in terms of its minimal polynomial over K, remained an open problem (and for general K it still is).

The field $\mathbb{Q}$ has no nontrivial unramified extensions (let alone abelian ones), so its Hilbert class field is not interesting (it is just $\mathbb{Q}$). After $\mathbb{Q}$, the simplest fields K to consider are imaginary quadratic fields. For an imaginary quadratic field K of discriminant D, the splitting field L of the Hilbert class polynomial $H_D(X)$ over K is a Galois extension of K with Galois group $\mathrm{Gal}(L/K) \simeq \mathrm{cl}(O_K)$. It follows from class field theory that L must be the Hilbert class field of K. The Hilbert class field of an imaginary quadratic field K can also be characterized as the minimal extension L/K over which there exists an elliptic curve E with CM by $\mathcal{O}_K$; in other words, $L = K(j(E))$.

What about the splitting field L of a Hilbert class polynomial $H_D(X)$ over $K = \mathbb{Q}(\sqrt{D})$ when D is the discriminant of a non-maximal order $\mathcal{O} \subsetneq \mathcal{O}_K$? These are called *ring class fields*. They are abelian extensions of K with Galois group $\mathrm{Gal}(L/K) \simeq \mathrm{cl}(\mathcal{O})$, but unlike the Hilbert class field of K, they are necessarily ramified at some primes. It follows from class field theory that ramified primes are not proper $\mathcal{O}$-ideals.

The ring class field L is characterized by the infinite set $\mathcal{S}_{L/\mathbb{Q}}$ of primes that split completely in L, and with finitely many exceptions, these are precisely the primes p that satisfy the equation $4p = t^2 - v^2 D$ for some $t, v \in \mathbb{Z}$, with $D = \mathrm{disc}(\mathcal{O})$; see [4, Thm. 9.2, Ex. 9.3]. Any extension M/K for which the set $\mathcal{S}_{M/\mathbb{Q}}$ matches $\mathcal{S}_{L/\mathbb{Q}}$ with only finitely many exceptions must in fact be equal to L, by [4, Thm. 8.19]. We thus have the following corollary of Theorem 21.5, which removes the assumption that p is unramified in L.

Corollary 21.10. *Let $\mathcal{O}$ be an order of discriminant D in an imaginary quadratic field K. The splitting field L of $H_D(X)$ over K is unramified at all primes that do not divide the conductor of $\mathcal{O}$. In particular, every rational prime $p \nmid D$ is unramified in L.*

Ring class fields allow us to explicitly construct infinitely many abelian extensions of a given imaginary quadratic field K. One might ask whether every abelian extension of K is contained in a ring class field. This is not the case, but by extending the ring class field of an order $\mathcal{O}$ by adjoining the x-coordinates of the n-torsion points of an elliptic curve with CM by $\mathcal{O}$ (or powers of them, when $\mathrm{disc}\,\mathcal{O} \in \{-3, -4\}$), one obtains what are known as *ray class fields*, which depend on the choice of both $\mathcal{O}$ and n. These are analogs of the cyclotomic extensions of $\mathbb{Q}$ (which is its own Hilbert class field because it has no unramified extensions). An analog of the Kronecker-Weber theorem then holds: every abelian extension of an imaginary quadratic field is contained in a ray class field. One can define ring class fields and ray class fields for arbitrary number fields, and obtain a similar result (this was started by Weber and finished by Takagi around 1920), but the constructions are not nearly as explicit as they are in the imaginary quadratic case.

21.5 The CM method

The equation

$$4p = t^2 - v^2 D$$

[4]This includes not only all prime $\mathcal{O}_K$-ideals, but also "infinite primes" of K, corresponding to embeddings of K into $\mathbb{C}$. For imaginary quadratic fields K this imposes no additional restrictions.

in part (iv) of Theorem 21.5 is known as the *norm equation*; it arises from the principal $\mathcal{O}$-ideal (λ) of norm p given by part (i), generated by a root $\lambda \in \mathcal{O} \subseteq \mathcal{O}_K$ of $x^2 - tx + p$, which has norm p and trace t. By the quadratic equation

$$\lambda = \frac{-t \pm \sqrt{t^2 - 4p}}{2} = \frac{-t \pm v\sqrt{D}}{2}.$$

Clearing denominators and taking norms yields the equation $\mathrm{N}(2\lambda) = 4\lambda\bar{\lambda} = 4p = t^2 - v^2 D$.

Let us assume this equation holds with $p \nmid D$ odd and $D < -4$. The prime p splits completely in the ring class field L for the order $\mathcal{O}$ of discriminant D, and we can completely factor $H_D(X)$ in both $\mathcal{O}_L[x]$ and $\mathbb{F}_p[x]$. If we now fix a prime $\mathfrak{q}$ lying above p, then $\mathrm{N}\mathfrak{q} = p$, by Theorem 21.5, and we have a reduction map $\mathcal{O}_L \to \mathcal{O}_L/\mathfrak{q} \simeq \mathbb{F}_p$ that we can apply to the roots of $H_D(X)$, equivalently, to the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C}) = \{j(E) \in \mathbb{C} : \mathrm{End}(E) \simeq \mathcal{O}\}$.

It follows that the j-invariant $j(E)$ of any elliptic curve $E/\mathbb{C}$ with CM by $\mathcal{O}$ can be reduced (modulo $\mathfrak{q}$) to the j-invariant of an elliptic curve $\overline{E}/\mathbb{F}_p$ that is the reduction of E: we can always pick a model $y^2 = x^3 + Ax + B$ for E with $A, B \in \mathcal{O}_L$ such that $\mathfrak{q} \nmid \Delta(E)$ because p is odd and the denominator of $j(E)$ has to be nonzero modulo $\mathfrak{q}$. Now we know that $\mathrm{End}(E) \simeq \mathcal{O}$, but what about $\mathrm{End}(\overline{E})$?

If $\varphi \in \mathrm{End}(E) \simeq \mathcal{O}$ is a nonzero endomorphism of E, then we can reduce the coefficients of the rational functions defining φ modulo $\mathfrak{q}$ to obtain a corresponding endomorphism $\bar{\varphi} \in \mathrm{End}(\overline{E})$. The endomorphism $\bar{\varphi}$ is nonzero because it must satisfy the characteristic equation $x^2 - [\mathrm{tr}\,\varphi]x + [\deg\varphi] = 0$ in $\mathrm{End}(\overline{E})$: multiplication-by-$n$ maps $[n]$ can always be reduced from $\mathrm{End}(E)$ to $\mathrm{End}(\overline{E})$, so $[\mathrm{tr}\,\varphi]$ and $[\deg\varphi]$ reduce to maps $[\mathrm{tr}\,\bar{\varphi}]$ and $[\deg\bar{\varphi}]$ that represent multiplication by the same integers. It follows that the reduction map induces an injective ring homomorphism

$$\mathrm{End}(E) \hookrightarrow \mathrm{End}(\overline{E}). \tag{3}$$

In fact this map is an isomorphism (see §21.6), but for the moment we will content ourselves with showing that it at least induces an isomorphism of endomorphism algebras. By Corollary 13.20 we know that $\mathrm{End}^0(\overline{E})$ is either an imaginary quadratic field or a quaternion algebra, depending on whether $\overline{E}$ is ordinary or supersingular.

Corollary 21.11. *Let $\mathcal{O}$ be an imaginary quadratic order with discriminant D and ring class field L, and let $p \nmid D$ be an odd prime satisfying $4p = t^2 - v^2 D$. Every $j(E) \in \mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is the j-invariant of an elliptic curve E/L with good reduction $\overline{E}$ modulo a prime $\mathfrak{q}$ of L lying above p. Provided $j(\overline{E}) \neq 0, 1728$, we have $\mathrm{tr}\,\pi_{\overline{E}} = \pm t \not\equiv 0 \bmod p$ and $\overline{E}$ is ordinary.*[5]

Proof. By Theorem 21.5 and its proof, p is the norm of a principal $\mathcal{O}$-ideal $\mathfrak{p} := (\lambda)$, where λ has norm p and trace t. As in the proof of Theorem 21.1, one of the isogenies $\phi_{\mathfrak{p}} \colon E \to \mathfrak{p}E$ and $\phi_{\bar{\mathfrak{p}}} \colon E \to \bar{\mathfrak{p}}E$ induces a purely inseparable isogeny $\phi \colon \overline{E} \to \overline{E}^{(p)} = \overline{E}$, which up to an automorphism, must be the Frobenius endomorphism $\pi_{\overline{E}}$. We have $\mathrm{tr}\,\phi = \mathrm{tr}\,\phi_{\mathfrak{p}} = \mathrm{tr}\,\phi_{\bar{\mathfrak{p}}} = t$, with $t \not\equiv 0 \bmod p$ by part (iv) of Theorem 21.5. For $j(\overline{E}) \neq 0, 1728$ the only automorphisms of $\overline{E}$ are ± 1, so $\mathrm{tr}\,\pi_{\overline{E}} = \pm\,\mathrm{tr}\,\phi = \pm t \not\equiv 0 \bmod p$ and $\overline{E}$ is ordinary. □

Corollary 21.11 gives us an explicit method for constructing elliptic curves over finite fields with a prescribed number of rational points. Let $D < -4$ be an imaginary quadratic discriminant and let $p \nmid D$ be an odd prime. In this case the norm equation $4p = t^2 - v^2 D$ determines t (and v) up to a sign, and we can efficiently compute a solution (t, v) using

[5] In fact $\overline{E}$ is also ordinary when $j(\overline{E}) \in \{0, 1728\}$, but this takes more work to prove.

Cornacchia's algorithm (see Problem Set 2). Given the Hilbert class polynomial $H_D(X)$, we can efficiently compute a root j_0 of $H_D(X)$ over $\mathbb{F}_p$ (using a randomized root-finding algorithm) and then write down the equation $y^2 = x^3 + Ax + B$ of an elliptic curve E with $j(E) = j_0$, using $A = 3j_0(1728 - j_0)$ and $B = 2j_0(1728 - j_0)^2$ (assuming $j_0 \neq 0, 1728$).

The Frobenius endomorphism π_E then satisfies $\operatorname{tr} \pi_E = \pm t$, and by Hasse's theorem,

$$\#E(\mathbb{F}_p) = p + 1 - \operatorname{tr}(\pi_E).$$

The sign of $\operatorname{tr} \pi_E$ can be explicitly determined using the formulas in [9]. Alternatively, one can simply pick a random point $P \in E(\mathbb{F}_p)$ and check whether $(p + 1 - t)P = 0$ or $(p + 1 + t)P = 0$ both hold (at least one must); if only one of these equations is satisfied, then $\operatorname{tr} \pi$ is determined (for large p this will almost always happen with the first P we try). Note that we can always change the sign of $\operatorname{tr} \pi$ by replacing E with its quadratic twist.

Now suppose that we wish to construct an elliptic curve E over some finite field $\mathbb{F}_p$ such that $\#E(\mathbb{F}_p) = N$, for some positive integer N. Provided we can factor N (typically N is prime and this is easy), we can use Cornacchia's algorithm to find a solution (a, v) to

$$4N = a^2 - v^2 D$$

for any particular imaginary quadratic discriminant D, whenever such a solution exists.[6] Given a solution (a, v), we put $t := a + 2$ and check whether $p := N - 1 + t$ is prime. If not, or if no solution (a, v) can be found, we just try a different discriminant D. In practice this will happen quite quickly; see [3] for a heuristic complexity analysis.

Once we have $p = N - 1 + t$ prime, we then observe that

$$4p = 4N - 4 + 4t = a^2 - v^2 D - 4 + 4a + 8 = (a + 2)^2 - v^2 D = t^2 - v^2 D,$$

so the norm equation is satisfied, and we can construct an elliptic curve $E/\mathbb{F}_p$ with $\operatorname{tr} \pi_E = \pm t$ using the Hilbert class polynomial $H_D(X)$ as described above, taking a quadratic twist if necessary to get $\operatorname{tr} \pi_E = t$. We then have $\#E(\mathbb{F}_p) = p + 1 - t = N$ as desired.

This method of constructing an elliptic curve $E/\mathbb{F}_p$ is known as the *CM method*. The CM method has many applications, one of which is an improved version of elliptic curve primality proving developed by Atkin and Morain [1]; see Problem Set 11.

Remark 21.12. It can happen that $H_D(X)$ has roots in $\mathbb{F}_p$ even when p does not split completely in the ring class field L. These roots cannot be j-invariants of elliptic curves $E/\mathbb{F}_p$ with $\operatorname{End}(E) = \mathcal{O}$, we must have $\mathcal{O} \subsetneq \operatorname{End}(E)$, and in fact the fraction field K of $\mathcal{O}$ must be properly contained in $\operatorname{End}^0(E)$. This means that $\operatorname{End}^0(E)$ has to be a quaternion algebra that contains the imaginary quadratic field K. This cannot happen when $p = \mathfrak{p}\overline{\mathfrak{p}}$ splits in K (which occurs exactly when $\left(\frac{D}{p}\right) = 1$), because L/K is Galois and the residue field extensions $\mathbb{F}_\mathfrak{q}/\mathbb{F}_\mathfrak{p}$ all have the same degree (so $H_D \bmod p$ either has no roots at all or splits completely and in the latter case p must split completely in the ring class field for $\mathcal{O}$). But if p is inert in K then $H_D(X)$ can easily have roots modulo p that must be j-invariants of supersingular elliptic curves. This actually provides a very efficient method for constructing supersingular elliptic curves; see [2] for details.

[6] We need to be able to factor N because Cornacchia's algorithm requires a square root of D modulo N; computing square roots modulo primes is easy, and if we know the factorization of N we can use the CRT to reduce to this case; in general, computing square roots modulo N is as hard as factoring N.

Remark 21.13. We have restricted our attention to prime fields $\mathbb{F}_p$ in order to simplify the exposition, but everything we have done generalizes to arbitrary finite fields $\mathbb{F}_q$ of prime power order q. If $\mathcal{O}$ is an imaginary quadratic order of discriminant D with ring class field L, in Theorem 21.5 we can replace $p \nmid D$ with $q \perp D$, replace $\left(\frac{D}{p}\right) = 1$ with the requirement that D is a square in $\mathbb{F}_q$ (automatic when q is a square), and rather than requiring p to split completely in L we require q to be the norm of a prime ideal $\mathfrak{q}$ in $\mathcal{O}_L$. The norm equation then becomes $4q = t^2 - v^2 D$ with $t \perp q$, and if it is satisfied with $D < -4$ the Hilbert class polynomial $H_D(X)$ splits completely in $\mathbb{F}_q[x]$ and its roots are j-invariants of elliptic curves $E/\mathbb{F}_q$ with $\operatorname{tr} \pi_E = \pm t$ (which in fact have $\operatorname{End}(E) = \mathcal{O}$).

The main limitation of the CM method is that it requires computing the Hilbert class polynomial $H_D(X)$, which becomes very difficult when $|D|$ is large. The degree of $H_D(X)$ is the class number $h(D) \approx \sqrt{|D|}$, and the size of its largest coefficient is on the order of $\sqrt{|D|} \log |D|$ bits.[7] Thus the total size of $H_D(X)$ is on the order of $|D| \log |D|$ bits, which makes it impractical to even write down if $|D|$ is large. An efficient algorithm for computing $H_D(X)$ is outlined in Problem Set 11, and with a suitably optimized implementation, it can practically handle discriminants with $|D|$ as large as 10^{13}, for which the size of $H_D(X)$ is several terabytes [11]. Using class polynomials associated to other modular functions discriminants up to $|D| \approx 10^{15}$ can be readily addressed [5], and with more advanced techniques, even $|D| \approx 10^{16}$ is feasible [12].

21.6 The Deuring lifting theorem

As noted in the previous section, the injective ring homomorphism $\operatorname{End}(E) \hookrightarrow \operatorname{End}(\overline{E})$ given by (3), where $\overline{E}/\mathbb{F}_p$ is the reduction of an elliptic curve E/L with CM by $\mathcal{O}$ over its ring class field L modulo an unramified prime $\mathfrak{q}$ of norm p, is actually an isomorphism. Moreover, every elliptic curve over $\mathbb{F}_p$ with CM by $\mathcal{O}$ arises as the reduction of an elliptic curve E/L, and this correspondence is a bijective at the level of j-invariants. These facts follow from results of Deuring that we won't take the time to prove, but record here for reference.

Theorem 21.14 (Deuring)**.** *Let $\mathcal{O}$ be an imaginary quadratic order of discriminant D with ring class field L, and let q be the norm of a prime ideal in $\mathcal{O}_L$ with $q \perp D$. Then $H_D(X)$ splits into distinct linear factors in $\mathbb{F}_q[X]$ and its roots form the set*

$$\operatorname{Ell}_{\mathcal{O}}(\mathbb{F}_q) := \{j(E) \in \mathbb{F}_q : \operatorname{End}(E) \simeq \mathcal{O}\}.$$

of j-invariants of elliptic curves $E/\mathbb{F}_q$ with CM by $\mathcal{O}$.

Proof. This follows from [6, Thm. 13]. □

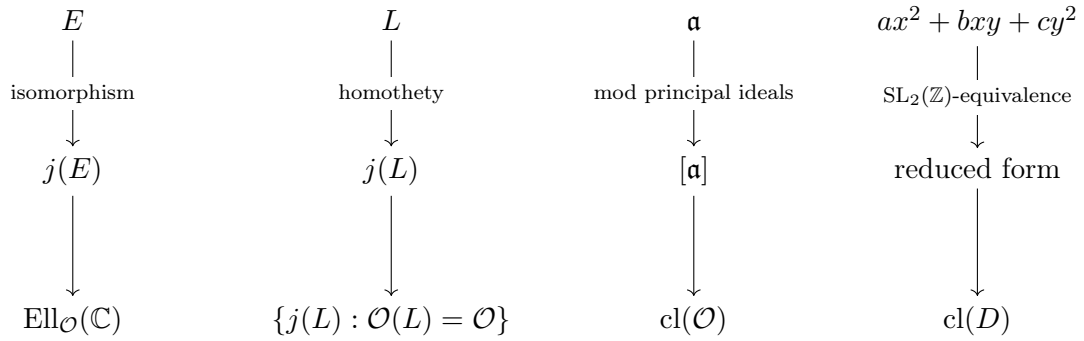
Theorem 21.15 (Deuring lifting theorem)**.** *Let $E/\mathbb{F}_q$ be an elliptic curve over a finite field and let $\phi \in \operatorname{End}(E)$ be nonzero. There exists an elliptic curve E^* over a number field L with an endomorphism $\phi^* \in \operatorname{End}(E^*)$ such that E^* has good reduction modulo a prime $\mathfrak{q}$ of L with residue field $\mathcal{O}_L/\mathfrak{q} \simeq \mathbb{F}_q$, and E and ϕ are the reductions modulo $\mathfrak{q}$ of E^* and ϕ^*.*

Proof. See [6, Thm. 14]. □

[7] Under the Generalized Riemann Hypothesis, these bounds are accurate to within an $O(\log \log |D|)$ factor.

21.7 Summing up the theory of complex multiplication

Let $\mathcal{O}$ be an imaginary quadratic order of discriminant D.



The figure above illustrates four different objects that have been our focus of study for the last several weeks:

1. Elliptic curves $E/\mathbb{C}$ with CM by $\mathcal{O}$.
2. Lattices L (which define tori $\mathbb{C}/L$ that correspond to elliptic curves).
3. Proper $\mathcal{O}$-ideals $\mathfrak{a}$ (which may be viewed as lattices).
4. Reduced primitive positive definite binary quadratic forms of discriminant D (which correspond to proper $\mathcal{O}$-ideals of norm a).

In each case we defined a notion of equivalence: isomorphism, homethety, equivalence modulo principal ideals, and equivalence modulo an $\mathrm{SL}_2(\mathbb{Z})$-action, respectively. Modulo this equivalence, we obtain a finite set of objects with the cardinality $h(\mathcal{O}) = h(D)$ in each case. The two sets on the right, $\mathrm{cl}(\mathcal{O})$ and $\mathrm{cl}(D)$, are finite abelian groups that act on the two sets on the left, both of which are equal to $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$. This action is free and transitive, so that $\mathrm{Ell}_{\mathcal{O}}(\mathbb{C})$ is a $\mathrm{cl}(\mathcal{O})$-torsor.

References

[1] A.O.L. Atkin and François Morain, *Elliptic curves and primality proving*, Mathematics of Computation **61** (1993), 29–68.

[2] Reinier Bröker, *Constructing supersingular elliptic curves*, Journal of Combinatorics and Number Theory **1** (2009), 269–273.

[3] Reinier Bröker and Peter Stevenhagen, *Efficient CM-constructions of elliptic curves over finite fields*, Mathematics of Computation **76** (2007), 2161–2179.

[4] David A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory, and complex multiplication*, Wiley, 1989.

[5] Andreas Enge and Andrew V. Sutherland, *Class invariants by the CRT method*, ANTS IX, LNCS 6197, Springer, 2010, pp. 142-156.

[6] Serge Lang, *Elliptic functions*, Springer, 1987.

[7] J. S. Milne, *Algebraic number theory*, course notes, 2020.

[8] Jürgen Neukirch, *Algebraic number theory*, Springer, 1999.

[9] Karl Rubin and Alice Silverberg, *Choosing the correct elliptic curve in the CM method*, Mathematics of Computation **79** (2010), 545–561.

[10] Joseph H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Springer, 1994.

[11] Andrew V. Sutherland, *Computing Hilbert class polynomials with the Chinese Remainder Theorem*, Mathematics of Computation **80** (2011), 501–538.

[12] Andrew V. Sutherland, *Accelerating the CM method*, LMS Journal of Computation and Mathematics **15** (2012), 172–204.

22 Isogeny volcanoes

We now want to shift our focus from elliptic curves over $\mathbb{C}$ to elliptic curves over other fields, finite fields in particular. As noted in Lecture 20, the moduli interpretation of the modular curve $X_0(N)$ parameterizing cyclic isogenies of degree N is valid over any field whose characteristic does not divide N; see Theorem 20.4. We can thus use the modular equation $\Phi_N \in \mathbb{Z}[X,Y]$ to identify pairs of N-isogenous elliptic curves using j-invariants in any field k. When k is not algebraically closed this determines the elliptic curves only up to a twist, but for finite fields there are generally only two twists to consider (assuming $j \neq 0, 1728$), and in many applications it suffices to work with $\bar{k}$ isomorphism classes of elliptic curves defined over k, equivalently the set of j-invariants of elliptic curves E/k, which by Theorem 13.12, is just the set k itself.

We are particularly interested in the case where N is a prime $\ell \neq \operatorname{char}(k)$. Every isogeny of degree ℓ is necessarily cyclic (since ℓ is prime), and for any fixed j-invariant $j_1 := j(E_1) \in k$, the k-rational roots of the polynomial

$$\phi_\ell(Y) = \Phi_\ell(j_1, Y)$$

are the j-invariants of the elliptic curves E_2/k that are ℓ-isogenous to E_1. More precisely, there is a bijection between the $\operatorname{Gal}(\bar{k}/k)$-invariant roots of $\phi_\ell(Y)$ in k and the $\operatorname{Gal}(\bar{k}/k)$-invariant order-$\ell$ cyclic subgroups of $E[\ell]$, provided we count roots of $\phi_\ell(Y)$ with multiplicity. Over $\bar{k}$ there are $\deg \phi_\ell = \ell + 1$ (not necessarily distinct) roots of ϕ_ℓ corresponding to $\ell + 1$ (necessarily distinct) cyclic subgroups of $E[\ell] \simeq \mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z}$ of order ℓ. Recall from Theorem 5.11 that every finite subgroup of $E(\bar{k})$ is the kernel of a separable isogeny that is uniquely determined up to composition with isomorphisms. As we are only interested in isogenies up to isomorphism, we consider separable isogenies to be distinct only when their kernels differ.

Throughout this lecture we assume $\ell \neq \operatorname{char}(k)$, so all the isogenies we shall consider are separable.

Definition 22.1. The *ℓ-isogeny graph* $G_\ell(k)$ is the directed graph with vertex set k and edges (j_1, j_2) present with multiplicity equal to the multiplicity of j_2 as a root of $\Phi_\ell(j_1, Y)$.

As noted in Remark 20.6, if $j_1 = j(E_1)$ and $j_2 = j(E_2)$ are the j-invariants of a pair of ℓ-isogenous elliptic curves, the ordered pair (j_1, j_2) does not uniquely determine an ℓ-isogeny $\varphi\colon E_1 \to E_2$; there may be multiple ℓ-isogenies from E_1 to E_2 with distinct kernels. The existence of the dual isogeny guarantees that (j_1, j_2) is an edge in $G_\ell(k)$ if and only if (j_2, j_1) is also an edge; provided that $j_1, j_2 \neq 0, 1728$ these edges have the same multiplicity, but in the exceptional case where one of j_1 and j_2 is 0 or 1728, this need not hold.

Remark 22.2. The exceptions for j-invariants $0 = j(\rho)$ and $1728 = j(i)$ arise from the fact that the corresponding elliptic curves $y^2 = x^3 + B$ and $y^2 = x^3 + Ax$ have automorphisms $\rho\colon (x,y) \mapsto (\rho x, y)$ and $i\colon (x,y) \mapsto (-x, iy)$, respectively, where ρ and i denote third and fourth roots of unity, respectively, in both $\operatorname{End}(E)$ and $\bar{k}$. The automorphism -1 does not pose a problem because it fixes every cyclic subgroup of $E[\ell]$, so for any ℓ-isogeny $\varphi\colon E_1 \to E_2$ the isogeny $\varphi \circ [-1] = [-1] \circ \varphi$ has the same kernel as φ; this does not apply to ρ and i, which fix only two cyclic subgroups of $E[\ell]$. If $j(E_1) = 0$ and $j(E_2) \neq 0$ then we cannot write "$\varphi \circ \rho = \rho \circ \varphi$" (the RHS does not even make sense, $\rho \notin \operatorname{End}(E_2)$) and the isogenies $\varphi, \varphi \circ \rho, \varphi \circ \rho^2$ all have different kernels, but the corresponding dual-isogenies all have the same kernel. In this situation the edge $(j(E_1), j(E_2))$ has multiplicity 3 in $G_\ell(k)$ but the

edge $(j(E_2), j(E_1))$ has multiplicity 1. The case where $j(E_1) = 1728$ and $j(E_2) \neq 1728$ is similar, except now $(j(E_1), j(E_2))$ has multiplicity 2.

Our objective in this lecture is to elucidate the structure of the graph $G_\ell(k)$ in the case that $k = \mathbb{F}_q$ is a finite field. Recall from Lecture 14 that elliptic curves over finite fields may be classified according to their endomorphism algebras and are either ordinary (meaning $\mathrm{End}^0(E)$ is an imaginary quadratic field) or supersingular (meaning $\mathrm{End}^0(E)$ is a quaternion algebra). Whether E is ordinary or supersingular is an isogeny invariant (by Theorem 13.2), so the graph $G_\ell(\mathbb{F}_q)$ can always be partitioned into ordinary and supersingular components. Since most elliptic curves are ordinary, we will focus on the ordinary components; you will have an opportunity to investigate the supersingular components on Problem Set 12.

22.1 Isogenies between elliptic curves with complex multiplication

Theorem 22.3. *Let $\varphi\colon E \to E'$ be an ℓ-isogeny of elliptic curves defined over a field k. Then $\mathrm{End}^0(E') \simeq \mathrm{End}^0(E)$, and if $\mathrm{End}^0(E) = K$ is an imaginary quadratic field then $\mathrm{End}(E) = \mathcal{O}$ and $\mathrm{End}(E') = \mathcal{O}'$ are orders in K such that one of the following holds:*

$$\text{(i) } \mathcal{O} = \mathcal{O}', \qquad \text{(ii) } [\mathcal{O} : \mathcal{O}'] = \ell, \qquad \text{(iii) } [\mathcal{O}' : \mathcal{O}] = \ell.$$

Proof. Let $\hat{\varphi}\colon E' \to E$ be the dual isogeny. If $\phi \in \mathrm{End}(E)$, the isogeny $\varphi \circ \phi \circ \hat{\varphi}\colon E' \to E'$ is an endomorphism $\phi' \in \mathrm{End}(E')$ with

$$\begin{aligned}
\mathrm{T}\phi' &= \phi' + \hat{\phi}' = \varphi \circ \phi \circ \hat{\varphi} + \varphi \circ \hat{\phi} \circ \hat{\varphi} = \varphi \circ [\mathrm{T}\phi] \circ \hat{\varphi} = \varphi \circ \hat{\varphi} \circ [\mathrm{T}\phi] = \ell\, \mathrm{T}\phi, \\
\mathrm{N}\phi' &= \phi' \circ \hat{\phi}' = \varphi \circ \phi \circ \hat{\varphi} \circ \varphi \circ \hat{\phi} \circ \hat{\varphi} = \varphi \circ \phi \circ [\ell] \circ \hat{\phi} \circ \hat{\varphi} = \varphi \circ [\ell\, \mathrm{N}\phi] \circ \hat{\varphi} = \ell^2 \mathrm{N}\phi,
\end{aligned}$$

and ϕ' is a root of $x^2 - (\mathrm{T}\phi')x + \mathrm{N}\phi' = x^2 - (\mathrm{T}\phi)(\ell x) + \ell^2 \mathrm{N}\phi = 0$. Thus $\phi'/\ell \in \mathrm{End}^0(E')$ is a root of $x^2 - (\mathrm{T}\phi)x + \mathrm{N}\phi$, and it follows that the characteristic polynomial of every $\phi \in \mathrm{End}(E)$ has a root in $\mathrm{End}^0(E')$ and therefore $\mathrm{End}(E) \subseteq \mathrm{End}^0(E')$. Applying the same argument in the reverse direction shows that $\mathrm{End}(E') \subseteq \mathrm{End}^0(E)$, so we must have $\mathrm{End}^0(E') = \mathrm{End}^0(E)$.

Assume $\mathrm{End}^0(E') \simeq \mathrm{End}^0(E)$ is an imaginary quadratic field, with $\mathrm{End}(E) = \mathcal{O} = [1, \tau]$ and $\mathrm{End}(E') = \mathcal{O}' = [1, \tau']$. Then $\varphi \circ \tau \circ \hat{\varphi} \in \mathrm{End}(E') = \mathcal{O}'$ has the same characteristic polynomial as $\ell\tau \in \mathcal{O}$, which implies $\ell\tau \in \mathcal{O}'$ (since $\mathcal{O}$ and $\mathcal{O}'$ lie in the same field K). We similarly find that $\ell\tau' \in \mathcal{O}$. Thus $[1, \ell\tau] \subseteq [1, \tau']$, and $[1, \ell\tau'] \subseteq [1, \tau]$, and therefore

$$[1, \ell^2\tau] \subseteq [1, \ell\tau'] \subseteq [1, \tau].$$

The index of $[1, \ell^2\tau]$ in $[1, \tau]$ is ℓ^2, so the index of $[1, \ell\tau']$ in $[1, \tau]$ must be $1, \ell$, or ℓ^2. These correspond to cases (iii), (i), and (ii) of the theorem, respectively. □

Definition 22.4. Theorem 22.3 allows us to distinguish ℓ-isogenies $\varphi\colon E \to E'$ of elliptic curves with CM by an imaginary quadratic field as follows:

(i) when $\mathcal{O} = \mathcal{O}'$ we say that φ is *horizontal*;

(ii) when $[\mathcal{O} : \mathcal{O}'] = \ell$ we say that φ is *descending*;

(iii) when $[\mathcal{O}' : \mathcal{O}] = \ell$ we say that φ is *ascending*.

We collectively refer to ascending and descending isogenies as *vertical* isogenies.

Theorem 22.5. *Let $E/\mathbb{C}$ be an elliptic curve with CM by an order $\mathcal{O}$ of discriminant D in an imaginary quadratic field K, and let ℓ be prime. If $\ell \nmid [\mathcal{O}_K : \mathcal{O}]$ then E admits $1 + \left(\frac{D}{\ell}\right)$ horizontal, $\ell - \left(\frac{D}{\ell}\right)$ descending, and no ascending ℓ-isogenies. Otherwise E admits no horizontal, ℓ descending, and one ascending ℓ-isogeny.*

Proof. We know that there are always $\ell + 1$ ℓ-isogenies in total, so it suffices to prove that the counts of horizontal and ascending ℓ-isogenies given in the theorem are correct.

Let us first consider the special case in which E corresponds to a torus $\mathbb{C}/L$ with $L := \ell\mathcal{O}$ homothetic to $\mathcal{O}$. As explained in Lecture 17 (see §17.5), every ℓ-isogeny $\varphi\colon E \to E'$ arises from a lattice inclusion $L \subseteq L'$ of index ℓ. The lattices L' containing $L = \ell\mathcal{O}$ with index ℓ are precisely the index-ℓ sublattices of $\mathcal{O}$. We then have

$$\mathrm{End}(E') \simeq \mathrm{End}(\mathbb{C}/L') = \mathcal{O}(L') := \{\alpha \in \mathbb{C} : \alpha L' \subseteq L'\},$$

and the inclusion $L \subseteq L'$ gives rise to a horizontal ℓ-isogeny if and only if $\mathcal{O}(L') = \mathcal{O}$, in other words, precisely when L' is a proper $\mathcal{O}$-ideal. By Corollary 21.9, if $\ell \nmid [\mathcal{O}_K : \mathcal{O}]$ there are $1 + \left(\frac{D}{\ell}\right)$ proper $\mathcal{O}$-ideals of norm ℓ, and otherwise there are none, which matches the claimed count of horizontal ℓ-isogenies.

When $\ell \nmid [\mathcal{O}_K : \mathcal{O}]$ there can be no ascending ℓ-isogenies, so it remains only to show that when ℓ divides $[\mathcal{O}_K : \mathcal{O}]$ there is exactly one ascending ℓ-isogeny. In this case $\mathcal{O}$ is an index-ℓ suborder of some order $\mathcal{O}'$ in $\mathcal{O}_K$, and we want to show that exactly one of the index ℓ sublattices L' of $\mathcal{O}$ (each of which contain $L = \ell\mathcal{O}$ with index ℓ) satisfies $\mathcal{O}(L') = \mathcal{O}'$. Let $\mathcal{O}' = [1, \omega]$. We can assume $\mathcal{O} = [1, \ell\omega]$, since this is clearly an index-ℓ suborder of $\mathcal{O}'$ and there is exactly one such suborder (by Theorem 17.18). The index-ℓ sublattices of $\mathcal{O}$ are $L_i := [\ell, \ell\omega + i]$ for $0 \le i < \ell$ and $L_\ell := [1, \ell^2\omega]$, by Lemma 20.2. Note that L_0 is homothetic to $\mathcal{O}'$, and we claim it is the unique index-ℓ sublattice L' for which $\mathcal{O}(L') = \mathcal{O}'$, equivalently, for which the inclusion $L \subseteq L'$ induces an ascending ℓ-isogeny $\varphi\colon E \to E'$.

We have $\mathcal{O}'L_0 = \mathcal{O}'\ell\mathcal{O}' = \ell\mathcal{O}' = L_0$, so $\mathcal{O}' \subseteq \mathcal{O}(L_0)$, and $\mathcal{O}(L_0)$ cannot be larger than $\mathcal{O}'$, since $\mathcal{O}'$ is the largest order possible for $\mathrm{End}(E')$, by Theorem 22.3, thus $\mathcal{O}(L_0) = \mathcal{O}'$ and the inclusion $L \subseteq L_0$ induces an ascending ℓ-isogeny. For $0 < i < \ell$ the element $\omega(\ell\omega + i) = \ell\omega^2 + i\omega \equiv i\omega \bmod \ell \notin [1, \ell\omega] = \mathcal{O}$ is not an element of L_i, so $\mathcal{O}' \not\subseteq \mathcal{O}(L_i)$, and for $i = \ell$ the element $\omega \cdot 1 = \omega \notin [1, \ell\omega] = \mathcal{O}$ is not an element of L_i and again $\mathcal{O}' \not\subseteq \mathcal{O}(L_i)$. Thus φ is an ascending ℓ-isogeny if and only if $L' = L_0$ and there is exactly one such φ.

We now consider the general case, in which L is homothetic to a proper $\mathcal{O}$-ideal $\mathfrak{a}$, which we can assume has prime norm $p \perp \ell[\mathcal{O}_K : \mathcal{O}]$ (by Theorem 20.11, every ideal class in $\mathrm{cl}(\mathcal{O})$ contains infinitely many ideals of prime norm). The CM action of $\mathfrak{a}$ is a horizontal p-isogeny $\varphi_{\mathfrak{a}}\colon E \to E_0$, with $E_0 \simeq \mathbb{C}/\mathcal{O}$. Let $\varphi\colon E \to E'$ be an ℓ-isogeny, let $\mathcal{O}' = \mathrm{End}(E')$, and define

$$\mathfrak{a}' := \begin{cases} \mathfrak{a} & \text{if } \varphi \text{ is horizontal,} \\ \mathfrak{a}\mathcal{O}' & \text{if } \varphi \text{ is ascending,} \\ \mathfrak{a} \cap \mathcal{O}' & \text{if } \varphi \text{ is descending.} \end{cases}$$

We must have $[\mathcal{O}' : \mathfrak{a}'] = [\mathcal{O}_K : \mathfrak{a}'\mathcal{O}_K] = [\mathcal{O}_K : \mathfrak{a}\mathcal{O}_K] = [\mathcal{O} : \mathfrak{a}] = p$, since p does divide $[\mathcal{O}_K : \mathcal{O}]$ or $[\mathcal{O}_K : \mathcal{O}']$; it follows that $\mathfrak{a}'$ is a proper $\mathcal{O}'$-ideal of norm p, and we have a horizontal p-isogeny $\varphi_{\mathfrak{a}'}\colon E' \to E'_0$ with $E' \simeq \mathbb{C}/\mathfrak{a}$ and $E'_0 \simeq \mathbb{C}/\mathcal{O}'$. Up to isomorphism, there is a unique ℓ-isogeny $\varphi_0\colon E_0 \to E'_0$ for which the diagram

$$\begin{array}{ccc} E & \xrightarrow{\varphi_{\mathfrak{a}}} & E_0 \\ \downarrow{\scriptstyle \varphi} & & \downarrow{\scriptstyle \exists!\varphi_0} \\ E' & \xrightarrow{\varphi_{\mathfrak{a}'}} & E'_0 \end{array}$$

commutes, namely the isogeny with kernel $\varphi_{\mathfrak{a}}(\ker(\varphi_{\mathfrak{a}'} \circ \varphi))$ given by Theorem 5.11. Since $\varphi_{\mathfrak{a}}$ and $\varphi_{\mathfrak{a}'}$ are both horizontal, the ℓ-isogeny φ_0 must be of the same type (horizontal, descending, or ascending) as φ. This reduces the general case to the special case above. □

Theorem 22.5 extends to any field whose characteristic is not ℓ (provided that one takes rationality into account: ℓ-isogenies admitted by E over $\bar{k}$ need not be defined over k). We won't prove this in full generality, but we can use Deuring's lifting theorem to address the case where k is a finite field $\mathbb{F}_q$.

For an imaginary quadratic order $\mathcal{O}$ with discriminant D and any field k we define

$$\mathrm{Ell}_{\mathcal{O}}(k) := \{j(E) \in k : \mathrm{End}(E) = \mathcal{O}\},$$

the set of j-invariants of elliptic curves over k with CM by $\mathcal{O}$; for $k = \mathbb{C}$ this is the same as the set of roots of the Hilbert class polynomial $H_D(X)$, whose cardinality is the class number $h(D) := \#\mathrm{Cl}(\mathcal{O})$, and a result of Deuring noted in the previous lecture (see Theorem 21.14) yields a similar statement for finite fields.

Lemma 22.6. *Let $\mathcal{O}$ be an imaginary quadratic order of discriminant D and let $\mathbb{F}_q$ be a finite field with $q \perp D$. The set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{F}_q)$ is either empty or has cardinality $h(D)$. If $\mathrm{Ell}_{\mathcal{O}}(\mathbb{F}_q)$ is nonempty, so is $\mathrm{Ell}_{\mathcal{O}'}(\mathbb{F}_q)$ for every imaginary quadratic order $\mathcal{O}'$ containing $\mathcal{O}$.*

Proof. If $\mathrm{Ell}_{\mathcal{O}}(\mathbb{F}_q)$ is nonempty then there is an elliptic curve $E/\mathbb{F}_q$ with CM by $\mathcal{O}$. Its Frobenius endomorphism π_E is an element of $\mathrm{End}(E) = \mathcal{O}$ with trace $t = \mathrm{tr}\,\pi_E$ and norm q, and we must have $t \perp q$, since E is ordinary, by Corollary 13.20. The discriminant of the characteristic polynomial $x^2 - tx + q$ has a root $\pi_E \in \mathcal{O}$ that is not in $\mathbb{Z}$ (because $t \neq \pm 2\sqrt{q}$), so its discriminant $t^2 - 4q$ is a square in $\mathcal{O} - \mathbb{Z}$, hence of the form $v^2 D$ for some $v \in \mathbb{Z}$. We then have $4q = t^2 - v^2 D$ with $t \not\equiv 0 \bmod p = \mathrm{char}\ \mathbb{F}_q$, and it follows from Theorem 21.5 and Remark 21.13 that q is the norm of a prime ideal in $\mathcal{O}_L$, where L is the ring class field of $\mathcal{O}$. By Theorem 21.14, the Hilbert class polynomial $H_D(X)$ of degree $h(D)$ splits into distinct linear factors in $\mathbb{F}_q[X]$ and its roots form the set $\mathrm{Ell}_{\mathcal{O}}(\mathbb{F}_q)$ of cardinality $h(D)$.

If $\mathcal{O}'$ is an order of discriminant D' that contains $\mathcal{O}$ with index u, then $D = u^2 D'$ and $4q = t^2 - u^2 v^2 D'$, so q is also the norm of a prime ideal in $\mathcal{O}_{L'}$, where L' is the ring class field of $\mathcal{O}'$, and we have $q \perp D'$, since $D' | D$. This implies that $\mathrm{Ell}_{\mathcal{O}'}(\mathbb{F}_q)$ is nonempty and has cardinality $h(D')$, by the same argument used above for $\mathcal{O}$. □

Corollary 22.7. *Let $E/\mathbb{F}_q$ be an elliptic curve with CM by an order $\mathcal{O}$ of discriminant $D \perp q$ in an imaginary quadratic field K, and let $\ell \nmid q$ be prime. If $\ell \nmid [\mathcal{O}_K : \mathcal{O}]$ then E admits $1 + \left(\frac{D}{\ell}\right)$ horizontal ℓ-isogenies and no ascending ℓ-isogenies, and if $\ell \mid [\mathcal{O}_K : \mathcal{O}]$ then E admits no horizontal ℓ-isogenies and one ascending ℓ-isogeny. The number of descending ℓ-isogenies admitted by E over $\mathbb{F}_q$ is either zero or $\ell - \left(\frac{D}{\ell}\right)$, depending on whether $\mathrm{Ell}_{\mathcal{O}'}(\mathbb{F}_q)$ is empty or not, where $\mathcal{O}'$ is the order of index ℓ in $\mathcal{O}$.*

Proof. This follows from Theorem 22.5, Lemma 22.6, and the Deuring lifting theorem (see Theorem 21.15). If $\varphi\colon E \to E'$ is an ℓ-isogeny of CM elliptic curves over $\mathbb{C}$ with $\mathrm{End}(E) = \mathcal{O}$ and $\mathrm{End}(E') = \mathcal{O}'$ and $\mathbb{F}_q$ is a finite field for which the sets $\mathrm{Ell}_{\mathcal{O}}(\mathbb{F}_q)$ and $\mathrm{Ell}_{\mathcal{O}'}(\mathbb{F}_q)$ are both nonempty, then we can view $\varphi\colon E \to E'$ as an isogeny of elliptic curves L, where L is the larger of the two ring class fields for $\mathcal{O}$ and $\mathcal{O}'$ (one must contain the other since either $\mathcal{O} \subseteq \mathcal{O}'$ or $\mathcal{O}' \subseteq \mathcal{O}$ and in both cases we can apply Corollary 21.6), and q is the norm of a prime ideal $\mathfrak{q}$ in $\mathcal{O}_L$. We can use the reduction map $\mathcal{O}_L \to \mathcal{O}_L/\mathfrak{q} = \mathbb{F}_q$ to reduce integral equations for E, E', and φ modulo $\mathfrak{q}$ to obtain a corresponding ℓ-isogeny $\overline{\varphi}\colon \overline{E} \to \overline{E}'$ of elliptic curves

over $\mathbb{F}_q$ with $\mathrm{End}(\overline{E}) = \mathrm{End}(E) = \mathcal{O}$, $\mathrm{End}(\overline{E}') = \mathrm{End}(E') = \mathcal{O}'$, and $\deg\overline{\varphi} = \deg\varphi = \ell$ (the degree of φ cannot change because $\ell \nmid q$, so $E[\ell] \simeq \overline{E}[\ell]$, which implies $\ker\varphi \simeq \ker\overline{\varphi}$, and $\overline{\varphi}$ must be separable).

Conversely, if $\overline{\varphi}\colon \overline{E} \to \overline{E}'$ is an ℓ-isogeny of elliptic curves over $\mathbb{F}_q$, we can lift $\overline{E}$ and $\overline{E}'$ to elliptic curves over L with $\mathrm{End}(E) = \mathrm{End}(\overline{E}) = \mathcal{O}$ and $\mathrm{End}(E') = \mathrm{End}(\overline{E}') = \mathcal{O}'$. There is then a corresponding ℓ-isogeny $\varphi\colon E \to E'$ whose kernel reduces to the kernel of $\overline{\varphi}$ (as above, the reduction map gives a bijection $E[\ell] \simeq \overline{E}[\ell]$ for $\ell \nmid q$). □

If $E/\mathbb{F}_q$ is an elliptic curve with CM by an imaginary quadratic order $\mathcal{O}$ and $\mathfrak{a}$ is a proper $\mathcal{O}$-ideal, then as in Definition 17.13 we have an $\mathfrak{a}$-torsion subgroup

$$E[\mathfrak{a}] := \{P \in E(\overline{\mathbb{F}}_q) : \alpha(P) = 0 \text{ for all } \alpha \in \mathfrak{a}\}.$$

Provided the norm of $\mathfrak{a}$ is prime to q, there is a corresponding separable isogeny $\varphi_\mathfrak{a}\colon E \to E'$ with $\ker\varphi_\mathfrak{a} = E[\mathfrak{a}]$ and $\deg\varphi_\mathfrak{a} = \mathrm{N}\mathfrak{a}$ uniquely determined up to isomorphism, by Theorem 5.11. As in the proof above we can lift the isogeny $\varphi_\mathfrak{a}\colon E \to E'$ to a number field $L \subseteq \mathbb{C}$ where it corresponds to the CM action of $\mathrm{cl}(\mathcal{O})$, which implies that we must have $\mathrm{End}(E') = \mathrm{End}(E) = \mathcal{O}$; if $\mathrm{N}\mathfrak{a}$ is a prime ℓ this means that $\varphi_\mathfrak{a}$ is a horizontal ℓ-isogeny. By Theorem 20.11, every ideal class in $\mathrm{cl}(\mathcal{O})$ contains infinitely many ideals of prime norm, and in particular, an ideal whose norm is prime to q. This allows us to define the CM action of $\mathrm{cl}(\mathcal{O})$ on the set $\mathrm{Ell}_\mathcal{O}(\mathbb{F}_q)$ in terms of horizontal ℓ-isogenies for various primes $\ell \nmid q$. As with the CM action on $\mathrm{Ell}_\mathcal{O}(\mathbb{C})$, the action of the inverse of an ideal $\mathfrak{a}$ is given by the dual isogeny $\hat{\varphi}_\mathfrak{a}$. We thus have the following corollary.

Corollary 22.8. *Let $\mathcal{O}$ be an imaginary quadratic order of discriminant D and let $\mathbb{F}_q$ be a finite field with $q \perp D$. If the set $\mathrm{Ell}_\mathcal{O}(\mathbb{F}_q)$ is nonempty then it is a $\mathrm{cl}(\mathcal{O})$-torsor in which the action of the ideal class of any proper $\mathcal{O}$-ideal of prime norm $\ell \nmid q$ is given by a horizontal ℓ-isogeny, and the inverse of this action is given by the dual isogeny.*

Remark 22.9. As noted above, every ideal class in $\mathrm{cl}(\mathcal{O})$ contains infinitely many proper $\mathcal{O}$-ideals of prime norm ℓ. This means that if we want to compute the action of a given proper $\mathcal{O}$-ideal $\mathfrak{l}_1$ of prime norm ℓ_1, we can compute this action using any other proper $\mathcal{O}$-ideal $\mathfrak{l}_2$ of prime norm ℓ_2 that lies is in the same ideal class. This has many practical applications: when ℓ_1 is large it allows us to use a much smaller ℓ_2. Indeed, under the Generalized Riemann Hypothesis, we can always find a prime ℓ_2 bounded by $O(\log^2|D|)$.

22.2 Isogeny volcanoes

Having determined the exact number of horizontal, ascending, and descending ℓ-isogenies that arise for an ordinary elliptic curve over a finite field, we can now completely determine the structure of the ordinary components of $G_\ell(\mathbb{F}_q)$. Figures 1 and 2 show top and side views of a typical example — with a bit of imagination, one can see a volcano.

Definition 22.10. An *ℓ-volcano* V is a connected undirected graph whose vertices are partitioned into one or more *levels* $V_0, \ldots, V_d$ such that the following hold:

1. The subgraph on V_0 (the *surface*) is a regular graph of degree at most 2.
2. For $i > 0$, each vertex in V_i has exactly one neighbor in level V_{i-1}, and this accounts for every edge not on the surface.
3. For $i < d$, each vertex in V_i has degree $\ell + 1$.

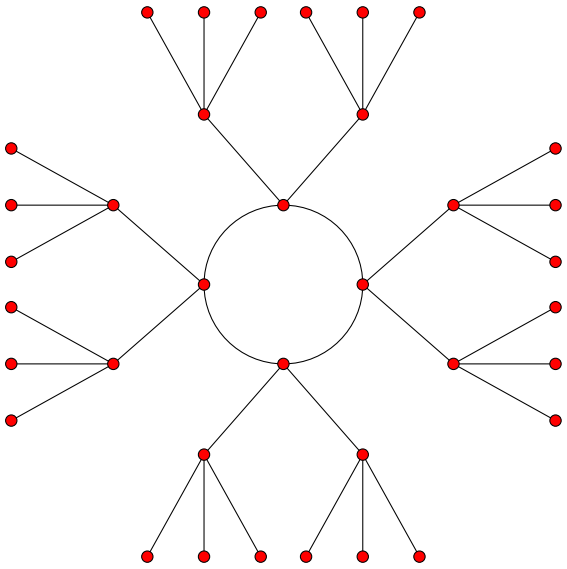


Figure 1: An ordinary component of $G_3(\mathbb{F}_p)$.

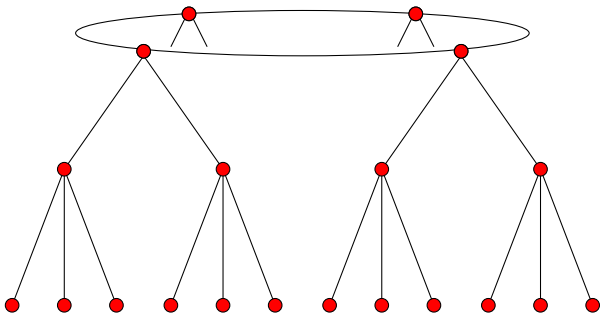


Figure 2: A 3-volcano of depth 2.

Level V_d is called the *floor* of the volcano; the floor and surface coincide when $d = 0$.

As with $G_\ell(\mathbb{F}_q)$, an ℓ-volcano may have multiple edges and self-loops, but it is an undirected graph. If the surface of an ℓ-volcano has more than two vertices, it must be a simple cycle. Two vertices may be connected by 1 or 2 edges, and a single vertex may have 0, 1, or 2 self-loops. As an abstract graph, an ℓ-volcano is determined by the integers ℓ, d, $|V_0|$.

If we ignore components that contain the two exceptional j-invariants 0 and 1728, the ordinary components of $G_\ell(\mathbb{F}_q)$ are all ℓ-volcanoes. This was proved by David Kohel in his Ph.D. thesis [6], although the term "volcano" was coined later by Fouquet and Morain in [3].

Theorem 22.11 (Kohel)**.** *Let $\mathbb{F}_q$ be a finite field, let $\ell \nmid q$ be a prime, and let V be an ordinary component of $G_\ell(\mathbb{F}_q)$ that does not contain the j-invariants 0 or 1728. Then V is an ℓ-volcano for which the following hold:*

(i) *The vertices in level V_i all have the same endomorphism ring $\mathcal{O}_i$.*

(ii) *The subgraph on V_0 has degree $1 + \left(\frac{D_0}{\ell}\right)$, where $D_0 = \operatorname{disc}(\mathcal{O}_0)$.*

(iii) *If $\left(\frac{D_0}{\ell}\right) \geq 0$, then $|V_0|$ is the order of $[\mathfrak{l}]$ in $\operatorname{cl}(\mathcal{O}_0)$; otherwise $|V_0| = 1$.*

(iv) *The depth of V is d, where $4q = t^2 - \ell^{2d} v^2 D_0$ with $\ell \nmid v$, $t^2 = (\operatorname{tr} \pi_E)^2$, for $j(E) \in V$.*

(v) *$\ell \nmid [\mathcal{O}_K : \mathcal{O}_0]$ and $[\mathcal{O}_i : \mathcal{O}_{i+1}] = \ell$ for $0 \leq i < d$.*

Proof. Let V be an ordinary component of $G_\ell(\mathbb{F}_q)$ that does not contain 0 or 1728. The only automorphisms admitted by elliptic curves E with $j(E) \neq 0, 1728$ are $\pm 1 \in \operatorname{End}(E)$, thus as explained in Remark 22.2, every edge (j_1, j_2) in V occurs with the same multiplicity as the edge (j_2, j_1), allowing us to view V as an undirected graph.

Since V is an ordinary component, every vertex is the j-invariant of an ordinary elliptic curve whose endomorphism ring is an order $\mathcal{O}$ in an imaginary quadratic field, by Corollary 13.20. It follows from Theorem 22.3 that the order $\mathcal{O}$ arising for elliptic curves with $j(E) \in V$ all lie in the same quadratic field K and differ only in the ℓ-adic valuation ν_ℓ of the conductor of $[\mathcal{O}_K : \mathcal{O}]$. By Corollary 22.7, every $j(E) \in V$ for which $\operatorname{End}(E) = \mathcal{O}$ has conductor divisible by ℓ admits an ascending ℓ-isogeny, and it follows that we can partition V into levels $V_0, \ldots, V_d$ with $j(E) \in V_i$ if and only if $v_\ell([\mathcal{O}_K : \mathcal{O}]) = i$; the set V is finite so d is bounded; this proves (i) and (v), and Corollary 22.7 also implies (ii) and that V is an ℓ-volcano as claimed.

If $\left(\frac{D_0}{\ell}\right) = -1$ then V_0 has degree 0 and we must have $|V_0| = 1$. Otherwise there exists a proper $\mathcal{O}_0$-ideal $\mathfrak{l}$ of norm ℓ, and its ideal class $[\mathfrak{l}] \in \operatorname{cl}(\mathcal{O})$ acts on V_0 via horizontal ℓ-isogenies, by Corollary 22.8. This proves (iii).

Part (iv) follows from Theorem 21.5 and Remark 21.13. If $4q = t^2 - v^2 \ell^{2d} D_0$ with $\ell \nmid v$, then the sets $\operatorname{Ell}_{\mathcal{O}_i}(k)$ must be non-empty for $0 \leq i \leq d$, but the set $\operatorname{Ell}_{\mathcal{O}_{d+1}}(k)$ must be empty since ℓ^{d+1} does not divide v. □

Remark 22.12. Theorem 22.11 can be extended to the case where V contains 0 or 1728 following Remark 22.2. Parts (i)-(v) still hold, the only necessary modification is the claim that V is an ℓ-volcano. When V contains 0, if V_1 is non-empty then it contains $\frac{1}{3}\left(\ell - \left(\frac{-3}{\ell}\right)\right)$ vertices, and each vertex in V_1 has three incoming edges from 0 but only one outgoing edge to 0. When V contains 1728, if V_1 is non-empty then it contains $\frac{1}{2}\left(\ell - \left(\frac{-1}{\ell}\right)\right)$ vertices, and each vertex in V_1 has two incoming edges from 1728 but only one outgoing edge to 1728. This 3-to-1 (resp. 2-to-1) discrepancy arises from the action of $\operatorname{Aut}(E)$ on the cyclic subgroups of $E[\ell]$ when $j(E) = 0$ (resp. 1728). Otherwise, V satisfies all the requirements

of an ℓ-volcano, and most of the algorithms designed for ℓ-volcanoes work just as well on ordinary components of $G_\ell(\mathbb{F}_q)$ that contain 0 or 1728.

22.3 Finding the floor

The vertices that lie on the floor of an ℓ-volcano V are distinguished by their degree.

Lemma 22.13. *Let v be a vertex in an ordinary component V of depth d in $G_\ell(\mathbb{F}_q)$. Either $\deg v \le 2$ and $v \in V_d$, or $\deg v = \ell + 1$ and $v \notin V_d$.*

Proof. If $d = 0$ then $V = V_0 = V_d$ is a regular graph of degree at most 2 and $v \in V_d$. Otherwise, either $v \in V_d$ and v has degree 1, or $v \notin V_d$ and v has degree $\ell + 1$. □

Given an arbitrary vertex $v \in V$, we would like to find a vertex on the floor of V. Our strategy is very simple: if $v_0 = j(E)$ is not already on the floor then we will construct a random path from v_0 to a vertex v_s on the floor. By a *path*, we mean a sequence of vertices $v_0, v_1, \dots, v_s$ such that each pair (v_{i-1}, v_i) is an edge and $v_i \neq v_{i-2}$ (no backtracking).

Algorithm FindFloor
Given an ordinary vertex $v_0 \in G_\ell(\mathbb{F}_q)$, find a vertex on the floor of its component.

1. If $\deg v_0 \le 2$ then output v_0 and terminate.
2. Pick a random neighbor v_1 of v_0 and set $s \leftarrow 1$.
3. While $\deg v_s > 1$: pick a random neighbor $v_{s+1} \neq v_{s-1}$ of v_s and increment s.
4. Output v_s.

Remark 22.14 (**Removing known roots**)**.** As a minor optimization, rather than picking v_{s+1} as a root of $\phi(Y) = \Phi_\ell(v_s, Y)$ in step 3 of the FindFloor algorithm, we may use $\phi(Y)/(Y - v_{s-1})^e$, where e is the multiplicity of v_{s-1} as a root of $\phi(Y)$. This is slightly faster and eliminates the need to check that $v_{s+1} \neq v_{s-1}$.

Notice that once FindFloor picks a descending edge (one leading closer to the floor), every subsequent edge must also be descending, because it is not allowed to backtrack along the single ascending edge and there are no horizontal edges below the surface. It follows that the expected length of the path chosen by FindFloor is $\delta + O(1)$, where δ is the distance from v_0 to the floor along a shortest path. With a bit more effort we can find a path of exactly length δ, a shortest path to the floor. The key to doing so is to observe that all but at most two of the $\ell + 1$ edges incident to any vertex above the floor must be descending edges. Thus if we construct *three* random paths from v_0 that all start with a different initial edge, then one of the initial edges must be a descending edge, which necessarily leads to a shortest path to the floor.

Algorithm FindShortestPathToFloor
Given an ordinary vertex $v_0 \in G_\ell(\mathbb{F}_q)$, find a shortest path to the floor of its component.

1. Let $v_0 = j(E)$. If $\deg v_0 \le 2$ then output v_0 and terminate.
2. Pick three neighbors of v_0 and extend paths from each of these neighbors in parallel, stopping as soon as any of them reaches the floor.[1]
3. Output a path that reached the floor.

[1] If v_0 does not have three distinct neighbors then just pick all of them.

The main virtue of FindShortestPathToFloor is that it allows us to compute δ, which tells us the level $V_{d-\delta}$ of $j(E)$ relative to the floor V_d. It effectively gives us an "altimeter" $\delta(v)$ that we may use to navigate V. We can determine whether a given edge (v_1, v_2) is horizontal, ascending, or descending, by comparing $\delta(v_1)$ to $\delta(v_2)$, and we can determine the exact level of any vertex.[2]

There are many practical applications of isogeny volcanoes, some of which you will explore on Problem Set 12. See the survey paper [8] for further details and references.

References

[1] Reinier Bröker, Kristin Lauter, and Andrew V. Sutherland, *Modular polynomials via isogeny volcanoes*, Mathematics of Computation **81**, 2012, 1201–1231.

[2] David A. Cox, *Primes of the form $x^2 + ny^2$: Fermat, class field theory, and complex multiplication*, Wiley, 1989.

[3] Mireille Fouquet and François Morain, *Isogeny volcanoes and the SEA algorithm*, Algorithmic Number Theory Fifth International Symposium (ANTS V), LNCS **2369**, Springer, 2002, 276–291.

[4] Sorina Ionica and Antoine Joux, *Pairing the volcano*, Mathematics of Computation **82** (2013), 581–603.

[5] Serge Lang, *Elliptic functions*, second edition, Springer, 1987.

[6] David Kohel, *Endomorphism rings of elliptic curves over finite fields*, PhD thesis, University of California at Berkeley, 1996.

[7] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer, 2009.

[8] Andrew V. Sutherland, *Isogeny volcanoes*, Algorithmic Number Theory 10th International Symposium (ANTS X), Open Book Series **1**, MSP 2013, 507–530.

[2] A more sophisticated approach that uses the Weil pairing (to be discussed in Lecture 23) can be found in [4]; the pairing based approach is more efficient when d is large, but in practice d is usually small.

23 Divisors and the Weil pairing

In this lecture we address a completely new topic, the Weil pairing, which has many practical and theoretical applications. In order to define the Weil pairing we first need to expand our discussion of the function field of a curve from Lecture 4. This requires a few basic results from commutative algebra and algebraic geometry that we will not take the time to prove; almost everything we need is summarized in the first two chapters of Silverman's book [7], which I recommend reviewing if you have not seen this material before.

23.1 Valuations on the function field of a curve

Let C/k be a smooth projective curve defined by a homogeneous polynomial $f_C(x,y,z) = 0$ that (as always) we assume is irreducible over $\bar{k}$.[1] For the sake of simplicity we assume throughout this section that k is a perfect field (every algebraic extension is separable).

In Lecture 4 we defined the *function field* $k(C)$ as the field of rational functions g/h, where $g, h \in k[x,y,z]$ are homogeneous polynomials of the same degree with $h \notin (f_C)$, modulo the equivalence relation

$$\frac{g_1}{h_1} \sim \frac{g_2}{h_2} \qquad \iff \qquad g_1 h_2 - g_2 h_1 \in (f_C).$$

Alternatively, we can view the function g/h as a rational map $(g : h)$ from C to $\mathbb{P}^1$. Our assumption that C is smooth implies that this rational map is actually a *morphism*, meaning that it is defined at every point $P \in C(\bar{k})$; this was stated as Theorem 4.15 and we will prove it below. This means that even though the rational map $(g_1 : h_1) : C \to \mathbb{P}^1$ associated to a particular representative g_1/h_1 of an element of $k(C)$ might not be defined at a point P (this occurs when $g_1(P) = h_1(P) = 0$, since $(0:0)$ is not a point in $\mathbb{P}^1$), there is always an equivalent g_2/h_2 representing the same element of $k(C)$ that *is* defined at P.

Example 23.1. Consider the function x/z on the elliptic curve $E\colon y^2z = x^3 + Axz^2 + Bz^3$. We can evaluate the map $(x : z)$ at any affine point, but not at the point $(0:1:0)$, where we get $(0:0)$. But the maps

$$(x : z) \sim (x^3 : x^2z) \sim (y^2z - Axz^2 - Bz^3 : x^2z) \sim (y^2 - Axz - Bz^2 : x^2)$$

all represent the same element of $k(E)$, and the last one sends $(0:1:0)$ to $(1:0) \in \mathbb{P}^1$, which is defined. Moreover, any other representative of the function x/z that is defined at $(0:1:0)$ will give the same value. Notice that the right-most map is also not defined everywhere, since it gives $(0:0)$ at the point $(0 : \sqrt{B} : 1)$. In general, there will typically not be a single representative for a function $f \in k(E)$ that can be used to evaluate the morphism $f\colon E \to \mathbb{P}^1$ at every point, even though the morphism is defined at every point.

Remark 23.2. It is often more convenient to write elements of the function field in affine form, just as we typically use the equation $y^2 = x^3 + Ax + B$ to refer to the projective curve defined by its homogenization; so we may write x instead of x/z, for example. In general, any time we refer to a function $r(x,y)$ as an element of $k(C)$ that is not a ratio

[1] Here we are assuming for simplicity that C is a plane curve (e.g. an elliptic curve in Weierstrass form). One can work more generally in $\mathbb{P}^n$ by replacing (f) with a homogeneous ideal I in $k[x_0,\dots,x_n]$ whose zero locus is a smooth absolutely irreducible projective variety of dimension one in $\mathbb{P}^n$. Everything in this section applies to any smooth projective (geometrically integral) curve, we use plane curves only for the sake of concreteness.

$g(x,y,z)/h(x,y,z)$ of two homogeneous polynomials g and h of the same degree, it should be understood that we mean the function one obtains by multiplying each term in the numerator and denominator of $r(x,y)$ by a suitable power of z to put it in the form g/h with g and h homogeneous polynomials of the same degree.

Definition 23.3. For any point $P \in C(\bar{k})$, we define the *local ring at* P (or the *ring of regular functions at* P) by

$$\mathcal{O}_P := \{f \in k(C) : f(P) \neq \infty\} \subseteq k(C),$$

where $\infty = (1:0) \in \mathbb{P}^1$. It is a principal ideal domain (PID) with a unique maximal ideal

$$\mathfrak{m}_P := \{f \in \mathcal{O}_P : f(P) = 0\}.$$

Any generator u_P for the principal ideal $\mathfrak{m}_P = (u_P)$ is called a *uniformizer* at P.

Definition 23.4. A *discrete valuation* on a field F is a surjective homomorphism $v\colon F^\times \to \mathbb{Z}$ that satisfies

$$v(x+y) \geq \min(v(x), v(y))$$

for all $x, y \in F^\times$ with $x+y \neq 0$ (one typically defines $v(0) = \infty$). If v is a discrete valuation on F, then the subring

$$R := \{x \in F : v(x) \geq 0\}$$

is a PID with the unique maximal ideal

$$\mathfrak{m} := \{x \in R : v(x) \geq 1\}.$$

Every nonzero ideal (x) of R is then of the form $\mathfrak{m}^n$, where $n = v(x)$. Any $u \in F$ with $v(u) = 1$ is a generator for $\mathfrak{m}$ and called a *uniformizer* for $\mathfrak{m}$.

Given a principal ideal domain R with a unique nonzero maximal ideal $\mathfrak{m} = (u)$, we can define a discrete valuation on its fraction field F via

$$v(x) := \min\{n \in \mathbb{Z} : u^{-n}x \in R\},$$

and we then have $R = \{x \in F : v(x) \geq 0\}$. Note that $v(x)$ does not depend on the choice of the uniformizer u. We call any such ring R a *discrete valuation ring* (DVR).

For the curve C/k, the local rings $\mathcal{O}_P$ are a family of DVRs that all have the same fraction field $k(C)$. We thus have a discrete valuation v_P for each point $P \in C(\bar{k})$ which we think of as measuring the "order of vanishing" of a function $f \in k(C)$ at P (one can formally expand f as a Laurent series in any uniformizer u_P for $\mathfrak{m}_P$, and the degree of the first nonzero term will be $v_P(f)$, just as with meromorphic functions over $\mathbb{C}$).

Remark 23.5. When k is not algebraically closed the function field $k(C)$ has many valuations that are not associated to rational points $P \in C(k)$. One can always work with $\bar{k}$-points as above (and in [7]), but a more natural approach is to work with *closed points*: $\mathrm{Gal}(\bar{k}/k)$-orbits in $C(\bar{k})$, which we also denote P (note that we have assumed k is a perfect field, so $\bar{k}/k$ is separable). Each closed point is a finite subset of $C(\bar{k})$ whose cardinality we denote $\deg P$; this is the same as the degree of the minimal extension of k over which all the points in P are defined (which is necessarily a finite Galois extension), and it is also the degree of the residue field $\mathcal{O}_P/\mathfrak{m}_P$ as an extension of k. Rational points (elements of $C(k)$) are closed points of degree one. Each closed point corresponds to a maximal ideal $\mathfrak{m}_P$ of the coordinate ring $k[C]$. Note that it still makes sense to "evaluate" a rational function $f \in k(C)$ at a closed point P; the result is a closed point $f(P)$ of $\mathbb{P}^1$ (because $f \in k(C)$ is, by definition, Galois invariant).

Now that we have valuations v_P and uniformizers u_P associated to each point P of a smooth projective curve we can easily prove Theorem 4.15, which was stated without proof.

Theorem 23.6. *Let C_1/k be a smooth projective curve and let $\phi\colon C_1 \to C_2$ be a rational map. Then ϕ is a morphism.*

Proof. Let $\phi = (\phi_0 : \cdots : \phi_m)$, let $P \in C_1(\bar{k})$ be any point, let u_P be a uniformizer at P, and let $n = \min_i v_P(\phi_i)$. Then

$$\phi = (u_P^{-n}\phi_0 : \cdots : u_P^{-n}\phi_m)$$

is defined at P because $v_P(u_P^{-n}\phi_i) \geq 0$ for all i and $v_P(u_P^{-n}\phi_i) = 0$ for at least one i. □

Remark 23.7. When C_1 is not smooth one can construct counterexamples to the theorem above. Smoothness guarantees that the local rings $\mathcal{O}_P$ are all DVRs, so that we have a valuation v_P to work with. Indeed, a curve is smooth if and only if all its local rings are DVRs; this gives an alternative criterion for smoothness that does not depend on the equation of the curve or even the dimension of the projective space in which it is embedded.

Example 23.8. For the function x on the elliptic curve $E\colon y^2 = x^3 + Ax + B$ we have

$$v_P(x) = \begin{cases} 0 & \text{if } P = (1 : * : *) \\ 1 & \text{if } P = (0 : \pm\sqrt{B} : 1) \quad (B \neq 0) \\ 2 & \text{if } P = (0 : 0 : 1) \qquad (B = 0) \\ -2 & \text{if } P = (0 : 1 : 0) \end{cases}$$

For the function y we have

$$v_P(y) = \begin{cases} 0 & \text{if } P = (* : y_0 : 1) \qquad (y_0 \neq 0) \\ 1 & \text{if } P = (x_0 : 0 : 1) \qquad (x_0^3 + Ax_0 + B = 0) \\ -3 & \text{if } P = (0 : 1 : 0) \end{cases}$$

You may wonder how we computed these valuations. In particular, how do we know that $v_\infty(x) = -2$ and $v_\infty(y) = -3$? There are a couple of ways to see this. One is to use the fact that for any $f \in k(C)$ we always have $\sum_P v_P(f) = 0$ (see below), so every function in $k(C)$ has the same number of zeros and poles. Thus if we know all the zeros (and the order of vanishing at each) and there is only one pole, we know its order.

A more general approach is to consider the *degree* of the morphism $f\colon C \to \mathbb{P}^1$. For non-constant functions f this is defined as

$$\deg f := [k(C) : f^*(k(\mathbb{P}^1))]$$

where $f^*\colon k(\mathbb{P}^1) \to k(C)$ is the morphism of function fields that sends $g \in k(\mathbb{P}^1)$ to the function $g \circ f$ in $k(C)$; for $f \in k^\times$ the convention is to define $\deg f = 0$. In explicit examples it is often obvious what the degree is, it is the cardinality of the fibers $f^{-1}(P)$ for all but finitely many $P \in \mathbb{P}^1(\bar{k})$. In our example, the function x defines a morphism of degree two from E to $\mathbb{P}^1$, because if we pick an arbitrary point on $\mathbb{P}^1$ there will generically be two points on E that get mapped to it (points with the same x-coordinate). Any time this is not the case, we have a *ramified* point, and in the case of a zero or pole the degree of ramification is what determines its multiplicity.

Whenever we have $f(P) = Q \in \mathbb{P}^1(\bar{k})$ and the size of the preimage $f^{-1}(Q)$ is the same as the degree of f as a morphism (which happens for all but finitely many Q), no ramification occurs and if $Q = 0$ or $Q = \infty$ then f has a simple zero or pole at P. More generally, we have the following theorem, which says that so long as we count points with multiplicity, every fiber of the morphism $f\colon C \to \mathbb{P}^1$ has the same size, equal to the degree of f.

Theorem 23.9. *Let C be a smooth projective curve over an algebraically closed field k and let $f \in k(C)^\times$ be an element of its function field (viewed as a morphism $f\colon C \to \mathbb{P}^1$). For every point $Q \in \mathbb{P}^1(k)$ we have*

$$\deg f = \sum_{f(P)=Q} v_P(u_Q \circ f),$$

where $u_Q \in k(\mathbb{P}^1)$ denotes any uniformizer for $\mathfrak{m}_Q$.

Proof. This is a special case of Proposition 2.6 in [7]. □

If t is our coordinate for $\mathbb{P}^1$ (which we may view as taking values in $k \cup \{\infty\}$), then we can take $u_Q := t - Q$ to be a simple translation. Computing $v_P(u_Q \circ f)$ then amounts to re-interpreting the order of "vanishing" at P with the order of "Q-ing" at P.

Corollary 23.10. *Let C be a smooth projective curve over an algebraically closed field k. For every $f \in k(C)^\times$ we have*

$$\sum_{P \in C(k)} v_P(f) = 0,$$

and $v_P(f) = 0$ for all but finitely many P; we have $v_P(f) = 0$ for all P if and only if $f \in k^\times$.

Proof. We have $v_P(f) \neq 0$ only when $f(P) = 0$ or $f(P) = \infty$. Applying Theorem 23.9 to $Q = 0$ using the uniformizer $u_0 = t$ yields

$$\deg f = \sum_{f(P)=0} v_P(f),$$

and if we apply it to $Q = \infty$ with uniformizer $u_\infty = 1/t$ we have

$$\deg f = \sum_{f(P)=\infty} v_P(u_\infty \circ f) = \sum_{f(P)=\infty} -v_P(f),$$

which implies $\sum v_P(f) = 0$. The cardinalities of $f^{-1}(0)$ and $f^{-1}(\infty)$ are each bounded by $\deg f$, hence finite, so $v_P(f) \neq 0$ for only finitely many P, and these cardinalities can be zero if and only if $f \in k^\times$, since otherwise $\deg f \geq 1$. □

Remark 23.11. When working with closed points over a non-algebraically closed field the formula in Theorem 23.9 needs to be modified to account for the degrees of the points. We then have

$$\deg f \deg Q = \sum_{f(P)=Q} v_P(u_Q \circ f) \deg P,$$

which holds for any closed point Q of $\mathbb{P}^1/k$; the formula in Corollary 23.10 becomes

$$\sum v_P(f) \deg P = 0,$$

where the sum is over closed points P.

Example 23.12. Another way to compute valuations is to work directly from the definition using a uniformizer u_P. We did not do this in Example 23.8 because we hadn't yet determined uniformizers for the points on an elliptic curve. But from the example it is clear that we can take

$$u_P = \begin{cases} x - x(P) & \text{if } y(P) \neq 0 \text{ and } P \neq (0:1:0) \\ y & \text{if } y(P) = 0 \\ x/y & \text{if } P = (0:1:0) \end{cases}$$

Note that $v_P(x/y) = v_P(x) - v_P(y) = -2 - (-3) = 1$. To check that $v_\infty(y) = -3$ using the uniformizer u_∞, for example, it suffices to show that $1/y$ and u_∞^3 generate the same ideal in $\mathcal{O}_\infty$: the function $s := y^2/x^3 = y^2/(y^2 - Ax - B)$ is a unit in $\mathcal{O}_\infty$ and we have $1/y = su_\infty^3$.

23.2 The divisor class group of a curve

As in the previous section, we continue to assume that C is a smooth projective curve over a perfect field k, and in this subsection we will temporarily assume k is algebraically closed (but we may still write $\bar{k}$ to emphasize this in places where it is especially important).

Definition 23.13. To each point $P \in C(\bar{k})$ we associate a formal symbol $[P]$. The *divisor group* of C is the free abelian group on the set $\{[P] : P \in C(\bar{k})\}$, denoted $\operatorname{Div} C$. Its elements are called *divisors*. Each is a finite sum of the form

$$D = \sum_P n_P [P]$$

in which the n_P are integers (so $n_P = 0$ for all but finitely many P).

Remark 23.14. Some authors write P rather than $[P]$ and rely on context to make it clear which is meant, but when C is an elliptic curve this makes it very hard to know whether $P + Q$ is meant to denote $[P] + [Q]$ or $[P + Q]$ (these two divisors are equivalent in a sense to be defined, but they are not the same divisor). It is also common to use (P) rather than $[P]$, but this can cause confusion when similar symbols are used for divisor coefficients and elements of $k(C)$, so we will avoid this notation.

The integer n_P is the *valuation* of D at P, also denoted by $v_P(D) := n_P$. For each divisor D the finite set

$$\operatorname{supp}(D) := \{P : v_P(D) \neq 0\}$$

is its *support*, and the integer

$$\deg D := \sum_P v_P(D)$$

is its *degree*. The degree map $D \mapsto \deg D$ is a surjective homomorphism of abelian groups whose kernel is the subgroup $\operatorname{Div}^0 C$ of divisors of degree zero. Associated to each function $f \in k(C)^\times$ there is a divisor

$$\operatorname{div} f := \sum_P v_P(f)[P],$$

which is called a *principal divisor*. Because each $v_P \colon k(C)^\times \to \mathbb{Z}$ is a group homomorphism, we have $\operatorname{div} fg = \operatorname{div} f + \operatorname{div} g$, and the map

$$\operatorname{div} \colon k(C)^\times \to \operatorname{Div} C$$

is a group homomorphism whose image is the subgroup of $\operatorname{Div} C$ consisting of principal divisors, and whose kernel consists of the nonzero constant functions $k^\times$, by Corollary 23.10.

The quotient group

$$\operatorname{Pic} C := \operatorname{Div} C / \operatorname{div}(k(C)^\times)$$

is the *Picard group* of C. We also have a degree map

$$\deg\colon \operatorname{Pic} C \to \mathbb{Z}$$

on divisor classes, and its kernel is the *reduced Picard group*

$$\operatorname{Pic}^0 C := \operatorname{Div}^0 C / \operatorname{div}(k(C)^\times)$$

also known as the (degree zero) *divisor class group* of C. We then have an exact sequence

$$1 \longrightarrow k^\times \longrightarrow k(C)^\times \longrightarrow \operatorname{Div}^0 C \longrightarrow \operatorname{Pic}^0 C \longrightarrow 0.$$

When $k \neq \bar{k}$ one instead defines divisors as sums over closed points (each of which is a Galois orbit of $\bar{k}$-points). The degree of a divisor is then $\deg D := \sum_P v_P(D) \deg P$, and if we let $\operatorname{Gal}(\bar{k}/k)$ act on $\operatorname{Div} C$ by defining $\sigma([P]) := [\sigma(P)]$ for $\sigma \in \operatorname{Gal}(\bar{k}/k)$, then the $\operatorname{Gal}(\bar{k}/k)$-invariant elements of $\operatorname{Div} C_{\bar{k}}$ and $\operatorname{Div}^0 C_{\bar{k}}$ are precisely those that are sums of closed points. So long as $C(k) \neq \emptyset$ (necessarily true when C is an elliptic curve), the $\operatorname{Gal}(\bar{k}/k)$-invariant elements of $\operatorname{Pic}^0 C_{\bar{k}}$ will all be represented by $\operatorname{Gal}(\bar{k}/k)$-invariant elements of $\operatorname{Div}^0 C_{\bar{k}}$ and we obtain the same exact sequence as above whether we start over $\bar{k}$ and take $\operatorname{Gal}(\bar{k}/k)$-invariants of each term or simply work over k (using closed points) throughout. But this is not automatic! In general it is possible for a coset of $\operatorname{div}(\bar{k}(C)^\times)$ in $\operatorname{Div}^0 C_{\bar{k}}$ to be $\operatorname{Gal}(\bar{k}/k)$-invariant without being a coset of $\operatorname{div}(k(C)^\times)$.

Remark 23.15. The condition $C(k) \neq \emptyset$ ensuring that every k-rational divisor class is represented by a k-rational divisor is stronger than necessary, any k-rational divisor of degree 1 will suffice, and such a divisor might involve closed points of higher degree.

Of the various groups defined above, the divisor class group $\operatorname{Pic}^0 C$ is the one of greatest interest to us, because it is intimately related to the curve C. Provided $C(k) \neq \emptyset$, the divisor class group $\operatorname{Pic}^0 C$ is isomorphic to the *Jacobian* of the curve C. Although this is not at all obvious from the definition above, in addition to its structure as an abelian group, $\operatorname{Pic}^0 C$ can be given the structure of an algebraic variety, making it an *abelian variety*. When $C(k) \neq \emptyset$ this variety is canonically isomorphic to the Jacobian of C in the category of abelian varieties over k. The formal definition of these objects is outside the scope of this course, but the details do not matter to us, because when C is an elliptic curve E we already know exactly what its Jacobian is: it is the abelian variety E.

Definition 23.16. Let C/k be a smooth projective curve with a rational point $0 \in C(k)$. The *Abel-Jacobi map* is the map $C(k) \to \operatorname{Pic}^0 C$ defined by

$$P \mapsto [P] - [0].$$

Although we will not prove this here, for a curve C/k of genus g, over an algebraically closed field the Abel-Jacobi map is surjective if and only if $g \leq 1$ and injective if and only if $g \geq 1$. As usual, genus $g = 1$ is the sweet spot, and we will prove in the next section that for smooth projective curves of genus 1 with a rational point (elliptic curves), the Abel-Jacobi map is an isomorphism.

23.3 The Jacobian of an elliptic curve

Definition 23.17. Let E/k be an elliptic curve with 0 as its distinguished point (for curves in Weierstrass form this is the projective point $(0:1:0)$, the point "at infinity"). For each pair of points $P, Q \in E(k)$ let $L_{P,Q} \in k(E)$ denote the function corresponding to the line $\overline{PQ}$, which we define as the tangent to the curve when $P = Q$. For example, if $P = (x_1, y_1)$ and $Q = (x_2, y_2)$ are distinct affine points then the point-slope formula tells us that

$$L_{P,Q} = (y - y_1)(x_2 - x_1) - (x - x_1)(y_2 - y_1),$$

which has zeros at P, Q, and $-(P + Q)$ where it intersects the curve E, but here we are thinking of $L_{P,Q} \in k(E)$ as a map $E \to \mathbb{P}^1$ that we can evaluate at any point R on E. We now define

$$G_{P,Q} := \frac{L_{P,Q}}{L_{P+Q,-(P+Q)}}.$$

The motivation for this is that $G_{P,Q}$ effectively encodes our geometric definition of the group law on E: to add P and Q we construct the line $\overline{PQ}$, which intersects the curve E at a third point $-(P+Q)$, and we then compute $P+Q$ as the point on the line through 0 and $-(P+Q)$; in the formula for $G_{P,Q}$ above this is the line $L_{P+Q,-(P+Q)}$ in the denominator.

To see this more clearly, let us compute the principal divisors corresponding to the functions $L_{P,Q}$ and $G_{P,Q}$. By definition, the function $L_{P,Q}$ has zeros at the points P, Q and $-(P + Q)$ (possibly with multiplicity if any of these points coincide); it has no other zeros and no poles at any affine points, so it must have a triple point at the point at infinity. Thus

$$\operatorname{div} L_{P,Q} = [P] + [Q] + [-(P + Q)] - 3[0].$$

We can then compute

$$\begin{aligned} \operatorname{div} G_{P,Q} &= [P] + [Q] + [-(P + Q)] - 3[0] - ([P + Q] + [-(P + Q)] + [0] - 3[0]) \\ &= [P] + [Q] - [P + Q] - [0]. \end{aligned}$$

Since $\operatorname{div} G_{P,Q}$ is a principal divisor, it follows that $[P] + [Q]$ and $[P + Q] + [0]$ represent the same equivalence class in $\operatorname{Pic} E$; such divisors are said to be *linearly equivalent*, and we use

$$[P] + [Q] \sim [P + Q] + [0] \tag{1}$$

to denote this relation.

Theorem 23.18. *Let E/k be an elliptic curve with distinguished point 0. The Abel-Jacobi map $E(k) \to \operatorname{Pic}^0 E$ defined by $P \mapsto [P] - [0]$ is a group isomorphism.*

Proof. By (1) we have

$$([P] - [0]) + ([Q] - [0]) \sim [P + Q] + [0] - 2[0] = [P + Q] - [0],$$

and clearly $[0] - [0] = 0$, so the Abel-Jacobi map is a group homomorphism.

To show surjectivity, let $D = \sum n_P[P]$ represent a divisor class in $\operatorname{Pic}^0 E$. By splitting D into separate sums with $n_P > 0$ and $n_P < 0$, we can write

$$D = \sum_{n_P > 0} n_P[P] - \sum_{n_P < 0} (-n_P)[P],$$

and by applying (1) repeatedly we obtain

$$D \sim \left[\sum_{n_P>0} n_P P\right] - \left[\sum_{n_P<0} (-n_P)P\right] + m[0],$$

for some integer m (note that the sums $\sum n_P P$ and $\sum(-n_P)P$ inside the brackets are sums of points in $E(k)$ that yield a single point in $E(k)$ in each case). Since D represents a class in $\operatorname{Pic}^0 E$, we have $\deg D = 0$, and computing degrees of both sides above yields

$$0 = 1 - 1 + m,$$

so $m = 0$. If we now let $Q = \sum_{n_P>0} n_P P$ and $R = \sum_{n_P<0}(-n_P)P$ then

$$D \sim [Q] - [R] = [Q] - [0] - ([R] - [0]) = [Q - R] - [0],$$

where we have used the fact that the Abel-Jacobi map is a group homomorphism to get the rightmost equality, which shows that D is in the image of the Abel-Jacobi map, which is thus surjective.

To show injectivity we need to show that the kernel of the Abel-Jacobi map is trivial, which amounts to showing that if $D = \sum n_P[P]$ is a principal divisor, then $\sum n_P P = 0$. As above, by applying (1) repeatedly we can obtain $D \sim [Q] - [R]$. By adding $G_{R,-Q}$ and negating, we obtain the principal divisor $[T] - [0]$, where $T = Q - R$.

We claim that $T = 0$, which implies $Q = R$ and therefore $\sum n_P P = 0$ as desired. Suppose not. Let $t \in k(E)^\times$ be a function with $\operatorname{div} t = [T] - [0]$ (in fact no such functions exist, we are supposing that $[T] - [0]$ is a principal divisor with $T \neq 0$ and this is going to lead to a contradiction). For any $f \in k(E)^\times - k^\times$, define

$$\tilde{f} := \prod_S (t - t(S))^{v_S(f)}$$

If f does not have a zero or pole at 0, then f and $\tilde{f}$ have the same divisor and f is a rational function of t. If f has a zero or pole at 0, we can replace f by $ft^{-v_0(f)}$, which does not have a zero or pole at 0, and we again find that f is a rational function of t. Thus every function in $k(E)$ is a rational function of t, so $k(E) = k(t)$. But $k(t) \simeq k(\mathbb{P}^1)$ and $\mathbb{P}^1$ has genus 0 while E has genus 1, a contradiction, so $T = 0$ as claimed. □

Corollary 23.19. *Let E/k be an elliptic curve and let $D = \sum_P n_P[P] \in \operatorname{Div}(E)$. Then D is a principal divisor if and only if $\sum n_P = 0$ and $\sum_P n_P P = 0$.*

Proof. It suffices to show that $D = \sum_P n_P[P] \in \operatorname{Div}^0(E)$ is principal if and only if $\sum_P n_P P = 0$. Let $\varphi\colon \operatorname{Pic}^0(E) \xrightarrow{\sim} E(\bar{k})$ denote the inverse of the Abel-Jacobi map. Then

$$D \sim 0 \iff \varphi(D) = 0 \iff \sum_P n_P \varphi([P]) = 0 \iff \sum_P n_P \varphi([P] - [0]) = 0 \iff \sum_P n_P P = 0,$$

where we have used $\sum_P n_P = 0$ to subtract $\sum_P n_P \varphi([0]) = 0$ in the third equivalence. □

23.4 Pullback maps

Let $\alpha\colon E \to E'$ be a separable isogeny of elliptic curves over k. Recall from Remark 4.32 that α induces a morphism of function fields

$$\begin{aligned} \alpha^*\colon k(E') &\to k(E) \\ f &\mapsto f \circ \alpha \end{aligned}$$

that allows us to view $k(E)$ as a separable field extension of $\alpha^*(k(E'))$ of degree equal to $\deg \alpha$. The isogeny α also induces a homomorphism of divisor groups defined by

$$\begin{aligned} \alpha^*\colon \operatorname{Div} E' &\to \operatorname{Div} E \\ [P] &\mapsto \sum_{\alpha(Q)=P} [Q], \end{aligned}$$

that is, given a divisor in $\operatorname{Div} E'$, replace each formal symbol corresponding to a point (or closed point) P with the sum of the formal symbols of each of its preimages under α. The two morphisms denoted α^* are compatible in that for every $f \in k(E')$ we have

$$\operatorname{div}(\alpha^*(f)) = \alpha^*(\operatorname{div} f).$$

Corollary 23.19 implies that α^* maps principal divisors to principal divisors, and it clearly maps divisors of degree 0 to divisors of degree 0, thus induces a homomorphism $\alpha^*\colon \operatorname{Pic}^0 E' \to \operatorname{Pic}^0 E$. When composed with the isomorphisms $E' \simeq \operatorname{Pic} E'$ and $\operatorname{Pic} E \simeq E$ given by the Abel-Jacobi maps, this yields the dual isogeny $\hat{\alpha}$. If $\beta\colon E' \to E''$ is another separable isogeny then $(\beta \circ \alpha)^* = \alpha^* \circ \beta^*$, consistent with $\widehat{\beta \circ \alpha} = \hat{\alpha} \circ \hat{\beta}$.

We also recall that each $P \in E(k)$ has an associated translation-by-P map τ_P that induces an automorphism τ_P^* of $k(E)$ defined by $f \mapsto f \circ \tau_P$. When α is separable and $E[\alpha] := \ker \alpha \subseteq E(k)$, each $P \in \ker \alpha$ induces an automorphism of $k(E)$ that fixes $\alpha^*(k(E'))$, making $k(E)/\alpha^*(k(E'))$ a Galois extension with Galois group $E[\alpha]^* = \{\tau_P^* : P \in \ker \alpha\}$. We may then view $\alpha^*(k(E'))$ as the fixed field of $k(E)$ under the action of $(\ker \alpha)^*$. It follows that the functions $g \in k(E)$ of the form $f \circ \alpha$ for some $f \in k(E')$ are precisely those for which $g \circ \tau_P = g$ for all $P \in E[\alpha]$.

We will apply these facts in the next section in the special case $E' = E$ with $\alpha \in \operatorname{End}(E)$ a multiplication-by-n map for some n not divisible by the characteristic of k.

Remark 23.20. When α is not necessarily separable the pullback map on divisors is defined by $[P] \mapsto \sum_{\alpha(Q)=P} e_\alpha(Q)[Q]$ where $e_\alpha(Q) := v_Q(\alpha^* u_P)$ is the *ramification index*, computed using a uniformizer u_P at $P = \alpha(Q)$. When α is separable we always have $e_\alpha(Q) = 1$.

23.5 The Weil pairing

Let E/k be an elliptic curve and let n be a positive integer not divisible by the characteristic of k. We will temporarily assume that $E[n] \subseteq E(k)$, which can be achieved by taking the base change of E to $k(E[n])$ if needed. For each $Q \in E[n]$ we define the divisor

$$D_Q := [n]^*([Q] - [0]) = \sum_{nQ'=Q} [Q'] - \sum_{nP=0} [P].$$

Note that our assumption $E[n] \subseteq E(k)$ ensures that D_Q is a k-rational divisor even if the Q' are not k-rational. For any $T \in E(\bar{k})$, if $nS_0 = T$ then

$$\sum_{nS=T} S = \sum_{R \in E[n]} (S_0 + R) = n^2 S_0 = nT,$$

and it follows that $\sum_{nQ'=Q} Q' = nQ = 0$ and $\sum_{P \in E[n]} P = 0$. Corollary 23.19 then implies that D_Q is a principal divisor, so there exists $g_Q \in k(E)^\times$ such that

$$\operatorname{div}(g_Q) = D_Q,$$

which is unique up to a scalar in $k^\times$. If we compose g_Q with a translation-by-P map τ_P for some $P \in E[n]$, this will not change its divisor, since the sums defining D_Q are both invariant under translation-by-P for any n-torsion point P. However, we may get a different element of $k(E)^\times$, which differs from g_Q by a scalar in $k^\times$. It is thus natural to consider the ratio

$$e_n(P,Q) := \frac{g_Q \circ \tau_P}{g_Q},$$

which does not depend on the choice of g_Q.

Warning 23.21. Some authors define $e_n(P,Q)$ as $\frac{g_P \circ \tau_Q}{g_P}$, reversing the roles of P and Q; see [8, §16] for example (but note that there is a typo in [8, Def. 16.2.1] which is corrected in [8, Eq.16.3]). By Theorem 23.23 below, this amounts to replacing $e_n(P,Q)$ with $e_n(Q,P) = e_n(P,Q)^{-1}$. Our definition matches the one used in Silverman [7].

Now consider $f_Q \in k(E)^\times$ with $\operatorname{div} f_Q = n[Q] - n[0]$ (via Corollary 23.19). We have

$$\operatorname{div}(f_Q \circ [n]) = n \sum_{nQ'=Q} [Q'] - n \sum_{nP=0} [P] = nD_Q = \operatorname{div}(g_Q^n)$$

and choose f_Q so $f_Q \circ [n] = g_Q^n$. It follows that $g_Q^n \in [n]^*(k(E))$ is invariant under composition with τ_P for all $P \in E[n]$, so $e_n(P,Q) = (g_Q \circ \tau_P)/g_Q \in k^\times$ is an nth root of unity.

Let $\mu_n \simeq \mathbb{Z}/n\mathbb{Z}$ denote the group of nth roots of unity in $\bar{k}^\times$. We now drop our assumption that $E[n] \subseteq E(k)$ and base change as required to define $e_n \colon E[n] \times E[n] \to \mu_n$.

Definition 23.22. Let E/k be an elliptic curve. For each $n \geq 1$ not divisible by the characteristic of k we define the *Weil pairing* to be the function $e_n \colon E[n] \times E[n] \to \mu_n$.

Theorem 23.23. *Let E/k be an elliptic curve and let m and n be positive integers not divisible by the characteristic of k. The Weil pairing satisfies the following properties.*

- *Bilinear: $e_n(P+Q,R) = e_n(P,R)e_n(Q,R)$ and $e_n(P,Q+R) = e_n(P,Q)e_n(P,R)$;*
- *Alternating: $e_n(P,P) = 1$ and $e_n(P,Q) = e_n(Q,P)^{-1}$;*
- *Nondegenerate: If $Q \neq 0$ then $e_n(P,Q) \neq 1$ for some $P \in E[n]$;*
- *Galois-equivariant: $e_n(P^\sigma, Q^\sigma) = e_n(P,Q)^\sigma$ for all $\sigma \in \operatorname{Gal}(\bar{k}/k)$;*
- *Compatibility: $e_{mn}(P,Q) = e_n(mP,Q)$ for all $P \in E[mn]$ and $Q \in E[n]$;*
- *Endomorphisms: $e_n(\phi(P),\phi(Q)) = e_n(P,Q)^{\deg \phi}$ for nonzero $\phi \in \operatorname{End}(E)$;*
- *Isogenies: $e_n(P, \hat{\alpha}(Q)) = e_n(\alpha(P), Q)$ for $\alpha \in \operatorname{Hom}(E,E')$, $P \in E[n]$, $Q \in E'[n]$;*
- *Surjective: for each $P \in E[n]$ we have $\{e_n(P,Q) : Q \in E[n]\} = \mu_r$, where $r := |P|$.*

Proof. We first note that our assumption that m and n are not divisible by the characteristic of k ensures the $\#E[m] = m^2$, $\#E[n] = n^2$, and $E[mn] = m^2n^2$, by Theorem 5.25.

Bilinear: We have

$$e_n(P+Q,R) = \frac{g_R \circ \tau_{P+Q}}{g_R} = \left(\frac{g_R \circ \tau_P}{g_R} \circ \tau_Q\right) \frac{g_R \circ \tau_Q}{g_R} = e_n(P,R)e_n(Q,R)$$

since the ratio $g_R \circ \tau_P / g_R$ lies in $\bar{k}^\times$ and is therefore invariant under composition. We now apply Corollary 23.19 to obtain $h \in \bar{k}(E)^\times$ with

$$\operatorname{div} h = [Q+R] - [Q] - [R] + [0]$$

so that $\operatorname{div} f_{Q+R} - \operatorname{div} f_Q - \operatorname{div} f_R = n[Q+R] - n[Q] - n[R] + n[0] = n \operatorname{div} h$. Then

$$\operatorname{div} \frac{g_{Q+R}}{g_Q g_R} = \frac{1}{n} \operatorname{div} \left(\frac{g_{Q+R}^n}{g_Q^n g_R^n}\right) = \frac{1}{n} \operatorname{div} \left(\frac{f_{Q+R}}{f_Q f_R} \circ [n]\right) = \operatorname{div}(h \circ [n]).$$

It follows that $g_{Q+R}/(g_Q g_R)$ is a scalar multiple of $h \circ [n]$, hence invariant under composition with τ_P for $P \in E[n]$, so $g_Q g_R (g_{Q+R} \circ \tau_P) = g_{Q+R}(g_Q g_R \circ \tau_P)$. We then have

$$\begin{aligned} e_n(P,Q+R) &= \frac{g_{Q+R} \circ \tau_P}{g_{Q+R}} = \left(\frac{g_Q g_R}{g_Q g_R}\right)\left(\frac{g_{Q+R} \circ \tau_P}{g_{Q+R}}\right) = \left(\frac{g_{Q+R}}{g_{Q+R}}\right)\left(\frac{g_Q g_R \circ \tau_P}{g_Q g_R}\right) \\ &= \frac{g_Q g_R \circ \tau_P}{g_Q g_R} = \frac{g_Q \circ \tau_P}{g_Q} \frac{g_R \circ \tau_P}{g_R} = e_n(P,Q)e_n(P,R). \end{aligned}$$

Alternating: Bilinearity implies $e_n(P+Q,P+Q) = e_n(P,P)e_n(P,Q)e_n(Q,P)e_n(Q,Q)$, so it suffices to show $e_n(P,P) = 1$ for $P \in E[n]$. Pick $R \in E[n^2]$ with $nR = P$ and let $F_P := \prod_{i=0}^{n-1}(f_P \circ \tau_{iP})$ and $G_P = \prod_{i=0}^{n-1}(g_P \circ \tau_{iR})$ with $f_P \circ \tau_{iP} \circ [n] = g_P \circ \tau_{iR}$ so $F_P = G_P^n$. We then have

$$n \operatorname{div} G_P = \operatorname{div} F_P = \sum_{i=0}^{n-1}(n[P+iP] - n[0+iP]) = n\sum_{i=1}^{n}[iP] - n\sum_{i=0}^{n-1}[iP] = 0,$$

so $G_P \in \bar{k}^\times$ is constant. Therefore $G_P = G_P \circ \tau_R$, and

$$\prod_{i=0}^{n-1} g_P \circ \tau_{iR} = \prod_{i=0}^{n-1} g_P \circ \tau_{(i+1)R}$$

implies $g_P = g_P \circ \tau_{nR} = g_P \circ \tau_P$. Thus $e_n(P,P) = (g_P \circ \tau_P)/g_P = 1$ as desired.

Nondegenerate: If $e_n(P,Q) = 1$ for all $P \in E[n]$ then $g_Q \circ \tau_P = g_Q$ for all $P \in E[n]$. Since n is not divisible by the characteristic, the multiplication-by-n map is separable, and the field extension $\bar{k}(E)/[n]^*(\bar{k}(E))$ is Galois, with the Galois group $E[n]^* = \{\tau_P^* : P \in E[n]\}$. So g_Q lies in the fixed field $\bar{k}(E)^{E[n]^*}$, hence of the form $h \circ [n]$ for some $h \in \bar{k}(E)^\times$. We then have $(h \circ [n])^n = g_Q^n = f_Q \circ [n]$, which implies $f_Q = h^n$ and $n \operatorname{div} h = \operatorname{div} f_Q = n[Q] - n[0]$ implies $\operatorname{div} h = [Q] - [0]$. But then either $h \in \bar{k}^\times$, in which case $\operatorname{div} h = 0$ and $Q = 0$, or $\deg h = \sum_{h(R)=0} v_R(h) = 1$, in which case h defines a map of degree one from E to $\mathbb{P}^1$. The latter is impossible, because $E \not\simeq \mathbb{P}^1$, so if $Q \neq 0$ then $e_n(P,Q) \neq 1$ for some $P \in E[n]$.

Galois equivariance: For any $\sigma \in \operatorname{Gal}(\bar{k}/k)$ and $P, Q \in E[n]$ we have

$$e_n(P^\sigma, Q^\sigma) = \frac{g_{Q^\sigma} \circ \tau_{P^\sigma}}{g_{Q^\sigma}} = \left(\frac{g_Q \circ \tau_P}{g_Q}\right)^\sigma = e_n(P,Q)^\sigma.$$

Compatibility: Let $P \in E[mn]$ and $Q \in E[n] \subseteq E[mn]$. Let us temporarily use the notation $g_{n,Q}, g_{mn,Q}$ and $f_{n,Q}, f_{mn,Q}$ to denote the functions g_Q and f_Q defined above to make the dependence on n explicit. We have

$$\mathrm{div}(g_{n,Q} \circ [m]) = [m]^* \mathrm{div} g_{n,Q} = [m]^*[n]^*([Q] - [0]) = [mn]^*([Q] - [0]) = \mathrm{div} g_{mn,Q}.$$

It follows that $g_{mn,Q} = \lambda(g_{n,Q} \circ [m])$ for some $\lambda \in \bar{k}^\times$, and this implies

$$e_{mn}(P,Q) = \frac{g_{mn,Q} \circ \tau_P}{g_{mn,Q}} = \frac{g_{n,Q} \circ [m] \circ \tau_P}{g_{n,Q} \circ [m]} = \frac{g_{n,Q} \circ \tau_{mP} \circ [m]}{g_{n,Q} \circ [m]} = e_n(mP,Q) \circ [m],$$

and $e_n(mP,Q) \circ [m] = e_n(mP,Q)$, since $e_n(mP,Q)$ is a constant function.

Endomorphisms: Let us fix a basis $E[n] = \langle R, S \rangle$. The action of any $\phi \in \mathrm{End}(E)$ on $E[n]$ is then described by a matrix $\gamma_\phi = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{M}_2(\mathbb{Z}/n\mathbb{Z})$ with $\deg \phi \equiv \det \gamma_\phi = (ad - bc) \bmod n$. For $P = uR + vS$ and $Q = wR + xS$ in $E[n]$, the alternating and bilinearity properties give

$$\begin{aligned}
e_n(\phi(P), \phi(Q)) &= e_n((au+bv)R + (cu+dv)S, (aw+bx)R + (cw+dx)S) \\
&= e_n((au+bv)R, (cw+dx)S) e_n((cu+dv)S, (aw+bx)R) \\
&= e_n(R,S)^{(au+bv)(cw+dx) - (aw+bx)(cu+dv)} \\
&= e_n(R,S)^{(ad-bc)(ux-vw)} \\
&= e_n(uR, xS)^{\det \gamma_\phi} e_n(vS, wR)^{\det \gamma_\phi} \\
&= e_n(uR + vS, wR + xS)^{\deg \phi} \\
&= e_n(P,Q)^{\deg \phi}
\end{aligned}$$

where we have used the fact that e_n has image in μ_n to apply equivalences modulo n.

Isogenies: We first note that it suffices to prove this for isogenies of prime degree, since if $\varphi = \psi \circ \phi$ then

$$e_n(P, \hat{\varphi}(Q)) = e_n(P, \hat{\phi}(\hat{\psi}(Q))) = e_n(\phi(P), \hat{\psi}(Q)) = e_n(\psi(\phi(P)), Q) = e_n(\varphi(P), Q)$$

follows once we know the claim holds for ψ and ϕ, and we can decompose any isogeny into a composition of isogenies of prime degree. So let ϕ be an isogeny of degree ℓ. If ℓ does not divide n then choose r so $\ell r \equiv 1 \bmod n$ and use endomorphism compatibility to obtain

$$e_n(P, \hat{\phi}(Q)) = e_n(P, \hat{\phi}(Q))^{\ell r} = e_n(\phi(P), \ell Q)^r = e_n(\phi(P), Q)^{\ell r} = e_n(\phi(P), Q).$$

If $n = \ell m$ we instead use

$$e_n(P, \hat{\phi}(Q)) = e_m(\ell P, \hat{\phi}(Q)) = e_m(P, \hat{\phi}(Q))^\ell = e_m(\phi(P), \ell Q) = e_n(\phi(P), Q).$$

Surjectivity: Fix any $P \in E[n]$. Bilinearity implies that $\{e_n(P,Q) : Q \in E[n]\}$ is a subgroup μ_m of μ_n. For all $Q \in E[n]$ we have

$$1 = e_n(P,Q)^m = e_n(mP, Q),$$

so by non-degeneracy, $mP = 0$ and m is a multiple of $r = |P|$. On the other hand, if $e_n(P,Q)$ has order m greater than r for any Q, then $e_n(rP,Q) = e_n(P,Q)^r \neq 1$, so $e_n(0,Q) \neq 1$, but $e_n(0,Q) = e_n(0,Q)e_n(Q,Q) = e_n(Q+0,Q) = e_n(Q,Q) = 1$, a contradiction. □

Corollary 23.24. *Let E/k be an elliptic curve and let n be a positive integer prime to the characteristic of k. If $E[n] \subseteq E(k)$ then $\mu_n \subseteq k^\times$. In particular, if $k = \mathbb{Q}$ then $E[n] \subseteq E(k)$ can occur only for $n \leq 2$, and if $k = \mathbb{F}_q$ then $E[n] \subseteq E(k)$ can occur only if $q \equiv 1 \bmod n$.*

Corollary 23.25. *Let E/k be an elliptic curve and let $P \in E(\bar{k})$ be a point of order n prime to the characteristic of k. For every $Q \in E[n]$ the order of $e_n(P,Q)$ in μ_n is the largest $m|n$ for which $E[m] \subseteq \langle P, Q\rangle$, equivalently, the least $m|n$ for which $mQ \in \langle P\rangle$. In particular, $e_n(P,Q) = 1$ if and only if $\langle P, Q\rangle$ is cyclic.*

Proof. Let us first suppose $m = n$, in which case $\langle P, Q\rangle = E[n] \simeq (\mathbb{Z}/n\mathbb{Z})^2$. By the surjectivity of $e_n \colon E[n] \times E[n] \to \mu_n$, we have $e_n(P, aP + bQ) = \zeta_n$ for some $a, b \in \mathbb{Z}$, and

$$\zeta_n = e_n(P, aP + bQ) = e_n(P,P)^a e_n(P,Q)^b = e_n(P,Q)^b$$

(by the bilinear and alternating properties of e_n), so $e_n(P,Q)$ generates $\mu_n = \langle \zeta_n\rangle \simeq \mathbb{Z}/n\mathbb{Z}$ and must have order n.

In the general case we have $mQ = aP$ with $0 \leq a < n$. The order of $aP = mQ$ is at most $r := n/m$, so a is divisible by m, and if we put $c = -a/m$ then $\langle rP, Q + cP\rangle = E[m]$. By the case we have already proved, $e_m(rP, Q + cP)$ has order m, and therefore

$$e_n(P,Q) = e_n(P,Q)e_n(P,P)^c = e_n(P, Q + cP) = e_{mr}(P, Q + cP) = e_m(rP, Q + cP)$$

also has order m (the last equality follows from compatibility). □

The isogeny compatibility of the Weil pairing provides an easy proof of Lemma 6.11.

Corollary 23.26. *Let E_1, E_2 be elliptic curves over a field k. For all $\alpha, \beta \in \operatorname{Hom}(E_1, E_2)$ we have $\widehat{\alpha + \beta} = \hat{\alpha} + \hat{\beta}$.*

Proof. We will show $\phi := \widehat{\alpha + \beta} - \hat{\alpha} - \hat{\beta} = 0$. Suppose not. Choose n coprime to the characteristic of k with $n^2 > \deg \phi$ so $E_2[n] \not\subseteq \ker \phi$. For any $P \in E_1[n]$, $Q \in E_2[n]$ we have

$$\begin{aligned} e_n(P, \widehat{\alpha + \beta}(Q)) &= e_n((\alpha + \beta)(P), Q) \\ &= e_n(\alpha(P) + \beta(P), Q) \\ &= e_n(\alpha(P), Q) e_n(\beta(P), Q) \\ &= e_n(P, \hat{\alpha}(Q)) e_n(P, \hat{\beta}(Q)) \\ &= e_n(P, \hat{\alpha}(Q) + \hat{\beta}(Q)), \end{aligned}$$

where the first and fourth equalities use isogeny compatibility. It follows that for every $P \in E_1[n]$ and $Q \in E_2[n]$ we have

$$e_n(P, \widehat{\alpha + \beta}(Q) - \hat{\alpha}(Q) - \hat{\beta}(Q)) = e_n(P, \phi(Q)) = 1,$$

and the non-degeneracy of e_n implies that this can only hold if $\phi(Q) = 0$ for all $Q \in E_2[n]$, which contradicts $E_2[n] \not\subseteq \ker \phi$. □

23.6 Computing the Weil pairing

For practical applications we want to be able to compute $e_n(P,Q)$ explicitly, in a computationally efficient manner. For this purpose we use the following sequence of functions proposed by Miller [5].

Definition 23.27. Let E/k be an elliptic curve and let $P \in E(k)$. For each integer n we recursively define the function $f_{n,P}$ via

$$f_{0,P} = f_{1,P} := 1, \qquad f_{n+1,P} := f_{n,P}G_{P,nP}, \qquad f_{-n,P} := (f_{n,P}G_{nP,-nP})^{-1},$$

where $G_{P,Q}$ is as in Definition 23.17.

We assume that the line functions $L_{P,Q}$ are all normalized (they will still be defined by an equation for the line $\overline{PQ}$); this implies that the functions $G_{P,Q}$ are also normalized, as are the functions $f_{n,P}$.

Lemma 23.28. *The functions $f_{n,P}$ satisfy the following properties:*

(i) $\operatorname{div} f_{n,P} = n[P] - (n-1)[0] - [nP]$;

(ii) $f_{m+n,P} = f_{m,P}f_{n,P}G_{mP,nP}$;

(iii) $f_{mn,P} = f_{m,P}^n f_{n,mP} = f_{n,P}^m f_{m,nP}$.

Proof. For (i) we proceed by induction on $n \geq 0$. For $n = 0, 1$ we have

$$\operatorname{div} f_{0,P} = 0 = 0[P] - (0-1)[0] - [0P] \quad \text{and} \quad \operatorname{div} f_{1,P} = 0 = 1[P] - (1-1)[0] - [1P],$$

and for $n > 1$ the inductive hypothesis yields

$$\begin{aligned}\operatorname{div} f_{n+1} &= \operatorname{div} f_{n,P} + \operatorname{div} G_{P,nP}\\ &= n[P] - (n-1)[0] - [nP] + [P] + [nP] - [P+nP] - [0]\\ &= (n+1)[P] - (n+1-1)[0] - [(n+1)P].\end{aligned}$$

We then note that

$$\begin{aligned}\operatorname{div} f_{-n,P} &= -\operatorname{div} f_{n,P} - \operatorname{div} G_{nP,-nP}\\ &= -n[P] + (n-1)[0] + [nP] - [nP] - [-nP] + [nP - nP] + [0]\\ &= -n[P] + (n-1)[0] - [-nP] + 2[0]\\ &= -n[P] - (-n-1)[0] - [-nP].\end{aligned}$$

which establishes (i) for all $n \in \mathbb{Z}$.

For (ii) we use (i) to compute

$$\begin{aligned}\operatorname{div} f_{m,P}f_{n,P}G_{mP,nP} &= (m+n)[P] - (m+n-2)[0] - [mP] - [nP]\\ &\qquad + [mP] + [nP] - [mP+nP] - [0]\\ &= (m+n)[P] - (m+n-1)[0] - [(m+n)P]\\ &= \operatorname{div} f_{m+n,P},\end{aligned}$$

and since these are all normalized functions, (ii) follows.

For (iii) we use (i) to compute

$$\begin{aligned}\operatorname{div} f_{m,P}^n f_{n,mP} &= n(m[P] - (m-1)[0] - [mP]) + n[mP] - (n-1)[0] - [mnP] \\ &= nm[P] - (nm-1)[0] - [mnP] \\ &= \operatorname{div} f_{mn,P}.\end{aligned}$$

which establishes the first equality in (iii), since these are normalized functions. The second equality is proved similarly. □

The key part of Lemma 23.28 is (ii), which allows us to efficiently compute $f_{n,P}$ using a double-and-add approach, or any generic exponentiation algorithm, in $O(\log n)$ steps. Lemma 23.28 allows us to reduce the computation of $f_{n,P}(Q)$ to computations of $G_{aP,bP}(Q)$, for various integers a and b. Computing $G_{aP,bP}(Q)$ involves evaluating the line functions $L_{aP,bP}$ and $L_{aP+bP,-(aP+bP)}$ at Q. Assuming we know the coordinates of the points aP and bP (which we will have computed in previous steps of an addition chain), this involves a single application of the group law on E to compute the coordinates of the point $aP + bP$ which we can then negate to compute $-(aP + bP)$ (for curves in short Weierstrass form, this means negating the y-coordinate), followed by $O(1)$ operations in k to evaluate the line functions at Q. Each group operation in $E(k)$ involves just $O(1)$ field operations, and we thus obtain the following corollary,

Corollary 23.29. *Let E/k be an elliptic curve and let n be a positive integer. For any $P, Q \in E(k)$ we can evaluate $f_{n,P}(Q)$ using $O(\log n)$ field operations in k.*

The following result allows us to use the Miller functions to efficiently compute the Weil pairing.

Theorem 23.30. *Let E/k be an elliptic curve with distinct points $P, Q \in E(k)[n]$, where $n \geq 1$ is prime to the characteristic of k. Then*

$$e_n(P,Q) = (-1)^n \frac{f_{n,Q}(P)}{f_{n,P}(Q)}.$$

Proof. This follows from [5, Prop. 8], since, as proved in [8, Prop. 16.1.1], the definition of $e_n(P,Q)$ used in [5] is equivalent to our definition of $e_n(Q,P)$ (the roles of P and Q are swapped in [8, Eq. 16.3] relative to our definition), and $e_n(P,Q) = e_n(Q,P)^{-1}$. □

Warning 23.31. The factor $(-1)^n$ is sometimes incorrectly omitted from this formula in the literature ([4, p. 387] is a notable example).

Note that the definition of $f_{n,P}$ does not require k to be algebraically closed, we just need to work over a field where P is defined, in which case all the points in the support of $\operatorname{div} f_{n,P}$ will be closed points of degree 1 and everything we have done over algebraically closed fields still applies. In particular, if P and Q are k-rational n-torsion points, then $e_n(P,Q)$ will be k-rational.

23.7 Applications of the Weil pairing

There are many applications of the Weil pairing, two of which you will have the opportunity to explore on Problem Set 13. These include an efficient algorithm to compute the structure of the group $E(\mathbb{F}_q)$, which was the original motivation of Miller's work in [5], and a method

for transferring the discrete logarithm problem on an elliptic curve $E/\mathbb{F}_q$ to the multiplicative group of an extension of $\mathbb{F}_q$ containing μ_n, where n is the cardinality of the subgroup of $E(\mathbb{F}_q)$ in which one wishes to compute a discrete logarithm. In most cases the minimal extension of $\mathbb{F}_q$ containing μ_n will be impractically large, but when this is not the case it may be easier to solve the discrete logarithm problem in this extension of $\mathbb{F}_q$ rather than in $E(\mathbb{F}_q)$. The degree of this minimal extension is known as the *embedding degree*, which we discuss in the next section. For cryptographic applications that depend on the difficulty of the discrete logarithm problem, it is important that the embedding degree is not too small. On the other hand, if the embedding degree is not too large, one can then use pairings to efficiently implement cryptographic protocols that would otherwise be impractical.

This brings us to the notion of *pairing-based cryptography*, a topic that we unfortunately do not have time to address in any detail. But we will give one example to demonstrate its utility: a one round tripartite Diffie-Hellman key exchange, due to Joux [4]. For the sake of presentation we will describe it in terms of the Weil pairing, but in practice one uses the more efficient Tate pairing defined in §23.9 below.

We assume that Alice, Bob, and Carol all know an elliptic curve $E/\mathbb{F}_q$ and two independent n-torsion points P and Q in $E[n]$. They want to agree on a random secret, and they would like to do this with a single round of messaging that does not require any back-and-forth communication.

To begin the protocol, Alice, Bob, and Carol individually generate random integers a, b, and c, respectively. Alice then sends $P_A := aP$ and $Q_A := aQ$ to Bob and Carol, Bob sends $P_B := bP$ and $Q_B := bQ$ to Alice and Carol, and Carol sends $P_C := cP$ and $Q_C := cQ$ to Alice and Bob.

Alice then computes

$$e_n(P_B, Q_C)^a = e_n(bP, cQ)^a = e_n(P, Q)^{bca},$$

Bob computes

$$e_n(P_A, Q_C)^b = e_n(aP, cQ)^b = e_n(P, Q)^{acb},$$

and Carol computes

$$e_n(P_A, Q_B)^c = e_n(aP, bQ)^c = e_n(P, Q)^{abc}.$$

The common value $e_n(P, Q)^{abc} \in \mu_n$ is now known to Alice, Bob, and Carol. If one assumes that the discrete logarithm problem is hard, an eavesdropper cannot readily determine the values of a, b, or c, and if one further assumes that the computational Diffie-Hellman problem is hard, an eavesdropper cannot readily determine the shared secret $e_n(P, Q)^{abc}$. The *computational Diffie-Hellman problem* is to compute abP, given P, aP, and bP; this can clearly be solved efficiently if one can compute discrete logarithms efficiently, but the converse is not known.

23.8 Embedding degree

For practical applications one typically applies Miller's algorithm to n-torsion points of an elliptic curve $E/\mathbb{F}_q$, where $\mathbb{F}_q$ is a finite field and n is a prime dividing $\#E(\mathbb{F}_q)$. While we typically will not have $E[n] \subseteq E(\mathbb{F}_q)$ (indeed, $E(\mathbb{F}_q)$ will often be cyclic), we can always choose an n that divides $\#E(\mathbb{F}_q)$, in which case we at least have a cyclic subgroup of $E[n]$ of order n that lies in $E(\mathbb{F}_q)$ (assuming n is prime). The remaining points in $E[n]$ will then lie in a finite extension of $\mathbb{F}_q$; as indicated in the previous section, the degree of this extension is a key parameter.

Definition 23.32. Let E/K be an elliptic curve over a field K and let n be a positive integer. The *embedding degree* of E with respect to n is the degree of the minimal extension L/K for which $E[n] \subseteq E(L)$.

An easy lower bound on the embedding degree k arises from the fact that the Weil pairing $E[n] \times E[n] \to \mu_n$ is surjective. If $E[n] \subseteq E(\mathbb{F}_{q^k})$ then we must have $\mu_n \subseteq \mathbb{F}_{q^k}^\times$. The group $\mathbb{F}_{q^k}^\times$ is cyclic, so this is the same as requiring n to divide $q^k - 1$, equivalently, $q^k \equiv 1 \bmod n$. When $E(\mathbb{F}_q)$ contains a cyclic group of order n, this necessary condition is also sufficient.

Lemma 23.33. *Let $E/\mathbb{F}_q$ be an elliptic curve, let $n \perp q$ be a prime divisor of $\#E(\mathbb{F}_q)$, and let $\pi_n \in \mathrm{End}(E[n]) \simeq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$ denote the restriction of the Frobenius endomorphism π_E to $E[n]$. Then either $E[n] \subseteq E(\mathbb{F}_q)$ or $E[n] \simeq \ker(\pi_n - 1) \oplus \ker(\pi_n - q)$, and the embedding degree of E with respect to n is the least integer $k > 0$ such that $q^k \equiv 1 \bmod n$.*

Proof. Let $t = \operatorname{tr} \pi_E$, so that $\#E(\mathbb{F}_q) = q+1-t$. Then $t \equiv q+1 \bmod n$ and the characteristic polynomial of π_E satisfies $x^2 - tx + q \equiv x^2 - (q+1)x + q \equiv (x-1)(x-q) \bmod n$. It follows that $(\pi_n - 1)(\pi_n - q) = 0$ in $\mathrm{End}(E[n])$. If $q \equiv 1 \bmod n$ then π_E acts trivially on $E[n]$ and $E[n] \subseteq E(\mathbb{F}_q)$; otherwise $\pi_n \in \mathrm{End}(E[n]) \simeq \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$ can be diagonalized and $E[n]$ can be decomposed as the sum of the distinct eigenspaces $\ker(\pi_n - 1)$ and $\ker(\pi_n - q)$ of π_n.

As observed above, the embedding degree e necessarily satisfies $q^e \equiv 1 \bmod n$, since $\mu_n \subseteq \mathbb{F}_{q^e}^\times$, so $e \geq k$. On the other hand, for $P \in \ker(\pi_n - 1)$ we have $P \in E(\mathbb{F}_q) \subseteq E(\mathbb{F}_{q^k})$, and for $P \in \ker(\pi_n - q)$ we have $\pi_n^k(P) = q^k P = P$ (since $q^k \equiv 1 \bmod n$), in which case P is fixed by π_E^k and lies in $E(\mathbb{F}_{q^k})$. It follows that $E[n] \subseteq E(\mathbb{F}_{q^k})$ and therefore $e \leq k$, so $e = k$ as claimed. □

Lemma 23.33 gives us an easy way to compute the embedding degree k when $n | \#E(\mathbb{F}_q)$. If we suppose E is chosen arbitrarily, we should expect q to be roughly equidistributed modulo n, and for most values of n this means it is likely that q is a primitive root modulo n, in which case we must have $k = n - 1$ (assuming n is prime). This is bad news for practical applications: if $k = n - 1$ it will take $\log_2(\#\mathbb{F}_{q^k}) = (n-1)\log_2 q \approx n \log n$ bits just to write down a typical n-torsion point, which is hopeless if n is of cryptographic size (say $n \approx 2^{256}$), since this will be more bits than there are atoms in the universe.

Practical applications of the Weil pairing are feasible only when k is small. It is possible to have k as small as 1 or 2 when E is supersingular (see Problem Set 12), but this is too small for cryptographic applications, as you will demonstrate on Problem Set 12, since one can transfer the discrete logarithm problem in $E(\mathbb{F}_q)$ to the discrete logarithm problem in $\mathbb{F}_{q^k}^\times$. Ideally one wants k to be around 10 or 20 to balance the difficulty of the discrete logarithm problems in $E(\mathbb{F}_q)$ and $\mathbb{F}_{q^k}^\times$; for $q \approx 2^{256}$ using $k = 12$ yields $\#\mathbb{F}_{q^k}^\times \approx 2^{3072}$, in which case the discrete logarithm problems have similar difficulty.

Elliptic curves with embedding degrees in this range are known as *pairing-friendly* curves. They are quite rare, far too rare to find by brute force search, but they can be constructed using the CM method. See [3] for an extensive survey of methods to compute suitable parameters q, n, k, D, where q and n are cryptographic size primes, k is small, $q^k \equiv 1 \bmod n$, and D is an imaginary quadratic discriminant with $|D|$ small enough so that the CM method can be used to construct an elliptic curve $E/\mathbb{F}_q$ so that n divides $\#E(\mathbb{F}_q)$.

23.9 Tate pairing

In most practical applications of pairings, rather than using the Weil pairing one instead uses the Tate pairing, or variations thereof, which can be computed much more efficiently.

Definition 23.34. Let $n > 2$ be an integer and let $E/\mathbb{F}_q$ be an elliptic curve over a finite field with embedding degree k with respect to n. The (modified) *Tate pairing* is the map $t_n\colon E[n] \times E[n] \to \mu_n$ defined by

$$t_n(P,Q) := \left(\frac{f_{n,P}(Q+T)}{f_{n,P}(T)}\right)^{(q^k-1)/n}$$

where $T \in E[n] - \{0, P, -Q, P-Q\}$.

The exponentiation by $(q^k-1)/n$ included in our definition of the Tate pairing means that if $P \in E[n]$ we can actually compute $t_n(P,Q)$ using any $Q \in E(\mathbb{F}_{q^k})$; the value of $t_n(P,Q)$ depends only on the image of $Q \in E(\mathbb{F}_{q^k})$ under the quotient map

$$E(\mathbb{F}_{q^k}) \to E(\mathbb{F}_{q^k})/nE(\mathbb{F}_{q^k}) \simeq E[n],$$

and we can view $Q \in E(\mathbb{F}_{q^k})$ as representing a coset of $nE(\mathbb{F}_{q^k})$ corresponding to an element of $E[n]$ (the Tate pairing is sometimes defined with this interpretation in mind).

Like the Weil pairing, the Tate pairing is a non-degenerate bilinear pairing that is surjective and Galois-equivariant. Unlike the Weil pairing, the Tate pairing is not alternating, and may have $t_n(P,P) \neq 1$; this is an advantage in many practical applications, because it means that the pairing may be non-trivial even when we restrict to points in a cyclic subgroup of $E[n]$, which is never true of the Weil pairing. Another advantage is that we only need to compute one Miller function $f_{n,P}$, rather than the two Miller functions $f_{n,P}$ and $f_{n,Q}$ required by the Weil pairing, and in the typical case where n is a prime dividing $\#E(\mathbb{F}_q)$, we can choose $P \in E(\mathbb{F}_q)$ to be rational, which greatly accelerates this computation.

In the practically interesting scenario where $n \perp q$ is a prime dividing $\#E(\mathbb{F}_q)$ and $k > 1$, Lemma 23.33 gives us a natural decomposition of $E[n] \simeq \ker(\pi_n - 1) \oplus \ker(\pi_n - q)$ into two cyclic subgroups of order n, the first of which is just $E(\mathbb{F}_q)[n]$. In many applications (and in many descriptions of the Tate pairing in the literature), one restricts the inputs of the Tate pairing to $P \in \ker(\pi_n - 1) = E(\mathbb{F}_q)[n]$ and $Q \in \ker(\pi_n - q) \subseteq E(\mathbb{F}_{q^k})$.

References

[1] Dan Boneh and Matthew Franklin, *Identity-based encryption from the Weil pairing*, SIAM J. Comput. **32** (2003), 586–615.

[2] Andreas Enge, *Elliptic curves and their applications to cryptography: An introduction*, Springer, 1999.

[3] David Freeman, Michael Scott, and Edlyn J. Teske, *A taxonomy of pairing-friendly elliptic curves*, J. Cryptology **23** (2010), 224-280.

[4] Antoine Joux, *A one round protocol for tripartite Diffie-Hellman*, Algorithmic Number Theory 4th International Symposium (ANTS IV), LNCS **1838** (2000), 385–394.

[5] Victor S. Miller, *The Weil pairing and its efficient calculation*, J. Cryptology **17** (2004), 235–261.

[6] Adi Shamir, *Identity based cryptosystems and signature schemes*, Advances in Cryptology – Proceedings of CRYPTO '84, LNCS **196** (1985), 47–53.

[7] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer, 2009.

[8] Katherine E. Stange, *Elliptic nets and elliptic curves*, PhD Thesis, Brown University, 2008.

[9] Lawrence C. Washington, *Elliptic curves: Number theory and cryptography*, second edition, Chapman and Hall/CRC, 2008.

24 Modular forms and L-functions

As we will prove in the next lecture, Fermat's Last Theorem is a corollary of the following theorem for elliptic curves over $\mathbb{Q}$ [19, 20].

Theorem 24.1 (Taylor-Wiles)**.** *Every semistable elliptic curve $E/\mathbb{Q}$ is modular.*

In fact, as a result of subsequent work [4], we now have the following stronger result.

Theorem 24.2 (Breuil-Conrad-Diamond-Taylor)**.** *Every elliptic curve $E/\mathbb{Q}$ is modular.*

In this lecture we will explain what it means for an elliptic curve over $\mathbb{Q}$ to be modular (we will also define the term semistable).

This requires us to delve briefly into the theory of modular forms. Our goal in doing so is simply to understand the definitions and the terminology; we will omit all but the most straightforward proofs.

24.1 Modular forms

Definition 24.3. A holomorphic function $f\colon \mathcal{H} \to \mathbb{C}$ is a *weak modular form* of *weight* k for a congruence subgroup Γ if

$$f(\gamma\tau) = (c\tau + d)^k f(\tau)$$

for all $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$.

Example 24.4. The j-function $j(\tau)$ is a weak modular form of weight 0 for $\mathrm{SL}_2(\mathbb{Z})$, and $j(N\tau)$ is a weak modular form of weight 0 for $\Gamma_0(N)$. For an example of a weak modular form of positive weight, recall the Eisenstein series

$$G_k(\tau) := G_k([1,\tau]) := \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}} \frac{1}{(m+n\tau)^k},$$

which, for $k \geq 3$, is a weak modular form of weight k for $\mathrm{SL}_2(\mathbb{Z})$. To see this, recall that $\mathrm{SL}_2(\mathbb{Z})$ is generated by the matrices $S = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, and note that

$$G_k(S\tau) = G_k(-1/\tau) = \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}} \frac{1}{(m-\frac{n}{\tau})^k} = \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}} \frac{\tau^k}{(m\tau-n)^k} = \tau^k G_k(\tau),$$

$$G_k(T\tau) = G_k(\tau+1) = G_k(\tau) = 1^k G_k(\tau).$$

If Γ contains $-I$, then any weak modular form f for Γ must satisfy $f(\tau) = (-1)^k f(\tau)$, since $-I$ acts trivially and $c\tau+d = -1$; this implies that when $-I \in \Gamma$ the only weak modular form of odd weight is the zero function. We are specifically interested in the congruence subgroup $\Gamma_0(N)$, which contains $-I$, so we will restrict our attention to modular forms of even weight, but we should note that for other congruence subgroups such as $\Gamma_1(N)$ that do not contain $-I$ (for $N > 2$) there are interesting modular forms of odd weight.

As we saw with modular functions (see Lecture 19), if Γ is a congruence subgroup of level N, meaning that it contains $\Gamma(N)$, then Γ contains the matrix $T^N = \left(\begin{smallmatrix} 1 & N \\ 0 & 1 \end{smallmatrix}\right)$, and every

weak modular form $f(\tau)$ for Γ must satisfy $f(\tau+N)=f(\tau)$ for $\tau\in\mathcal{H}$, since for T^N we have $(c\tau+d)^k=1^k=1$. It follows that $f(\tau)$ has a q-expansion of the form

$$f(\tau)=f^*(q^{1/N})=\sum_{n=-\infty}^{\infty}a_nq^{n/N}\qquad(q:=e^{2\pi i\tau}).$$

We say that f is *holomorphic at* ∞ if f^* is holomorphic at 0, equivalently, $a_n=0$ for $n<0$. We say that f is *holomorphic at the cusps* if $f(\gamma\tau)$ is holomorphic at ∞ for all $\gamma\in\mathrm{SL}_2(\mathbb{Z})$. As with modular functions, we only need to check this condition at a (finite) set of cusp representatives for Γ (if f is holomorphic at a particular cusp in $\mathbb{P}^1(\mathbb{Q})$ then it is necessarily holomorphic at every Γ-equivalent cusp). We should note that a weak modular form of positive weight is not Γ-invariant, so even when it is holomorphic on a cusp orbit, it may take on different values at cusps in the same orbit (but if it vanishes at a particular cusp then it vanishes at every Γ-equivalent cusp; this is relevant to the *cusp forms* defined below).

Definition 24.5. A *modular form* f is a weak modular form that is holomorphic at the cusps. Equivalently, f is a weak modular form that extends to a holomorphic function on the extended upper half plane $\mathcal{H}^*=\mathcal{H}\cup\mathbb{P}^1(\mathbb{Q})$.

If Γ is a congruence subgroup that contains the matrix $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ then every modular form for Γ has a q-series expansion at ∞ (or any cusp) of the form

$$f(\tau)=f^*(q)=\sum_{n\ge0}a_nq^n$$

that contains only integer powers of q, regardless of the level N. This includes the congruence subgroups $\Gamma_0(N)$ and $\Gamma_1(N)$ of interest to us. The coefficients a_n in the q-series for f are also referred to as the *Fourier coefficients* of f.

The only modular forms of weight 0 are constant functions. This is the main motivation for introducing the notion of weight, it allows us to generalize modular functions in an interesting way, by strengthening their analytic properties (holomorphic on $\mathcal{H}^*$, not just meromorphic) at the expense of weakening their congruence properties (modular forms of positive weight are not Γ-invariant due to the factor $(c\tau+d)^k$).

The j-function is not a modular form, since it has a pole at ∞, but the Eisenstein functions $G_k(\tau)$ are nonzero modular forms of weight k for $\mathrm{SL}_2(\mathbb{Z})$ for all even $k\ge4$. For $\Gamma=\mathrm{SL}_2(\mathbb{Z})$ there is only one cusp to check and it suffices to note that

$$\lim_{\mathrm{im}\,\tau\to\infty}G_k(\tau)=\lim_{\mathrm{im}(\tau)\to\infty}\sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}}\frac{1}{(m+n\tau)^k}=2\sum_{n=1}^{\infty}\frac{1}{n^k}=2\zeta(k)<\infty,$$

(recall that the series converges absolutely, which justifies rearranging its terms).

Definition 24.6. A modular form is a *cusp form* if it vanishes at all the cusps. Equivalently, its q-expansion at every cusp has constant coefficient $a_0=0$.

Example 24.7. For even $k\ge4$ the Eisenstein series $G_k(\tau)$ is not a cusp form, but the discriminant function

$$\Delta(\tau)=g_2(\tau)^3-27g_3(\tau)^2,$$

with $g_2(\tau)=60G_4(\tau)$ and $g_3(\tau)=140G_6(\tau)$, is a cusp form of weight 12 for $\mathrm{SL}_2(\mathbb{Z})$; to see that it vanishes at ∞, note that $j(\tau)=g_2(\tau)^3/\Delta(\tau)$ has a pole at ∞ and $g_2(\tau)$ does not, so $\Delta(\tau)$ must vanish (see the proof of Theorem 15.11).

The set of modular forms of weight k for Γ is closed under addition and multiplication by constants $\lambda \in \mathbb{C}$ and thus forms a $\mathbb{C}$-vector space $M_k(\Gamma)$ that contains the cusp forms $S_k(\Gamma)$ as a subspace. We also note that if $f_1 \in M_{k_1}(\Gamma)$ and $f_2 \in M_{k_2}(\Gamma)$ then $f_1 f_2 \in M_{k_1+k_2}(\Gamma)$, but we will not use this fact.

Remarkably, the dimensions of the vector spaces $M_k(\Gamma)$ and $S_k(\Gamma)$ are finite, and can be explicitly computed in terms of invariants of the corresponding modular curve $X(\Gamma) = \mathcal{H}^*/\Gamma$.

As in Problem Set 10, let $\nu_2(\Gamma)$ count the number of Γ-inequivalent $\mathrm{SL}_2(\mathbb{Z})$-translates of i fixed by some $\gamma \in \Gamma$ other than $\pm I$ (elliptic points of period 2), and similarly define $\nu_3(\Gamma)$ in terms of $\rho = e^{2\pi i/3}$ (elliptic points of period 3). Let ν_∞ denote the number of cusp orbits, and let $g(\Gamma)$ be the genus of $X(\Gamma)$.

Theorem 24.8. *Let Γ be a congruence subgroup. For $k = 0$ we have $\dim M_k(\Gamma) = 1$ and $\dim S_k(\Gamma) = 0$. For any even integer $k > 0$ we have*

$$\dim M_k(\Gamma) = (k-1)(g(\Gamma)-1) + \left\lfloor \frac{k}{4} \right\rfloor \nu_2 + \left\lfloor \frac{k}{3} \right\rfloor \nu_3 + \frac{k}{2}\nu_\infty,$$

and if $k > 2$ we also have

$$\dim S_k(\Gamma) = (k-1)(g(\Gamma)-1) + \left\lfloor \frac{k}{4} \right\rfloor \nu_2 + \left\lfloor \frac{k}{3} \right\rfloor \nu_3 + \left(\frac{k}{2} - 1\right)\nu_\infty.$$

For $k = 2$ we have $\dim S_k(\Gamma) = g(\Gamma)$.

Proof. See [6, Thm. 3.5.1] □

We are specifically interested in the vector space $S_2(\Gamma_0(N))$ of dimension $g(\Gamma_0(N))$.

Remark 24.9. Those who know a bit of algebraic geometry may suspect that there is a relationship between the space of cusp forms $S_2(\Gamma_0(N))$ and the space of regular differentials for the modular curve $X_0(N)$, since their dimensions coincide; this is indeed the case.

24.2 Hecke operators

In order to understand the relationship between modular forms and elliptic curves we want to construct a canonical basis for $S_2(\Gamma_0(N))$. To help with this, we now introduce the *Hecke operators* T_n on $M_k(\Gamma_0(N))$; these are linear operators that fix the subspace $S_k(\Gamma_0(N))$.[1]

In order to motivate the definition of the Hecke operators on modular forms, we first define them in terms of lattices, following the presentation in [13, VII.5.1]. As in previous lectures, a lattice (in $\mathbb{C}$) is an additive subgroup of $\mathbb{C}$ that is a free $\mathbb{Z}$-module of rank 2 containing an $\mathbb{R}$-basis for $\mathbb{C}$.

For each positive integer n, the Hecke operator T_n sends each lattice $L = [\omega_1, \omega_2]$ to the formal sum of its index-n sublattices.

$$T_n L := \sum_{[L:L']=n} L'. \tag{1}$$

Here we are working in the free abelian group $\operatorname{Div}\mathcal{L}$ generated by the set $\mathcal{L}$ of all lattices; we extend T_n linearly to an endomorphism of $\operatorname{Div}\mathcal{L}$ (this means $T_n \sum L := \sum T_n L$). Another family of endomorphisms of $\operatorname{Div}\mathcal{L}$ are the homothety operators R_λ defined by

$$R_\lambda L := \lambda L, \tag{2}$$

[1] One can define Hecke operators more generally on $M_k(\Gamma_1(N))$, which contains $M_k(\Gamma_0(N))$, but the definition is more involved and not needed here.

for any $\lambda \in \mathbb{C}^\times$. This setup might seem overly abstract, but it allows one to easily prove some essential properties of the Hecke operators that are applicable in many settings. When defined in this generality the Hecke operators are also sometimes called *correspondences*.

Remark 24.10. Recall that if $E/\mathbb{C}$ is the elliptic curve isomorphic to the torus $\mathbb{C}/L$, the index-n sublattices of L correspond to n-isogenous elliptic curves. The fact that the Hecke operators average over sublattices is related to the fact that the relationship between modular forms and elliptic curves occurs at the level of isogeny classes.

Theorem 24.11. *The operators T_n and R_λ satisfy the following:*

(i) $T_n R_\lambda = R_\lambda T_n$ *and* $R_\lambda R_\mu = R_{\lambda\mu}$.

(ii) $T_{mn} = T_m T_n$ *for all* $m \perp n$.

(iii) $T_{p^{r+1}} = T_{p^r} T_p - p T_{p^{r-1}} R_p$ *for all primes p and integers $r \geq 1$.*

Proof. (i) is clear, as is (ii) if we note that for $m \perp n$ there is a bijection between index-mn sublattices L'' of L and pairs (L', L'') with $[L : L'] = n$ and $[L' : L''] = m$. For (iii), the first term on the RHS counts pairs (L', L'') with $[L : L'] = p$ and $[L' : L''] = p^r$, and the second term corrects for over counting; see [13, Prop. VII.10] for details. □

Corollary 24.12. *The subring of* $\mathrm{End}(\mathrm{Div}\,\mathcal{L})$ *generated by* $\{R_p, T_p : p \text{ prime}\}$ *is commutative and contains all the Hecke operators T_n.*

Proof. By recursively applying (iii) we can reduce any T_{p^r} to a polynomial in T_p and R_p, and any two such polynomials commute (since T_p and R_p commute, by (i)). Moreover, (i) and (ii) imply that for distinct primes p and q, polynomials in T_p, R_p commute with polynomials in T_q, R_q. Using (ii) and (iii) we can reduce any T_n to a product of polynomials in T_{p_i}, R_{p_i} for distinct primes p_i and the corollary follows. □

Any function $F\colon \mathcal{L} \to \mathbb{C}$ extends linearly to a function $F\colon \mathrm{Div}\,\mathcal{L} \to \mathbb{C}$ to which we may apply any operator $T \in \mathrm{End}(\mathrm{Div}\,\mathcal{L})$, yielding a new function $TF\colon \mathrm{Div}\,\mathcal{L} \to \mathbb{C}$ defined by $TF\colon D \mapsto F(T(D))$; restricting TF to $\mathcal{L} \subseteq \mathrm{Div}\,\mathcal{L}$ then gives a function $TF\colon \mathcal{L} \to \mathbb{C}$ that we regard as the transform of our original function F by T. This allows us to apply the Hecke operators T_n and homothety operators R_λ to any function that maps lattices to complex numbers. We will work this out explicitly for the Hecke operators acting on modular forms for $\mathrm{SL}_2(\mathbb{Z})$ in the next section.

24.3 Hecke operators for modular forms of level one

We now define the action of the Hecke operators T_n on $M_k(\mathrm{SL}_2(\mathbb{Z})) = M_k(\Gamma_0(1))$. The case $M_k(\Gamma_0(N))$ is analogous, but the details are more involved, so let us assume $N = 1$ for the sake of presentation and address $N > 1$ in remarks.

Let $f\colon \mathcal{H} \to \mathbb{C}$ be a modular form of weight k. We can view $f(\tau)$ as a function on lattices $[1, \tau]$, which we extend to arbitrary lattices $L = [\omega_1, \omega_2]$ by defining

$$f([\omega_1, \omega_2]) := f(\omega_1[1, \omega_2/\omega_1]) := \omega_1^{-k} f([1, \omega_2/\omega_1]),$$

we assume ω_1 and ω_2 are ordered so that ω_2/ω_1 is in the upper half plane. Conversely, any function $F\colon \mathcal{L} \to \mathbb{C}$ on lattices induces a function $\tau \mapsto F([1, \tau])$ on the upper half plane. Viewing our modular form f as a function $\mathcal{L} \to \mathbb{C}$, we can transform this function by any

$T \in \mathrm{End}(\mathrm{Div}\,\mathcal{L})$ as described above, thereby obtaining a new function $\mathcal{L} \to \mathbb{C}$ that induces a function $Tf\colon \mathcal{H} \to \mathbb{C}$ on the upper half plane. In general the function Tf need not be a modular form, but for $f \in M_k(\Gamma_0(1))$ it is (we will verify this in the cases of interest to us).

Motivated by the discussion above, for $f \in M_k(\Gamma_0(1))$ we define

$$R_\lambda f(\tau) := f(\lambda[1,\tau]) = \lambda^{-k} f(\tau),$$

which clearly lies in $M_k(\Gamma_0(1))$, and if f is a cusp form then so is $R_\lambda f$.

We define $T_n f$ similarly, but introduce a scaling factor of n^{k-1} that simplifies the formulas that follow. An easy generalization of Lemma 20.2 shows that for each integer $n \geq 1$, the index n sublattices of $[1,\tau]$ are given by

$$\Big\{[d, a\tau + b] : ad = n,\ 0 \leq b < d\Big\};$$

see [13, Lem. VII.5.2], for example. If we rescale by d^{-1} to put them in the form $[1,\omega]$, we have $\omega = (a\tau + b)/d$. For $f \in M_k(\Gamma_0(1))$ we thus define $T_n f$ as

$$T_n f(\tau) := n^{k-1} \sum_{[[1,\tau]:L]=n} f(L) = n^{k-1} \sum_{ad=n,\, 0\leq b<d} d^{-k} f\left(\frac{a\tau+b}{d}\right),$$

which is also clearly an element of $M_k(\Gamma_0(1))$, and if f is a cusp form, so is $T_n f$. It is clear from the definition that T_n acts linearly, so it is a linear operator on the vector spaces $M_k(\Gamma_0(1))$ and $S_k(\Gamma_0(1))$. Theorem 24.11 then yields the following corollary.

Corollary 24.13. *The Hecke operators T_n for $M_k(\Gamma_0(1))$ satisfy $T_{mn} = T_m T_n$ for $m \perp n$ and $T_{p^{r+1}} = T_{p^r} T_p - p^{k-1} T_{p^{r-1}}$ for p prime.*

Proof. The first equality is clear; the second term on the RHS of the second equality arises from the fact that $pT_{p^{r-1}} R_p f = p^{k-1} T_{p^{r-1}} f$. □

The corollary implies that we may restrict our attention to the Hecke operators T_p for p prime. Let us compute the q-series expansion of $T_p f$, where $f(\tau) = \sum_{n=1}^\infty a_n q^n$ is a cusp form of weight k for $\Gamma_0(1)$. We have

$$\begin{aligned}
T_p f(\tau) &= p^{k-1} \sum_{\substack{ad=p\\0\leq b<d}} d^{-k} f\left(\frac{a\tau+b}{d}\right)\\
&= p^{k-1} f(p\tau) + p^{-1} \sum_{b=0}^{p-1} f\left(\frac{\tau+b}{p}\right)\\
&= p^{k-1} \sum_{n=1}^\infty a_n e^{2\pi i n p\tau} + p^{-1} \sum_{b=0}^{p-1} \sum_{n=1}^\infty a_n e^{2\pi i n(\tau+b)/p}\\
&= p^{k-1} \sum_{n=1}^\infty a_n q^{np} + p^{-1} \sum_{b=0}^{p-1} \sum_{n=1}^\infty a_n \zeta_p^{bn} q^{n/p}\\
&= p^{k-1} \sum_{n=1}^\infty a_{n/p} q^n + p^{-1} \sum_{n=1}^\infty a_n \left(\sum_{b=0}^{p-1} \zeta_p^{bn}\right) q^{n/p}\\
&= \sum_{n=1}^\infty \left(a_{np} + p^{k-1} a_{n/p}\right) q^n,
\end{aligned}$$

where $\zeta_p = e^{2\pi i/p}$ and $a_{n/p}$ is defined to be 0 when $p \nmid n$. This calculation yields the following theorem and corollary, in which we use $a_n(f)$ to denote the coefficient of q^n in the q-expansion of f.

Theorem 24.14. *For any $f \in S_k(\Gamma_0(1))$ and prime p we have*

$$a_n(T_p f) = \begin{cases} a_{np}(f) & \text{if } p \nmid n, \\ a_{np}(f) + p^{k-1} a_{n/p}(f) & \text{if } p \mid n. \end{cases}$$

Corollary 24.15. *For any modular form $f \in S_k(\Gamma_0(1))$ and integers $m \perp n$ we have $a_m(T_n f) = a_{mn}(f)$; in particular, $a_1(T_n f) = a_n(f)$.*

Proof. The corollary follows immediately from Theorem 24.14 for n prime. For composite n (and any $m \perp n$), we proceed by induction on n. If $n = cd$ with $c \perp d$ both greater than 1, then by Theorem 24.14 and the inductive hypothesis we have

$$a_m(T_n f) = a_m(T_c T_d f) = a_{mc}(T_d f) = a_{mcd} = a_{mn}.$$

For $n = p^{r+1}$, applying Theorem 24.14, Corollary 24.13, and the inductive hypothesis yields

$$\begin{aligned} a_m(T_{p^{r+1}} f) &= a_m(T_{p^r} T_p f) - p^{k-1} a_m(T_{p^{r-1}} f) \\ &= a_{mp^r}(T_p f) - p^{k-1} a_{mp^{r-1}}(f) \\ &= a_{mp^{r+1}}(f) + p^{k-1} a_{mp^{r-1}}(f) - p^{k-1} a_{mp^{r-1}}(f) \\ &= a_{mn}(f), \end{aligned}$$

as desired. □

Remark 24.16. All the results in this section hold for $f \in S_k(\Gamma_0(N))$ if we restrict to Hecke operators T_n with $n \perp N$, which is all that we require, and the key result $a_1(T_n f) = a_n(f)$ holds in general. For $p|N$ the definition of T_p (and T_n for $p|n$) needs to change and the formulas in Corollary 24.13 and Theorem 24.14 must be modified. The definition of the Hecke operators is more complicated (in particular, it depends on the level N), but some of the formulas are actually simpler (for example, for $p|N$ we have $T_{p^r} = T_p^r$).

24.4 Eigenforms for the Hecke operators

The Hecke operators T_n defined in the previous section form an infinite family of linear operators on the vector space $S_k(\Gamma_0(1))$. We are interested in the elements $f \in S_k(\Gamma_0(1))$ that are simultaneous eigenvectors for all the Hecke operators; this means that for every $n \geq 1$ we have $T_n f = \lambda_n f$ for some eigenvalue $\lambda_n \in \mathbb{C}$ of T_n. When such an f also satisfies $a_1(f) = 1$, we call it a (normalized) *eigenform*. It is not immediately obvious that such f exist, but we will prove that they do, and that they provide a canonical basis for $S_k(\Gamma_0(1))$.

Given an eigenform f, we can read off the corresponding Hecke eigenvalues λ_n from its q-expansion $f = \sum a_n q^n$: if $T_n f = \lambda_n f$ then we must have

$$\lambda_n = \lambda_n a_1 = a_1(T_n f) = a_n(f) = a_n,$$

by Corollary 24.15. Corollary 24.13 implies that the a_n then satisfy

$$\begin{aligned} a_{mn} &= a_m a_n && (m \perp n), \\ a_{p^r} &= a_p a_{p^{r-1}} - p^{k-1} a_{p^{r-2}} && (p \text{ prime}). \end{aligned} \tag{3}$$

In particular, the coefficients a_n are completely determined by the values a_p at primes p.

Remark 24.17. For $k = 2$ the recurrence for a_{p^r} should look familiar: it is the same recurrence satisfied by the Frobenius traces $a_{p^r} := p^r + 1 - \#E(\mathbb{F}_{p^r})$ of an elliptic curve $E/\mathbb{F}_p$, as shown in Problem Set 7.

Our goal in this section is to construct a basis of eigenforms for $S_k(\Gamma_0(1))$, and prove that it is unique. In order to do so, we need to introduce the *Petersson inner product*, which defines a Hermitian form on the $\mathbb{C}$-vector spaces $S_k(\Gamma)$ (for any congruence subgroup Γ). Recall that for $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$, we have $\operatorname{im}\gamma\tau = \operatorname{im}\tau/|c\tau+d|^2$, thus for any $f, g \in S_k(\Gamma)$ we have

$$f(\gamma\tau)\overline{g(\gamma\tau)}(\operatorname{im}\gamma\tau)^k = (c\tau+d)^k f(\tau)(c\bar{\tau}+d)^k g(\tau)\left(\frac{\operatorname{im}\tau}{|c\tau+d|^2}\right)^k = f(\tau)\overline{g(\tau)}(\operatorname{im}\tau)^k.$$

The function $f(\tau)\overline{g(\tau)}(\operatorname{im}\tau)^k$ is thus Γ-invariant. If we parameterize the upper half-plane $\mathcal{H}$ with real parameters $x = \operatorname{re}\tau$ and $y = \operatorname{im}\tau$, so $\tau = x + iy$, it is straight-forward to check that the measure

$$\mu(U) = \iint_U \frac{dxdy}{y^2}$$

is $\mathrm{SL}_2(\mathbb{Z})$-invariant (hence Γ-invariant), that is, $\mu(\gamma U) = \mu(U)$ for all measurable sets $U \subseteq \mathcal{H}$. This motivates the following definition.

Definition 24.18. The *Petersson inner product* on $S_k(\Gamma)$ is defined by

$$\langle f, g\rangle = \int_{\mathcal{F}} f(\tau)\overline{g(\tau)}y^{k-2}dxdy, \tag{4}$$

where the integral ranges over points $\tau = x + yi$ in a fundamental region $\mathcal{F} \subseteq \mathcal{H}$ for Γ. It is easy to check that $\langle f, g\rangle$ is a positive definite Hermitian form: it is sesquilinear in f and g, it satisfies $\langle f, g\rangle = \overline{\langle g, f\rangle}$, and $\langle f, f\rangle \geq 0$ with equality only when $f = 0$. It thus defines an inner product on the $\mathbb{C}$-vector space $S_k(\Gamma)$.

One can show that the Hecke operators for $S_k(\Gamma_0(1))$ are self-adjoint with respect to the Petersson inner product, that is, they satisfy $\langle f, T_n g\rangle = \langle T_n f, g\rangle$. The T_n are thus Hermitian (normal) operators, and we know from Corollary 24.13 that they all commute with each other. This makes it possible to apply the following form of the Spectral Theorem.

Lemma 24.19. *Let V be a finite-dimensional $\mathbb{C}$-vector space equipped with a positive definite Hermitian form, and let $\alpha_1, \alpha_2, \ldots$ be a sequence of commuting Hermitian operators. Then $V = \bigoplus_i V_i$, where each V_i is an eigenspace of every α_n.*

Proof. The matrix for α_1 is Hermitian, therefore diagonalizable,[2] so we can decompose V as a direct sum of eigenspaces for α_1, writing $V = \bigoplus_i V(\lambda_i)$, where the λ_i are the distinct eigenvalues of α_1. Because α_1 and α_2 commute, α_2 must fix each subspace $V(\lambda_i)$, since for each $v \in V(\lambda_i)$ we have $\alpha_1\alpha_2 v = \alpha_2\alpha_1 v = \alpha_2\lambda_i v = \lambda_i\alpha_2 v$, and therefore $\alpha_2 v$ is an eigenvector for α_1 with eigenvalue λ_i, so $\alpha_2 v \in V(\lambda_i)$. Thus we can decompose each $V(\lambda_i)$ as a direct sum of eigenspaces for α_2, and may continue in this fashion for all the α_n. □

By Lemma 24.19, we may decompose $S_k(\Gamma_0(1)) = \bigoplus_i V_i$ as a direct sum of eigenspaces for the Hecke operators T_n. Let $f(\tau) = \sum a_n q^n$ be a nonzero element of V_i. We then have $a_1(T_n f) = a_n$, by Corollary 24.13, and also $T_n f = \lambda_n f$, for some eigenvalue λ_n of T_n which

[2]This fact is also sometimes called the Spectral Theorem and proved in most linear algebra courses.

is determined by V_i, so $a_n = \lambda_n a_1$. This implies $a_1 \neq 0$, since otherwise $f = 0$, and if we normalize f so that $a_1 = 1$ (which we can do, since f is nonzero and V_i is a $\mathbb{C}$-vector space), we then have $a_n = \lambda_n$ for all $n \geq 1$, and f completely determined by the sequence of Hecke eigenvalues λ_n for V_i. It follows that every element of V_i is a multiple of f, so $\dim V_i = 1$ and the eigenforms in $S_k(\Gamma_0(1))$ form a basis.

Theorem 24.20. *The space of cusp forms $S_k(\Gamma_0(1))$ is a direct sum of one-dimensional eigenspaces for the Hecke operators T_n and has a unique basis of eigenforms $f(\tau) = \sum a_n q^n$, where each a_n is the eigenvalue of T_n on the one-dimensional subspace spanned by f.*

The analog of Theorem 24.20 fails for $S_k(\Gamma_0(N))$ for two reasons, both of which are readily addressed. First, as in Remark 24.16, we need to restrict our attention to the Hecke operators T_n with $n \perp N$ (when n and N have a common factor T_n is not necessarily a Hermitian operator with respect to the Petersson inner product). We can then proceed as above to decompose $S_k(\Gamma_0(N))$ into eigenspaces for the Hecke operators T_n with $n \perp N$. We then encounter the second issue, which is that these eigenspaces need not be one-dimensional. In order to obtain a decomposition into one-dimensional eigenspaces we must restrict our attention to a particular subspace of $S_k(\Gamma_0(N))$.

Note that for any $M|N$ the space $S_k(\Gamma_0(M))$ is a subspace of $S_k(\Gamma_0(N))$ (since $\Gamma_0(M)$-invariance implies $\Gamma_0(N)$-invariance for $M|N$). We say that a cusp form $f \in S_k(\Gamma_0(N))$ is *old* if it also lies in the subspace $S_k(\Gamma_0(M))$ for some M properly dividing N. The oldforms in $S_k(\Gamma_0(N))$ generate a subspace $S_k^{\text{old}}(\Gamma_0(N))$, and we define $S_k^{\text{new}}(\Gamma_0(N))$ as the orthogonal complement of $S_k^{\text{old}}(\Gamma_0(N))$ in $S_k(\Gamma^0(N))$ (with respect to the Petersson inner product), so that

$$S_k(\Gamma_0(N)) = S_k^{\text{old}}(\Gamma_0(N)) \oplus S_k^{\text{new}}(\Gamma_0(N)),$$

and we call the eigenforms in $S_k^{\text{new}}(\Gamma_0(N))$ *newforms* (normalized so $a_1 = 1$). One can show that the Hecke operators T_n with $n \perp N$ preserve both $S_k^{\text{old}}(\Gamma_0(N))$ and $S_k^{\text{new}}(\Gamma_0(N))$. If we then decompose $S_k^{\text{new}}(\Gamma_0(N))$ into eigenspaces with respect to these operators, the resulting eigenspaces are all one-dimensional, moreover, each is actually generated by an eigenform (a simultaneous eigenvector for *all* the T_n, not just those with $n \perp N$ that we used to obtain the decomposition); this is a famous result of Atkin and Lehner [3, Thm. 5]. Note that $S_k^{\text{new}}(\Gamma_0(1)) = S_k(\Gamma_0(1))$, and we thus have the following generalization of Theorem 24.20.

Theorem 24.21. *The space $S_k^{\text{new}}(\Gamma_0(N))$ is a direct sum of one-dimensional eigenspaces for the Hecke operators T_n and has a unique basis of newforms $f(\tau) = \sum a_n q^n$, where each a_n is the eigenvalue of T_n on the one-dimensional subspace spanned by f.*

24.5 The *L*-function of a modular form

Our interest in cusp forms is that each has an associated *L-function*, which is defined in terms of a particular *Dirichlet series*.

Definition 24.22. A *Dirichlet series* is a series of the form

$$L(s) = \sum_{n \geq 1} a_n n^{-s},$$

where the a_n are complex numbers and s is a complex variable. Provided the a_n satisfy a polynomial growth bound of the form $|a_n| = O(n^\sigma)$ (as $n \to \infty$), the series $L(s)$ converges locally uniformly in the right half plane $\text{Re}(s) > 1 + \sigma$ and defines a holomorphic function

in this region (which may extend to a holomorphic or meromorphic function on a larger region).

Example 24.23. The most famous Dirichlet series is the *Riemann zeta function*

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$$

which converges locally uniformly to a holomorphic function on $\mathrm{re}(s) > 1$. It has three properties worth noting:

- **analytic continuation**: $\zeta(s)$ extends to a meromorphic function on $\mathbb{C}$ (with a simple pole at $s = 1$ and no other poles);
- **functional equation**: the *completed zeta function*[3] $\hat{\zeta}(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$ satisfies
$$\hat{\zeta}(s) = \hat{\zeta}(1-s);$$
- **Euler product**: we can write $\zeta(s)$ as a product over primes (for $\mathrm{re}(s) > 1$) via
$$\zeta(s) = \prod_p (1-p^{-s})^{-1} = \prod_p (1 + p^{-s} + p^{-2s} + \dots) = \sum_{n=1}^{\infty} n^{-s}.$$

Definition 24.24. The *L-function* (or *L-series*) of a cusp form $f(\tau) = \sum_{n=1}^{\infty} a_n q^n$ of weight k is the complex function defined by the Dirichlet series

$$L(f,s) := \sum_{n=1}^{\infty} a_n n^{-s},$$

which converges locally uniformly to a holomorphic function on $\mathrm{re}(s) > 1 + k/2$.

Theorem 24.25 (Hecke)**.** *Let $f \in S_k(\Gamma_0(N))$. The L-function $L(f,s)$ extends analytically to a holomorphic function on $\mathbb{C}$, and the normalized L-function*

$$\tilde{L}_f(s) = N^{s/2}(2\pi)^{-s}\Gamma(s)L(f,s)$$

satisfies the functional equation

$$\tilde{L}_f(s) = \pm\tilde{L}_f(k-s).$$

Remark 24.26. There are more explicit versions of this theorem that also determine the sign in the functional equation above.

For newforms we also get an Euler product.

Theorem 24.27. *Let $f \in S_k^{\mathrm{new}}(\Gamma_0(N))$. The L-function $L(f,s)$ has the Euler product*

$$L(f,s) = \sum_{n=1}^{\infty} a_n n^{-s} = \prod_p (1 - a_p p^{-s} + \chi(p)p^{k-1}p^{-2s})^{-1}, \tag{5}$$

where $\chi(p) = 0$ for $p|N$ and $\chi(p) = 1$ otherwise.

The function χ in Theorem 24.27 is the principal Dirichlet character of conductor N, a periodic function $\mathbb{Z} \to \mathbb{C}$ supported on $(\mathbb{Z}/N\mathbb{Z})^\times$ that defines a group homomorphism $(\mathbb{Z}/N\mathbb{Z})^\times \to \mathbb{C}$ (the adjective "principal" indicates that the homomorphism is trivial).

[3]Here $\Gamma(s) := \int_0^\infty e^{-t}t^{s-1}dt$ is Euler's gamma function.

24.6 The L-function of an elliptic curve

What does all this have to do with elliptic curves? Like eigenforms, elliptic curves over $\mathbb{Q}$ also have an L-function with an Euler product. In fact, with elliptic curves, we use the Euler product to define the L-function.

Definition 24.28. The *L-function of an elliptic curve* $E/\mathbb{Q}$ is given by the Euler product

$$L(E,s) = \prod_p L_p(p^{-s})^{-1} = \prod_p \left(1 - a_p p^{-s} + \chi(p) p p^{-2s}\right)^{-1}, \tag{6}$$

where $\chi(p)$ is 0 if E has bad reduction at p, and 1 otherwise.[4] For primes p where E has good reduction (all but finitely many), $a_p := p + 1 - \#E_p(\mathbb{F}_p)$ is the trace of Frobenius, where E_p denotes the reduction of E modulo p. Equivalently, $L_p(T)$ is the numerator of the zeta function

$$Z(E_p;T) = \exp\left(\sum_{n=1}^{\infty} \#E_p(\mathbb{F}_{p^n})\frac{T^n}{n}\right) = \frac{1 - a_p T + pT^2}{(1-T)(1-pT)},$$

that appeared in the special case of the Weil conjectures that you proved in Problem Set 7. For primes p where E has bad reduction, the polynomial $L_p(T)$ is defined by

$$L_p(T) = \begin{cases} 1 & \text{if } E \text{ has } \textit{additive} \text{ reduction at } p. \\ 1 - T & \text{if } E \text{ has } \textit{split multiplicative} \text{ reduction at } p. \\ 1 + T & \text{if } E \text{ has } \textit{non-split multiplicative} \text{ reduction at } p. \end{cases}$$

according to the type of bad reduction E has at p, as explained in the next section. This means that $a_p \in \{0, \pm 1\}$ at bad primes.

The L-function $L(E,s)$ converges to a holomorphic function on $\mathrm{re}(s) > 3/2$.

24.7 The reduction type of an elliptic curve

When computing $L(E,s)$, it is important to use a *minimal* Weierstrass equation for E, one that has good reduction at as many primes as possible. To see why this is necessary, note that if $y^2 = x^3 + Ax + B$ is a Weierstrass equation for E, then, up to isomorphism, so is $y^2 + u^4 Ax + u^6 B$, for any integer u, and this equation will have bad reduction at all primes $p|u$. Moreover, even though the equation $y^2 = x^3 + Ax + B$ always has bad reduction at 2, there may be an equation for E in general Weierstrass form that has good reduction at 2. For example, the elliptic curve defined by $y^2 = x^3 + 16$ is isomorphic to the elliptic curve defined by $y^2 + y = x^3$ (replace x by $4x$, divide by 64, and then replace y by $y + 1/2$), which does have good reduction at 2.

Definition 24.29. Let $E/\mathbb{Q}$ be an elliptic curve. A (global) *minimal model* for E is an integral Weierstrass equation

$$y^2 + a_1 xy + a_3 y = x^3 + a_2 x^2 + a_4 x + a_6,$$

with $a_1, a_2, a_3, a_4, a_6 \in \mathbb{Z}$ that defines an elliptic curve isomorphic to E whose discriminant $\Delta_{\min}(E)$ divides the discriminant of every integral Weierstrass equation for E.

[4] As explained in §24.7, this assumes we are using a minimal Weierstrass equation for E.

It is not immediately obvious that minimal models necessarily exist, but for elliptic curves over $\mathbb{Q}$ this is so; see [16, Prop. VII.1.3].[5] One can construct a minimal model in Sage using `E.minimal_model()`; see [9] for an explicit algorithm.

We now address the three types of bad reduction. To simplify the presentation, we will ignore the prime 2, but the three cases described below also occur at 2. For any odd prime p of bad reduction we can represent the singular curve $E_p/\mathbb{F}_p$ by an equation of the form $y^2 = f(x)$, for some cubic $f \in \mathbb{F}_p[x]$ that has a repeated root r. The repeated root r is necessarily rational, and by replacing x with $x-r$ we can assume $r = 0$, so $y^2 = x^3 + ax^2$ for some $a \in \mathbb{F}_p$. The projective curve $y^2z = x^3 + ax^2z$ has exactly one singular point $(0 : 0 : 1)$ and is smooth elsewhere (including the point $(0 : 1 : 0)$ at infinity).

If we exclude the singular point $(0 : 0 : 1)$, the standard formulas for the group law on $E_p(\mathbb{F}_p)$ still make sense, and the set

$$E_p^{\mathrm{ns}}(\mathbb{F}_p) := E_p(\mathbb{F}_p) - \{(0 : 0 : 1)\}$$

of non-singular points of $E_p(\mathbb{F}_p)$ is closed under the group operation.[6] Thus $E_p^{\mathrm{ns}}(\mathbb{F}_p)$ is a finite abelian group. We now define

$$a_p := p - \#E_p^{\mathrm{ns}}(\mathbb{F}_p).$$

This is analogous to the good reduction case in which $a_p = p + 1 - \#E_p(\mathbb{F}_p)$; we have removed the (necessarily rational) singular point, so we reduce a_p by one.

There are two cases to consider, depending on whether $f(x)$ has a double or triple root at 0; these two cases give rise to three possibilities for the group $E_p^{\mathrm{ns}}(\mathbb{F}_p)$.

- **Case 1: triple root** ($y^2 = x^3$)

  We have the projective curve $zy^2 = x^3$. After removing the singular point $(0 : 0 : 1)$, every other projective point has non-zero y coordinate, so we can fix $y = 1$, and work with the affine curve $z = x^3$. There are p solutions to this equation (including $x = 0$ and $z = 0$, which corresponds to the projective point $(0 : 1 : 0)$ at infinity). It follows that $E_p^{\mathrm{ns}}(\mathbb{F}_p)$ is a cyclic group of order p, which is necessarily isomorphic to the additive group of $\mathbb{F}_p$; see [18, §2.10] for an explicit isomorphism. In this case we have $a_p = 0$ and say that E has *additive reduction* at p.

- **Case 2: double root** ($y^2 = x^3 + ax^2$, $a \neq 0$).

  We have the projective curve $zy^2 = x^3 + ax^2z$, and the point $(0 : 1 : 0)$ at infinity is the only non-singular point on the curve whose x or z coordinate is zero. Excluding the point at infinity for the moment, let us divide both sides by x^2, introduce the variable $t = y/x$, and fix $z = 1$. This yields the affine curve $t^2 = x + a$, and the number of

[5] For an elliptic curve E over a number field K one defines $\Delta_{\min}(E)$ as the $\mathcal{O}_K$-ideal generated by the discriminants of all integral models for E (with $a_1, a_2, a_3, a_4, a_6 \in \mathcal{O}_K$); if the class number of $\mathcal{O}_K$ is greater than one this ideal need not be a principal ideal, in which case E cannot have a minimal model over K.

[6] To see this geometrically, note that any line in $\mathbb{P}^2$ intersecting a plane cubic in two non-singular points cannot also intersect it in a singular point; when we count intersections with multiplicity the total must be three, by Bezout's theorem, but singular points contribute multiplicity greater than one.

points with $x \neq 0$ is

$$\begin{aligned}\sum_{x\neq 0}\left(1+\left(\frac{x+a}{p}\right)\right) &= \sum_{x}\left(1+\left(\frac{x+a}{p}\right)\right)-\left(1+\left(\frac{a}{p}\right)\right)\\ &= \sum_{x}\left(1+\left(\frac{x}{p}\right)\right)-1-\left(\frac{a}{p}\right)\\ &= p-1-\left(\frac{a}{p}\right)\end{aligned}$$

where $\left(\frac{a}{p}\right)$ is the Kronecker symbol. If we now add the point at infinity into our total we get $p-\left(\frac{a}{p}\right)$, so $a_p = p-(p-\left(\frac{a}{p}\right)) = \left(\frac{a}{p}\right) = \pm 1$. In this case we say that E has *multiplicative reduction* at p, and distinguish the cases $a_p = 1$ and $a_p = -1$ as *split* and *non-split* respectively. One can show that in the split case $E_p^{\mathrm{ns}}(\mathbb{F}_p)$ is isomorphic to the multiplicative group $\mathbb{F}_p^\times$, and in the non-split case it is isomorphic to the multiplicative subgroup of $\mathbb{F}_{p^2} = \mathbb{F}_p[x]/(x^2-a)$ consisting of the norm 1 elements; see [18, §2.10].

To sum up, there are three possibilities for $a_p = p - \#E_p^{\mathrm{ns}}(\mathbb{F}_p)$:

$$a_p = \begin{cases} 0 & \text{additive reduction,}\\ +1 & \text{split multiplicative reduction,}\\ -1 & \text{non-split multiplicative reduction.}\end{cases}$$

It can happen that the reduction type of E changes when we consider E as an elliptic curve over a finite extension $K/\mathbb{Q}$ (in which case we are then talking about reduction modulo primes $\mathfrak{p}$ of K lying above p). It turns out that this can only happen when E has additive reduction at p, which leads to the following definition.

Definition 24.30. An elliptic curve $E/\mathbb{Q}$ is *semistable* if it does not have additive reduction at any prime.

As we shall see in the next lecture, for the purposes of proving Fermat's Last Theorem, we can restrict our attention to semistable elliptic curves.

24.8 *L*-functions of elliptic curves versus *L*-functions of modular forms

Although we defined the L-function of an elliptic curve using an Euler product, we can always expand this product to obtain a Dirichlet series

$$L(E,s) = \prod_p \left(1-a_p p^{-s}+\chi(p)pp^{-2s}\right)^{-1} = \sum_{n=1}^{\infty} a_n n^{-s}.$$

We now observe that the integer coefficients a_n in the Dirichlet series for $L(E,s)$ satisfy the recurrence relations listed in (3) for an eigenform of weight $k=2$. We have $a_1 = 1$, $a_{mn} = a_m a_n$ for $m \perp n$, and $a_{p^{r+1}} = a_p a_{p^r} - p a_{p^{r-1}}$ for all primes p of good reduction, as you proved on Problem Set 7. For the primes of bad reduction we have $a_p \in \{0, \pm 1\}$ and it is easy to check that $a_{p^r} = a_p^r$, which applies to the coefficients of an eigenform in $S_k^{\mathrm{new}}(\Gamma_0(N))$ when $p|N$ (see Remark 24.16).

So it now makes sense to ask, given an elliptic curve $E/\mathbb{Q}$, is there a modular form f for which $L(E,s) = L(f,s)$? Or, to put it more simply, let $L(E,s) = \sum_{n=1}^{\infty} a_n n^{-s}$, and define

$$f_E(\tau) = \sum_{n=1}^{\infty} a_n q^n \qquad (q := e^{2\pi i \tau})$$

Our question then becomes: is $f_E(\tau)$ a modular form?

It's clear from the recurrence relation for a_{p^r} that if $f_E(\tau)$ is a modular form, then it must be a modular form of weight 2; but there are additional constraints. For $k = 2$ the equations (5) and (6) both give the Euler product

$$\prod_p \left(1 - a_p p^{-s} + \chi(p) p p^{-2s}\right)^{-1},$$

and it is essential that $\chi(p)$ is the same in both cases. For newforms $f \in S_k^{\text{new}}(\Gamma_0(N))$ we have $\chi(p) = 0$ for primes $p|N$, while for elliptic curves $E/\mathbb{Q}$ we have $\chi(p) = 0$ for primes $p|\Delta_{\min}(E)$. No elliptic curve over $\mathbb{Q}$ has good reduction at every prime, so we cannot use eigenforms of level 1, we need to consider newforms of some level $N > 1$.

This suggests we take N to be the product of the prime divisors of $\Delta_{\min}(E)$, but note that any N with the same set of prime divisors would have the same property, so this doesn't uniquely determine N. For semistable elliptic curves, it turns out that taking the product of the prime divisors of $\Delta_{\min}(E)$ is the correct choice, and this is all we need for the proof of Fermat's Last Theorem.

Definition 24.31. Let $E/\mathbb{Q}$ be a semistable elliptic curve. The *conductor* N_E of E is the product of the prime divisors of its minimal discriminant $\Delta_{\min}(E)$.

In general, the conductor N_E of an elliptic curve $E/\mathbb{Q}$ is always divisible by the product of the primes $p|\Delta_{\min}(E)$, and N_E is squarefree if and only if E is semistable. For primes p where E has multiplicative reduction (split or non-split) $p|N_E$ but $p^2 \nmid N_E$, and when E has additive reduction at p then $p^2|N_E$ and if $p > 3$ then $p^3 \nmid N_E$. The primes 2 and 3 require special treatment (as usual): the maximal power of 2 dividing N_E may be as large as 2^8, and the maximal power of 3 dividing N_E may be as large as 3^5, see [17, IV.10] for the details, which are slightly technical.

We can now say precisely what it means for an elliptic curve over $\mathbb{Q}$ to be modular.

Definition 24.32. An elliptic curve $E/\mathbb{Q}$ is *modular* if f_E is a modular form.

If $E/\mathbb{Q}$ is modular, the modular form f_E is necessarily a newform in $S_2^{\text{new}}(\Gamma_0(N_E))$ with an integral q-expansion; this follows from the Eichler-Shimura Theorem (see Theorem 24.37).

Theorem 24.33 (Modularity Theorem)**.** *Every elliptic curve $E/\mathbb{Q}$ is modular.*

Proof. This is proved in [4], which extends the results in [19, 20] to all elliptic curves $E/\mathbb{Q}$. □

Prior to its proof, the conjecture that every elliptic curve $E/\mathbb{Q}$ is modular was variously known as the Shimura-Taniyama-Weil conjecture, the Taniyama-Shimura-Weil conjecture, the Taniyama-Shimura conjecture, the Shimura-Taniyama conjecture, the Taniyama-Weil conjecture, or the Modularity Conjecture, depending on the author. Thankfully, everyone is now happy to call it the Modularity Theorem!

24.9 BSD and the parity conjecture

When E is modular, the L-function of E is necessarily the L-function of a modular form, and this implies that $L(E,s)$ has an analytic continuation and satisfies a functional equation, since this holds for the L-function of a modular form, by Theorem 24.25. Prior to the proof of the modularity theorem, this was an open question known as the Hasse-Weil conjecture; we record it here as a corollary to the Modularity Theorem.

Corollary 24.34. *Let E be an elliptic curve over $\mathbb{Q}$. Then $L(E,s)$ has an analytic continuation to a holomorphic function on $\mathbb{C}$, and the normalized L-function*

$$\tilde{L}_E(s) := N_E^{s/2}(2\pi)^{-s}\Gamma(s)L(E,s)$$

satisfies the functional equation

$$\tilde{L}_E(s) = w_E\tilde{L}_E(2-s),$$

where $w_E = \pm 1$.

The sign w_E in the functional equation is called the *root number* of E. If $w_E = -1$ then the functional equation implies that $\tilde{L}_E(s)$, and therefore $L(E,s)$, has a zero at $s=1$; in fact it is easy to show that $w_E = 1$ if and only if $L(E,s)$ has a zero of even order at $s=1$.

The conjecture of Birch and Swinnerton-Dyer (BSD) relates the order of vanishing of $L(E,s)$ at $s=1$ to the rank of $E(\mathbb{Q})$. Recall that

$$E(\mathbb{Q}) \simeq E(\mathbb{Q})_{\mathrm{tor}} \times \mathbb{Z}^r,$$

where $E(\mathbb{Q})_{\mathrm{tor}}$ denotes the torsion subgroup of $E(\mathbb{Q})$ and r is the *rank* of E.

Conjecture 24.35 (Weak BSD)**.** *Let $E/\mathbb{Q}$ be an elliptic curve of rank r. Then $L(E,s)$ has a zero of order r at $s=1$.*

The strong version of the BSD conjecture makes a more precise statement that expresses the leading coefficient of the Taylor expansion of $L(E,s)$ at $s=1$ in terms of various invariants of E. A proof of even the weak form of the BSD conjecture is enough to claim the Millennium Prize offered by the Clay Mathematics Institute. There is also the Parity Conjecture, which simply relates the root number w_E in the functional equation for $L(E,s)$ to the parity of r as implied by the BSD conjecture.

Conjecture 24.36 (Parity Conjecture)**.** *Let $E/\mathbb{Q}$ be an elliptic curve of rank r. Then the root number is given by $w_E = (-1)^r$.*

24.10 Modular elliptic curves

The relationship between elliptic curves and modular forms is remarkable and not at all obvious. It is reasonable to ask why people believed the modular conjecture in the first place. Probably the most compelling reason is that every newform of weight 2 with an integral q-series gives rise to an elliptic curve $E/\mathbb{Q}$.

Theorem 24.37 (Eichler-Shimura, Carayol)**.** *Let $f = \sum a_n q^n \in S_2^{\mathrm{new}}(\Gamma_0(N))$ be a newform with $a_n \in \mathbb{Z}$. There exists an elliptic curve $E/\mathbb{Q}$ of conductor N for which $f_E = f$.*

See [10, V.6] for an overview of how to construct the elliptic curve given by the theorem, which was known long before the modularity theorem was proved.[7] For a more detailed (but still very accessible) exposition, see [12].

The elliptic curve E whose existence is guaranteed by the Eichler-Shimura theorem is determined only up to isogeny.[8] This is due to the fact that isogenous elliptic curves E and E' over $\mathbb{Q}$ necessarily have the same L-function, which implies $f_E = f_{E'}$. If E and E' are isogenous over $\mathbb{Q}$ then their reductions modulo any prime p where they both have good reduction are necessarily isogenous, and as you showed on Problem Set 7, they must have the same trace of Frobenius a_p; it turns out that in fact E and E' must have the same reduction type at every prime so their L-functions are actually identical. The converse also holds; in fact, something even stronger is true; this follows from work begun by Tate and completed by Faltings in 1983 [7]; see [10, Thm. V.4.1] for further details.

Theorem 24.38 (Faltings-Tate)**.** *Let E and E' be elliptic curves over $\mathbb{Q}$ with L-function $L(E,s) = \sum a_n n^{-s}$ and $L(E',s) = \sum a'_n n^{-s}$, respectively. If $a_p = a'_p$ for sufficiently many primes p of good reduction for E and E', then E and E' are isogenous.*

What "sufficiently many" means depends on E and E', but it is a finite number. In particular, all but finitely many is always enough, which is all we need for the next lecture.

Corollary 24.39. *Elliptic curves $E, E'/\mathbb{Q}$ are isogenous if and only if $L(E,s) = L(E',s)$, equivalently, if and only if E_p and E'_p are isogenous modulo sufficiently many good primes p.*

The fact that isogenous elliptic curves have the same L-functions while distinct newforms have distinct L-functions means that the correspondence between elliptic curves and weight-2 newforms with $a_n \in \mathbb{Z}$ is many-to-one, not one-to-one; there can be up to 8 isomorphism classes of elliptic curves $E/\mathbb{Q}$ in the same isogeny class (but no more than 8, this is a result of Kenku [8]). But the modularity theorem implies that there is a one-to-one correspondence between isogeny classes of elliptic curves over $\mathbb{Q}$ and weight-2 newforms with $a_n \in \mathbb{Z}$.

For any given value of N, one can effectively enumerate the newforms in $S_2^{\text{new}}(\Gamma_0(N))$ with integral q-expansions; this is a finite list. It is also possible (but not easy)[9] to determine the isogeny classes of all elliptic curves of a given conductor N for suitable values of N, without assuming these elliptic curves are modular; this is also a finite list. When this was done for many small values of N, it was found that the two lists always matched perfectly. It was this matching that made the modularity conjecture truly compelling. Much of this matching was done before Theorems 24.37 and 24.38 had been completely proved, but they were both conjectured (and partially proved) much earlier.

References

[1] M. K. Agrawal, John H. Coates, David C. Hunt, Alfred J. van der Poorten, *Elliptic curves of conductor 11*, Math. Comp. **35** (1980), 991–1002.

[2] Amod Agashe, Kenneth Ribet, and William A. Stein, *The Manin constant*, Pure and Applied Mathematics Quarterly **2** (2006), 617–636.

[7] The original results of Eichler and Shimura [14] proved $a_p(E) = a_p(f)$ only for primes of good reduction and did not address the correspondence between the level and the conductor. The correspondence between the level and conductor was conjectured by Weil but not rigorously proved until 1986 by Carayol [5, §0.8].

[8] But there is an *optimal* representative for each isogeny class; see John Cremona's appendix to [2].

[9] This requires enumerating all solutions to certain Diophantine equations; see [1] and [11] for examples.

[3] A.O.L. Atkin and J. Lehner, *Hecke operators on $\Gamma_0(m)$*, Mathematische Annalen **185** (1970), 134–160.

[4] Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor, *On the modularity of elliptic curves over $\mathbb{Q}$: wild 3-adic exercises*, Journal of the AMS **14** (2001), 843–939.

[5] Henri Carayol, *Sur les représentations l-adiques associées aux formes modulaires de Hilbert*, Ann. Sci. École Norm. Sup. (4) **19** (1986), 409–468.

[6] Fred Diamond and Jerry Shurman, *A first course in modular forms*, Springer, 2005.

[7] Gerd Faltings, *Finiteness theorems for abelian varieties over number fields*, Inventiones **73** (1983), 349–366.

[8] M. A. Kenku, *On the number of $\mathbb{Q}$-isomorphism classes of elliptic curves in each $\mathbb{Q}$-isogeny class*, Journal of Number Theory **15** (1982), 199–202.

[9] Michael Laska, *An algorithm for finding a minimal Weierstrass equation for an elliptic curve*, Mathematics of Computation **38** (1982), 257-260.

[10] J. S. Milne, *Elliptic curves*, BookSurge Publishers, 2006.

[11] Andrew P. Ogg, *Abelian curves of small conductor*, J. Reine Angew. Math. **224** (1967), 204–215.

[12] Corentin Perent-Gentil, *Associating abelian varieties to weight-2 modular forms: the Eichler-Shimura construction*, Master's thesis, EPF Lausanne, 2014.

[13] Jean-Pierre Serre, *A course in arithmetic*, Springer, 1973.

[14] Goro Shimura, *Correspondances modulaires et les fonctions ζ de courbes algébriques*, Journal of the Mathematical Society of Japan, **10** (1958), 1–28.

[15] Goro Shimura, *Introduction to the arithmetic theory of automorphic functions*, Publications of the Mathematical Society of Japan **11**, 1971.

[16] Joseph H. Silverman, *The arithmetic of elliptic curves*, second edition, Springer, 2009.

[17] Joseph H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Springer, 1994.

[18] Lawrence C. Washington, *Elliptic curves: Number theory and cryptography*, second edition, Chapman and Hall/CRC, 2008.

[19] Richard Taylor and Andrew Wiles, *Ring-theoretic properties of certain Hecke algebras*, Annals of Mathematics **141** (1995), 553–572.

[20] Andrew Wiles, *Modular elliptic curves and Fermat's last theorem*, Annals of Mathematics **141** (1995), 443-551.

25 Fermat's Last Theorem

In this final lecture we give an overview of the proof of Fermat's Last Theorem. Our goal is to explain what Andrew Wiles [23], with the assistance of Richard Taylor [21], proved, and why it implies Fermat's Last Theorem. This implication is a consequence of earlier work by several mathematicians, including Richard Frey, Jean-Pierre Serre, and Ken Ribet. We will say very little about the details of Wiles' proof, which are beyond the scope of this course, but we will provide references for those who wish to learn more.

25.1 Fermat's Last Theorem

In 1637, Pierre de Fermat famously wrote in the margin of a copy of Diophantus' *Arithmetica* that the equation

$$x^n + y^n = z^n$$

has no integer solutions with $xyz \neq 0$ and $n > 2$, and claimed to have a remarkable proof of this fact. As with most of Fermat's work, he never published this claim (mathematics was a hobby for Fermat, he was a lawyer by trade). Fermat's marginal comment was apparently discovered only after his death, when his son Samuel was preparing to publish Fermat's mathematical correspondence, but it soon became well known and is included as commentary in later printings of *Arithmetica*.

Fermat did prove the case $n = 4$, using a descent argument. It then suffices to consider only cases where n is an odd prime, since if $p|n$ and (x_0, y_0, z_0) is a solution to $x^n + y^n = z^n$, then $(x_0^{n/p}, y_0^{n/p}, z_0^{n/p})$ is a solution to $x^p + y^p = z^p$.

A brief chronology of the progress made toward proving Fermat's Last Theorem prior to Wiles' work is listed below.

1637	Fermat makes his conjecture and proves it for $n = 4$.
1753	Euler proves FLT for $n = 3$ (his proof has a fixable error).
1800s	Sophie Germain proves FLT for $n \nmid xyz$ for all $n < 100$.
1825	Dirichlet and Legendre complete the proof for $n = 5$.
1839	Lamé addresses $n = 7$.
1847	Kummer proves FLT for all primes $n \nmid h(\mathbb{Q}(\zeta_n))$, called *regular* primes. This leaves 37, 59, and 67 as the only open cases for $n < 100$.
1857	Kummer addresses 37, 59, and 67, but his proof has gaps.
1926	Vandiver fills the gaps and addresses all irregular primes $n < 157$.
1937	Vandiver and assistants handle all irregular primes $n < 607$.
1954	Lehmer, Lehmer, and Vandiver introduce techniques better suited to mechanical computation and use a computer to address all $n < 2521$.
1954-1993	Computers verify FLT for all $n < 4,000,000$.

All of the results above are based on work in algebraic number theory, none of it uses elliptic curves.[1] The first person to suggest a connection between elliptic curves and Fermat's Last Theorem was Yves Hellegouarch. In his 1972 doctoral thesis [7], Hellegouarch associates

[1] Work in this direction continued even after FLT was proved. We now know that the Kummer-Vandiver conjecture $p \nmid h(\mathbb{Q}(\zeta_p)^+)$ holds for $p \leq 2^{31}$ [6]. This conjecture is a key ingredient to approaches to proving FLT using algebraic number theory (in particular, the theory of cyclotomic fields); see [22, Ch. 9] for details. We still do not know if the Kummer-Vandiver conjecture is true or not (but we do know FLT is true).

to any non-trivial solution (a,b,c) of $x^p + y^p = z^p$, with p an odd prime, the elliptic curve

$$E_{a,b,c}\colon \qquad y^2 = x(x-a^p)(x+b^p).$$

Without loss of generality we assume $\gcd(a,b,c) = 1$, which implies that a, b, c must be pairwise relatively prime, and that $a \equiv 3 \bmod 4$ and $b \equiv 0 \bmod 2$ (we can always swap a and b and/or multiply both sides by -1 in order to achieve this). Proving Fermat's Last Theorem then amounts to showing that no such elliptic curve $E_{a,b,c}$ can exist.

Hellegouarch did not make much progress with this, but in 1984 Gerhard Frey suggested that the elliptic curve $E_{a,b,c}$, if it existed, could not possibly be modular [5]. Shortly thereafter, Jean-Pierre Serre [16] reduced Frey's conjecture to a much more precise statement about modular forms and Galois representations, known as the *epsilon conjecture*, which was proved by Ken Ribet a few years later [14]. With Ribet's result in hand, it was then known that the modularity conjecture, which states that every elliptic curve over $\mathbb{Q}$ is modular, implies Fermat's Last Theorem: it guarantees that $E_{a,b,c}$, and therefore the solution (a,b,c) to $x^p + y^p = z^p$, cannot exist. At that time no one expected the modularity conjecture to be proved any time soon; indeed, the fact that it implies Fermat's Last Theorem was taken as evidence of how difficult it would be to prove the modularity conjecture.

25.2 A strange elliptic curve

To get a sense of what makes the elliptic curve $E_{a,b,c}$ so strange that one might question its very existence, let us compute its discriminant:

$$\Delta(E_{a,b,c}) = -16(0-a^p)^2(0+b^p)^2(a^p+b^p)^2 = -16(abc)^{2p}.$$

As explained in the last lecture, the definition of the L-series of an elliptic curve E requires us to determine the minimal discriminant of E and its reduction type at each prime dividing the minimal discriminant (additive, split multiplicative, or non-split multiplicative). It turns out that the discriminant Δ is not quite minimal, the minimal discriminant is

$$\Delta_{\min}(E_{a,b,c}) = 2^{-8}(abc)^{2p},$$

(assuming $p > 3$, which we know must be the case), which differs from Δ only at 2.

On the other hand, the conductor of $E_{a,b,c}$ is much smaller than its minimal discriminant. Recall from the previous lecture that for odd primes ℓ an elliptic curve $E : y^2 = f(x)$ can have additive reduction at ℓ only if the cubic $f \in \mathbb{Z}[x]$ has a triple root modulo ℓ. This is clearly not the case for the curve $E_{a,b,c}\colon y^2 = f(x) = x(x-a^p)(x+b^p)$, since 0 is always a root modulo ℓ, but a and b are relatively prime and cannot both be divisible by ℓ, so 0 is not a triple root. One can also show that $E_{a,b,c}$ does not have additive reduction at 2. This implies that $E_{a,b,c}$ is semistable, so its conductor is the squarefree integer

$$N_{E_{a,b,c}} = \prod_{\ell \mid abc} \ell,$$

which we note is divisible by 2 (since b is).

For the elliptic curve $E_{a,b,c}$ the ratio $\Delta_{\min}(E_{a,b,c})/N_{E_{a,b,c}}$ grows exponentially with p. But it is very unusual (conjecturally impossible) for the minimal discriminant of an elliptic

curve to be so much larger than its conductor. Szpiro's conjecture [18], which is closely related to the ABC conjecture,[2] states that for every $\epsilon > 0$ there is a constant c_ϵ such that

$$\Delta_{\min}(E) \le c_\epsilon N_E^{6+\epsilon}$$

for every elliptic curve $E/\mathbb{Q}$. This cannot possibly be true for $E_{a,b,c}$ if p is sufficiently large. This does not imply that $E_{a,b,c}$ cannot be modular, but it suggests that there is something very strange about this elliptic curve (so strange that one might expect it cannot exist).

25.3 Galois representations

Let E be an elliptic curve over $\mathbb{Q}$, let ℓ be a prime, and let $K := \mathbb{Q}(E[\ell])$ be its *ℓ-torsion field*, the extension of $\mathbb{Q}$ obtained by adjoining the coordinates of all the points in $E[\ell]$ to $\mathbb{Q}$. The field K is a Galois extension of $\mathbb{Q}$ (it is either the splitting field of the ℓth division polynomial, or a quadratic extension of it), and its Galois group acts on the ℓ-torsion subgroup $E[\ell]$ via its action on the coordinates of each point. This yields a group representation

$$\rho\colon \mathrm{Gal}(K/\mathbb{Q}) \to \mathrm{Aut}(E[\ell]) \simeq \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z}),$$

that maps each $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ to the automorphism of $E[\ell] \simeq \mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z}$ given by applying σ to the coordinates of each ℓ-torsion point (all of which lie in $K = \mathbb{Q}(E[\ell])$, by definition). We consider two representations $\rho, \rho'\colon \mathrm{Gal}(K/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$ to be isomorphic if there exists $A \in \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$ such that $\rho'(\sigma) = A\rho(\sigma)A^{-1}$ for all $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$, in which case we write $\rho \simeq \rho'$.

Let S be the finite set of primes consisting of ℓ and the primes of bad reduction for E. Every prime $p \notin S$ is unramified in K. As explained in Lecture 20, this means that the $\mathcal{O}_K$-ideal generated by p factors into a product of distinct prime ideals:

$$p\mathcal{O}_K = \mathfrak{p}_1 \cdots \mathfrak{p}_r.$$

The Galois group $\mathrm{Gal}(K/\mathbb{Q})$ acts transitively on the set $\{\mathfrak{p}|p\} := \{\mathfrak{p}_1, \dots, \mathfrak{p}_r\}$, and for each prime ideal $\mathfrak{p}|p$ we have a corresponding *decomposition group*

$$D_\mathfrak{p} := \{\sigma \in \mathrm{Gal}(K/\mathbb{Q}) : \sigma(\mathfrak{p}) = \mathfrak{p}\}$$

equipped with an isomorphism

$$\begin{aligned} \varphi\colon D_\mathfrak{p} &\xrightarrow{\sim} \mathrm{Gal}(\mathbb{F}_\mathfrak{p}/\mathbb{F}_p) \\ \sigma &\mapsto \bar\sigma \end{aligned}$$

where $\mathbb{F}_\mathfrak{p} := \mathcal{O}_K/\mathfrak{p}$ is the residue field at $\mathfrak{p}$ and the automorphism $\bar\sigma$ is defined by $\bar\sigma(\bar x) = \overline{\sigma(x)}$, where $\bar x$ denotes the image of $x \in \mathcal{O}_K$ in the quotient $\mathcal{O}_K/\mathfrak{p} = \mathbb{F}_\mathfrak{p}$. The Galois group $\mathrm{Gal}(\mathbb{F}_\mathfrak{p}/\mathbb{F}_p)$ is cyclic, generated by the p-power Frobenius automorphism $\pi_p\colon x \mapsto x^p$, and we define the *Frobenius element*

$$\mathrm{Frob}_\mathfrak{p} := \varphi^{-1}(\pi_p) \in D_\mathfrak{p} \subseteq \mathrm{Gal}(K/\mathbb{Q}).$$

[2] The ABC conjecture states that for all $\epsilon > 0$ there is a constant c_ϵ such that only finitely many integer solutions to $a + b = c$ satisfy $\mathrm{rad}(abc)^{1+\epsilon} < c_\epsilon$, where $\mathrm{rad}(abc)$ denotes the squarefree part of abc. This is equivalent to a modified version of Szpiro's conjecture in which one replaces $\Delta_{\min}(E)$ with $\max(|A|^3, B^2)$, where A and B are the coefficients in a short Weierstrass equation for $E\colon y^2 = x^3 + Ax + B$. Mochizuki announced a proof of the ABC conjecture in 2012 that was finally published in 2021, but as of this writing, most number theorists do not consider the ABC conjecture to have been proved.

Different choices of $\mathfrak{p}|p$ yield conjugate $\mathrm{Frob}_\mathfrak{p}$ (and every conjugate of $\mathrm{Frob}_\mathfrak{p}$ arises for $\mathfrak{p}|p$), and we let Frob_p denote this conjugacy class; as an abuse of terminology we may speak of the Frobenius element Frob_p as an element of $\mathrm{Gal}(K/\mathbb{Q})$ representing this conjugacy class, with the understanding that Frob_p is determined only up to conjugacy.

Thus for each prime $p \notin S$ we get a Frobenius element $\mathrm{Frob}_p \in \mathrm{Gal}(K/\mathbb{Q})$, and may consider its image $A_p := \rho(\mathrm{Frob}_p) \in \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$ under the Galois representation ρ. The characteristic polynomial of A_p (which depends only on the conjugacy class of Frob_p) is

$$\det(\lambda I - A_p) = \lambda^2 - (\operatorname{tr} A_p)\lambda + \det A_p,$$

with

$$\operatorname{tr} A_p \equiv a_p \bmod \ell \qquad \text{and} \qquad \det A_p \equiv p \bmod \ell.$$

Here $a_p := p + 1 - \#E_p(\mathbb{F}_p)$ is the trace of the Frobenius endomorphism of the reduction $E_p/\mathbb{F}_p$ of E modulo p, equivalently, the pth coefficient in the Dirichlet series of the L-function $L(E,s) = \sum_{n\geq 1} a_n n^{-s}$ of the elliptic curve E.

For any positive integer n we can similarly consider the Galois representation

$$\rho\colon \mathrm{Gal}(\mathbb{Q}(E[\ell^n])/\mathbb{Q}) \to \mathrm{Aut}(E[\ell^n]) \simeq \mathrm{GL}_2(\mathbb{Z}/\ell^n\mathbb{Z}).$$

For primes $p \notin S$ with $4\sqrt{p} \leq \ell^n$, the value of the integer $a_p \equiv \operatorname{tr}\rho(\mathrm{Frob}_p) \bmod \ell^n$ is uniquely determined. Note that this holds no matter which auxiliary prime ℓ we pick.

The discussion above applies not only to $\mathbb{Q}(E[\ell^n])$, but to any Galois extension K of $\mathbb{Q}$ containing $\mathbb{Q}(E[\ell^n])$. Even if the extension $K/\mathbb{Q}$ is ramified at primes outside of S, the image of $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$ under ρ depends only on the restriction of the automorphism σ to $\mathbb{Q}(E[\ell^n])$, so given a Galois representation $\rho\colon \mathrm{Gal}(K/\mathbb{Q}) \to \mathrm{Aut}(E[\ell^n]) \simeq \mathrm{GL}_2(\mathbb{Z}/\ell^n\mathbb{Z})$ we can determine $\rho(\mathrm{Frob}_p) \in \mathrm{GL}_2(\mathbb{Z}/\ell^n\mathbb{Z})$ up to conjugacy. Here we use $\mathrm{Frob}_p \in \mathrm{Gal}(K/\mathbb{Q})$ to denote any element whose restriction to $\mathrm{Gal}(\mathbb{Q}(E[\ell^n])/\mathbb{Q})$ lies in the conjugacy class represented by the Frobenius element $\mathrm{Frob}_p \in \mathrm{Gal}(\mathbb{Q}(E[\ell^n])/\mathbb{Q})$. The conjugacy class of $\rho(\mathrm{Frob}_p)$ in $\mathrm{GL}_2(\mathbb{Z}/\ell^n\mathbb{Z})$, and in particular its trace, is independent of this choice.

We now define the *ℓ-adic Tate module*

$$T_\ell(E) \coloneqq \varprojlim_n E[\ell^n]$$

as the projective limit of the inverse system

$$E[\ell] \xleftarrow{[\ell]} E[\ell^2] \xleftarrow{[\ell]} \cdots \xleftarrow{[\ell]} E[\ell^n] \xleftarrow{[\ell]} E[\ell^{n+1}] \xleftarrow{[\ell]} \cdots,$$

whose connecting homomorphisms are multiplication-by-ℓ maps. Elements of $T_\ell(E)$ are infinite sequences of points $(P_1, P_2, P_3, \ldots)$ with $P_n \in E[\ell^n]$ such that $\ell P_{n+1} = P_n$.

We now let $G_\mathbb{Q} := \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and define the *ℓ-adic Galois representation*

$$\rho_{E,\ell}\colon G_\mathbb{Q} \to \mathrm{Aut}(T_\ell(E)) \simeq \mathrm{GL}_2(\mathbb{Z}_\ell),$$

where $\mathbb{Z}_\ell = \varprojlim \mathbb{Z}/\ell^n\mathbb{Z}$ is the *ring of ℓ-adic integers*, which contains $\mathbb{Z}$ as a subring.[3] Each $\sigma \in G_\mathbb{Q}$ acts on $(P_1, P_2, P_3, \ldots) \in T_\ell(E)$ via its action on the coordinates of each $P_n \in E[\ell^n]$.

[3]You can view elements of $\mathbb{Z}_\ell$ as infinite sequences of integers $(a_1, a_2, a_3, \ldots)$ with $a_n \equiv a_{n+1} \bmod \ell^n$, and ring operations defined coordinate-wise. We embed $\mathbb{Z}$ in $\mathbb{Z}_\ell$ via the map $a \mapsto (a, a, a, \ldots)$. Note that $\mathbb{Z}_\ell$ has characteristic 0 but comes equipped with reduction maps to the positive characteristic rings $\mathbb{Z}/\ell^n\mathbb{Z}$.

For primes $p \notin S$ we now use $\mathrm{Frob}_p \in G_{\mathbb{Q}}$ to denote an element whose restriction to $\mathrm{Gal}(\mathbb{Q}(E[\ell^n])/\mathbb{Q})$ is conjugate to $\mathrm{Frob}_p \in \mathrm{Gal}(\mathbb{Q}(E[\ell^n])/\mathbb{Q})$ for each $n \geq 1$; this amounts to choosing a compatible sequence of Frobenius elements $\mathrm{Frob}_{p,n} \in \mathrm{Gal}(\mathbb{Q}(E[\ell^n])/\mathbb{Q})$ such that $\mathrm{Frob}_{p,n}$ is the restriction of $\mathrm{Frob}_{p,n+1}$ to $\mathbb{Q}(E[\ell^n])$. The conjugacy class of $\rho(\mathrm{Frob}_p)$ in $\mathrm{GL}_2(\mathbb{Z}_\ell)$ is independent of these choices; in particular its trace in $\mathbb{Z}_\ell$ is well defined.

We then have $\operatorname{tr}\rho_{E,\ell}(\mathrm{Frob}_p) = a_p$, as elements of $\mathbb{Z} \subseteq \mathbb{Z}_\ell$. The representation $\rho_{E,\ell}$ thus determines the coefficients a_p of the L-series $L_E(s)$ at all primes $p \notin S$. By the Tate-Faltings Theorem (see Theorem 24.38), this determines E up to isogeny, and therefore determines the entire L-function $L_E(s)$, including the values of a_p for $p \in S$.

We also have the *mod-ℓ Galois representation*

$$\bar{\rho}_{E,\ell} \colon G_{\mathbb{Q}} \to \mathrm{Aut}(E[\ell]) \simeq \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z}),$$

which is equivalent to composing $\rho_{E,\ell}$ with the map from $\mathrm{GL}_2(\mathbb{Z}_\ell)$ to $\mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$ that reduces each matrix coefficient modulo ℓ.

25.4 Serre's modularity conjecture

Let us forget about elliptic curves for a moment and consider an arbitrary[4] ℓ-adic Galois representation $\rho \colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_\ell)$ with $\ell > 3$ prime. We say that ρ is *modular* (of weight k and level N), if there is a modular form $f_\rho = \sum a_n q^n$ in $S_k(\Gamma_1(N))$ with $a_n \in \mathbb{Z}$ such that[5]

$$\operatorname{tr}\rho(\mathrm{Frob}_p) = a_p$$

for all primes $p \nmid \ell N$ (if $\rho = \rho_{E,\ell}$ and $N = N_E$ this excludes the same finite set of primes S as the previous section). Similarly, if we have a mod-ℓ representation $\bar{\rho} \colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$, we say that $\bar{\rho}$ is modular if

$$\operatorname{tr}\bar{\rho}(\mathrm{Frob}_p) \equiv a_p \bmod \ell$$

for all primes $p \nmid \ell N$.

Let $c \in G_{\mathbb{Q}}$ be the automorphism of $\overline{\mathbb{Q}} \subseteq \mathbb{C}$ corresponding to complex conjugation. The automorphism c has order 2, so $\det\rho(c) = \pm 1$. We say that a Galois representation ρ is *odd* when $\det\rho(c) = -1$. This is necessarily the case if $\rho = \rho_{E,\ell}$ is a Galois representation associated to an elliptic curve. One way to see this is to base change E to $\mathbb{C}$ and view $E_{\mathbb{C}}$ as isomorphic to a torus $\mathbb{C}/L$ for some lattice $L = [1, \tau]$. For a suitable choice of basis (P, Q) for the ℓ^n-torsion subgroup of $\mathbb{C}/L$ in which P has real coordinates, complex conjugation fixes P and sends Q to $-Q$ (this is easy to see when $\operatorname{re}\tau = 0$ and holds in general). Since we already know that every $f = \sum a_n q^n$ in $S_2^{\mathrm{new}}(\Gamma_0(N))$ with $a_n \in \mathbb{Z}$ gives rise to an elliptic curve (see Theorem 24.37), this constraint necessarily applies to Galois representations associated to modular forms of weight 2 with integral q-series.

We want to impose a further constraint on the Galois representations we shall consider that is not always satisfied by the representation $\bar{\rho}_{E,\ell}$ associated to an elliptic curve $E/\mathbb{Q}$, but usually is (always for $\ell > 163$). We call a Galois representation $\rho \colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$ *irreducible* if its image does not fix any one-dimensional subspace of $(\mathbb{Z}/\ell\mathbb{Z})^2$; equivalently,

[4] As profinite groups, both $G_{\mathbb{Q}} = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and $\mathrm{GL}_2(\mathbb{Z}_\ell)$ are topological groups and we always require ℓ-adic Galois representations to be continuous with respect to this topology; this is automatically true for the representation $\rho_{E,\ell}$ of interest to us.

[5] In the previous lecture we focused on $S_k(\Gamma_0(N))$, which suffices for everything we need in the sections that follow (and we only need $k = 2$), but in order to state Serre's conjecture we temporarily work in greater generality; note that $\Gamma_1(N) \subseteq \Gamma_0(N)$ implies $S_k(\Gamma_0(N)) \subseteq S_k(\Gamma_1(N))$.

its image is not conjugate to a group of upper triangular matrices in $\mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$. For an elliptic curve $E/\mathbb{Q}$, the mod-ℓ Galois representation $\bar{\rho}_{E,\ell}$ is irreducible if and only if E does not admit a rational ℓ-isogeny. Mazur's isogeny theorem [12] implies that this necessarily holds for $\ell \notin \{2, 3, 5, 7, 11, 13, 17, 19, 37, 43, 67, 163\}$ (the cases $19, 43, 67, 163$ can arise only when E has complex multiplication).

In 1975 Serre made the following remarkable conjecture, which he refined in [16]. This conjecture is now a theorem, proved in 2008 by Khare and Wintenberger [8, 9], but this work came long after the proof of Fermat's Last Theorem (and built on the modularity lifting techniques used to prove it).

Conjecture 25.1 (Serre's modularity conjecture)**.** *Every odd irreducible Galois representation $\bar{\rho}\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$ is modular.*[6]

Serre gave an explicit recipe for what the optimal weight $k(\bar{\rho})$ and level $N(\bar{\rho})$ of the corresponding modular form should be. Given a newform $f \in S_2^{\mathrm{new}}(\Gamma_0(N))$ with Fourier coefficients $a_n \in \mathbb{Z}$, the Eichler-Shimura Theorem (see Theorem 24.37) gives us a corresponding elliptic curve $E/\mathbb{Q}$ whose mod-ℓ Galois representation $\rho_{E,\ell}$ is modular of weight 2 and level $N = N_E$, and $\bar{\rho}_{E,\ell}$ will typically also be irreducible. The weight 2 agrees with the optimal weight $k(\bar{\rho}_{E,\ell})$ conjectured by Serre (at least when $\ell \nmid N_E$), but the optimal level $N(\bar{\rho}_{E,\ell})$ may properly divide N_E. In certain (rare) circumstances, distinct newforms of weight 2 with different levels may have Fourier coefficients a_n that are congruent modulo ℓ.

The mod-ℓ Galois representation associated to the "strange" elliptic curve $E_{a,b,c}$ arising from a Fermat solution $a^\ell + b^\ell = c^\ell$ gives rise to one of these rare circumstances. For an irreducible mod-ℓ Galois representations $\bar{\rho}_{E,\ell}$ arising from a semistable elliptic curve $E/\mathbb{Q}$, Serre's optimal level $N(\bar{\rho}_{E,\ell})$ is a product of primes p for which $v_p(\Delta_{\min}(E)) \not\equiv 0 \bmod \ell$, where $v_p(\cdot)$ denotes the p-adic valuation.

For the elliptic curve $E_{a,b,c}$ we have

$$N_{E_{a,b,c}} = \prod_{p|abc} p, \qquad\qquad \Delta_{\min}(E_{a,b,c}) = 2^{-8}(abc)^{2\ell},$$

which means that for every odd prime $p|N_{E_{a,b,c}}$ we have $v_p(\Delta_{\min}(E_{a,b,c})) \equiv 0 \bmod \ell$, in which case Serre's optimal level is $N(\bar{\rho}_{E_{a,b,c},\ell}) = 2$. But there are no (nonzero) modular forms of weight 2 and level 2, because $\dim S_2(\Gamma_1(2)) = \dim S_2(\Gamma_0(2)) = g(X_0(2)) = 0$. We must have $\ell > 163$, since Fermat's Last Theorem has long been known for $\ell \le 163$, so $E_{a,b,c}$ cannot admit a rational ℓ-isogeny, by Mazur's isogeny theorem, which means that $\bar{\rho}_{E_{a,b,c},\ell}$ must be irreducible. Thus if $E_{a,b,c}$ is modular, then $\bar{\rho}_{E_{a,b,c},\ell}$ represents a counterexample to Serre's conjecture. Serre's *epsilon-conjecture*, proved by Ribet in 1986, implies that this cannot happen. Below is a form of Ribet's theorem [14] that suffices to prove this.

Theorem 25.2 (Ribet)**.** *Let ℓ be prime, let E be an elliptic curve of conductor $N = mN'$, where m is the product of all primes $p|N$ such that $v_p(N) = 1$ and $v_p(\Delta_{\min}(E)) \equiv 0 \bmod \ell$. If E is modular and $\bar{\rho}_{E,\ell}$ is irreducible, then $\bar{\rho}_{E,\ell}$ is modular of weight 2 and level N'.*

Corollary 25.3. *The elliptic curve $E_{a,b,c}$ is not modular.*

[6] In fact Serre made his conjecture for all odd irreducible representations $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_{\ell^n})$, which includes the special case considered here with $\mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z}) \simeq \mathrm{GL}_2(\mathbb{F}_\ell)$.

25.5 The modularity lifting theorem

The final and by far the most difficult step to proving Fermat's Last Theorem is to show that if the elliptic curve $E_{a,b,c}$ exists, then it *is* modular. Andrew Wiles, with the assistance of Richard Taylor,[7] proved the stronger statement that every semistable elliptic curve over $\mathbb{Q}$ is modular (recall that $E_{a,b,c}$ is semistable).

A key element of Wiles' proof is a technique now known as *modularity lifting*. Let E be an elliptic curve over $\mathbb{Q}$ and let ℓ be a prime. Wiles uses modularity lifting to show that if the mod-ℓ Galois representation $\bar{\rho}_{E,\ell}$ of semistable elliptic curve $E/\mathbb{Q}$ is modular, then the ℓ-adic representation $\rho_{E,\ell}$ is also modular, which in turn implies that E is modular.

Given a representation $\rho_0\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$, a representation $\rho_1\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_\ell)$ whose reduction modulo ℓ is equal to ρ_0 is called a *lift* of ρ_0. More generally, if R is a suitable ring[8] with a reduction map to $\mathbb{Z}/\ell\mathbb{Z}$, and $\rho_1\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(R)$ is a representation whose reduction is equal to ρ_0, then we say that ρ_1 is a lift of ρ_0 (to R). Two lifts of ρ_0 are said to be *equivalent* if they are conjugate via an element in the kernel of the reduction map from $\mathrm{GL}_2(R)$ to $\mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$. A *deformation* of ρ_0 is an equivalence class of lifts of ρ_0 to the ring R, which is sometimes called the *deformation ring*.

Building on work by Mazur, Hida, and others that established the existence of certain *universal deformations* $\rho_{\mathbb{T}}\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{T})$, where $\mathbb{T}$ is a certain Hecke algebra, Taylor and Wiles were able to show that if ρ_0 is modular, then *every* lift of ρ_0 satisfying a specified list of properties is modular (this result and generalizations of it are now known as "$R = \mathbb{T}$" theorems), and Wiles was able to show that this list of properties is satisfied by the representation $\rho_{E,\ell}$ associated to a semistable elliptic curve $E/\mathbb{Q}$.

We are intentionally glossing over a massive amount of detail that is beyond the scope of this course. We refer the interested reader to [3], which contains not only a detailed overview of the proof, but many chapters devoted to the background necessary to understand these details, and also the lecture notes from 2009-2010 Modularity lifting seminar held at Stanford [2] which cover refinements of the Taylor-Wiles method and subsequent results.

Theorem 25.4 (Taylor-Wiles)**.** *Let $E/\mathbb{Q}$ be a semistable elliptic curve. If $\bar{\rho}_{E,\ell}$ is modular, then $\rho_{E,\ell}$ is also modular (and therefore E is modular).*

25.6 Proof of Fermat's Last Theorem

It remains only to find a modular representation $\rho_0\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}/\ell\mathbb{Z})$ that we can lift to $\rho_{E,\ell}$. The obvious candidate is $\bar{\rho}_{E,\ell}$, for some suitable choice of ℓ. It is not clear that proving the modularity of $\bar{\rho}_{E,\ell}$ is necessarily any easier than proving the modularity of $\rho_{E,\ell}$, but thanks to work of Langlands [10] and Tunnell [19] on a special case of Langlands' Reciprocity Conjecture [3, Ch. 6], we have the following result for $\ell = 3$.

Theorem 25.5 (Langlands-Tunnell)**.** *Let E be an elliptic curve over $\mathbb{Q}$. If $\bar{\rho}_{E,3}$ is irreducible, then it is modular.*

The one remaining difficulty is that $\bar{\rho}_{E,3}$ need not be irreducible; indeed there are infinitely many semistable elliptic curves $E/\mathbb{Q}$ that admit a rational 3-isogeny, and for these curves $\bar{\rho}_{E,3}$ is not irreducible. However, if E is semistable and $\bar{\rho}_{E,3}$ is reducible then $\bar{\rho}_{E,5}$

[7] Wiles' initial proof (announced in summer 1993) contained a significant gap. Richard Taylor helped Wiles circumvent this gap, which was the last critical step required to obtain a complete proof; see [4] for an accessible account.

[8] A complete local Noetherian ring with residue field $\mathbb{F}_\ell$.

must be irreducible. This follows from the fact that if neither $\bar{\rho}_{E,3}$ nor $\bar{\rho}_{E,5}$ is irreducible then E admits both a rational 3-isogeny and a rational 5-isogeny; the cyclic group of order 15 generated by their kernels is then the kernel of a rational 15-isogeny, but this cannot be the case if E is semistable.

Theorem 25.6. *No semistable elliptic curve $E/\mathbb{Q}$ admits a rational 15-isogeny.*

Proof. Let $E/\mathbb{Q}$ be an elliptic curve that admits a rational 15-isogeny. Let $\langle P\rangle \subseteq E(\overline{\mathbb{Q}})$ be the kernel of this isogeny, which we note is necessarily cyclic. The pair $(E, \langle P\rangle)$ corresponds to a non-cuspidal $\mathbb{Q}$-rational point on $X_0(15)$, the modular curve that parameterizes $\overline{\mathbb{Q}}$-isomorphism classes of 15-isogenies. The modular curve $X_0(15)$ is a smooth projective curve of genus 1, and it has a rational point (take the cusp at infinity, for example), so it can be viewed as an elliptic curve. A minimal Weierstrass model for $X_0(15)$ is given by

$$X_0(15)\colon y^2 + xy + y = x^3 + x^2 - 10x - 10.$$

Additional information about this curve can be found on its home page in the LMFDB [11]. This information includes the fact that $X_0(15)$ has rank 0 and a torsion subgroup of order 8. Its 8 rational points include 4 cusps and 4 non-cuspidal points that represent $\overline{\mathbb{Q}}$-isomorphism classes $(E, \langle P\rangle)$ of elliptic curves $E/\mathbb{Q}$ that admit a rational 15-isogeny with kernel $\langle P\rangle$. None of these elliptic curves E has j-invariant 0 or 1728, so each isomorphism class is a family of quadratic twists. Any family of quadratic twists of elliptic curves over $\mathbb{Q}$ contains a minimal representative whose conductor divides the conductor of all others; for the 4 non-cuspidal points on $X_0(15)$ these minimal quadratic twists all have conductor $50 = 2\cdot 5^2$ (you can find a list of them and the 15-isogenies they admit here). None of these curves is semistable, since 50 is not squarefree, nor are any of their quadratic twists. The theorem follows. □

There is unfortunately no analog of the Langlands-Tunnell theorem for $\ell = 5$. Indeed, the case $\ell = 3$ is quite special: the group $\mathrm{GL}_2(\mathbb{Z}/3\mathbb{Z})$ is solvable, which is not true for any prime $\ell > 3$ (and $\ell = 2$ has other problems). So we would seem to be stuck. But Wiles cleverly proved the following result, which is now known as the *three-five trick*.

Theorem 25.7 (Wiles)**.** *Let $E/\mathbb{Q}$ be a semistable elliptic curve for which $\bar{\rho}_{E,5}$ is irreducible. There exists a semistable elliptic curve $E'/\mathbb{Q}$ such that*

- *$\bar{\rho}_{E',3}$ is irreducible,*
- *$\bar{\rho}_{E',5} \simeq \bar{\rho}_{E,5}$.*

Now we are in business.

Theorem 25.8 (Wiles)**.** *Let $E/\mathbb{Q}$ be a semistable elliptic curve. Then E is modular.*

Proof. There are two cases. If $\bar{\rho}_{E,3}$ is irreducible then:

- $\bar{\rho}_{E,3}$ is modular, by the Langlands-Tunnell theorem,
- $\rho_{E,3}$ is modular, by the modularity lifting theorem,
- E is modular, since $f_E = f_{\rho_{E,3}}$.

On the other hand, if $\bar{\rho}_{E,3}$ is reducible, then:

- $\bar{\rho}_{E,5}$ is irreducible, because no semistable $E/\mathbb{Q}$ admits a rational 15-isogeny,

- there exists a semistable $E'/\mathbb{Q}$ with $\bar{\rho}_{E',3}$ irreducible and $\bar{\rho}_{E',5} \simeq \bar{\rho}_{E,5}$, by the 3-5 trick,
- $\bar{\rho}_{E',3}$ is modular, by the Langlands-Tunnell theorem,
- $\rho_{E',3}$ is modular, by the modularity lifting theorem,
- E' is modular, since $f_{E'} = f_{\rho_{E',3}}$,
- $\rho_{E',5}$ is modular, since $f_{\rho_{E',5}} = f_{E'}$, and therefore $\bar{\rho}_{E',5}$ is modular,
- $\bar{\rho}_{E,5} \simeq \bar{\rho}_{E',5}$ is modular,
- $\rho_{E,5}$ is modular, by the modularity lifting theorem,
- E is modular, since $f_E = f_{\rho_{E,5}}$.

Q.E.D. □

Corollary 25.9. *$x^n + y^n = z^n$ has no integer solutions with $xyz \neq 0$ for $n > 2$.*

References

[1] Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor, *On the modularity of elliptic curves over $\mathbb{Q}$: wild 3-adic exercises*, Journal of the American Mathematical Society **14** (2001), 843–939.

[2] B. Baran, R. Bellovin, B. Conrad, S. Dasgupta, B. Levin, S. Lichtenstein, M. Lipnowski, A. Paulin, N. Mok, A. Snowden, D. Trotabas, A. Venkatesh, and M. Weissman, *Modularity lifting seminar*, lecture notes from Stanford seminar, 2009–2010.

[3] Gary Cornell, Joseph H. Silverman, Glenn Stevens, *Modular forms and Fermat's Last Theorem*, Springer, 1998.

[4] Gerd Faltings, *The proof of Fermat's last theorem by R. Taylor and A. Wiles*, Notices of the American Mathematical Society **42** (1995), 743–746.

[5] Gerhard Frey, *Links between stable elliptic curves and certain diophantine equations*, Annales Universitatis Saraviensis. Series Mathematicae **1** (1986), 1–40.

[6] William Hart, David Harvey, and W. Ong, *Irregular primes to two billion*, Math. Comp. **86** (2017), 3031–3049.

[7] Yves Hellegouarch, *Courbes elliptiques et équation de Fermat*. Thése, Besançon, (1972).

[8] Chandrashekhar Khare and Jean-Pierre Wintenberger, *Serre's modularity conjecture (I)*, Inventiones Mathematicae **178** (2009), 485–504.

[9] Chandrashekhar Khare and Jean-Pierre Wintenberger, *Serre's modularity conjecture (II)*, Inventiones Mathematicae **178** (2009), 505–586.

[10] Robert P. Langlands, *Base change for* GL(2), Annals of Mathematical Studies **96**, Princeton University Press, 1980.

[11] The LMFDB collaboration, *The L-functions and Modular Forms Database*, published electronically at `https://www.lmfdb.org`, accessed May 19, 2021.

[12] Barry Mazur, *Rational isogenies of prime degree*, Invent. Math. **44** (1978), 129–162.

[13] J. S. Milne, *Elliptic curves*, BookSurge Publishers, 2006.

[14] Kenneth Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ *arising from modular forms*, Inventiones Mathematicae **100** (1990), 431–476.

[15] Kenneth Ribet, *Galois representations and modular forms*, Bulletin of the AMS **32** (1995), 375–402.

[16] Jean-Pierre Serre, *Sur les représentations modulaires de degré* 2 *de* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, Duke Mathematics Journal **54** (1987), 179–230.

[17] Joseph H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Springer, 1994.

[18] Lucien Szpiro, *Discriminant et conducteur des courbes elliptiques*, in *Séminaire sur les Pinceaux de Courbe Elliptiques (Paris, 1988)*, Astérisque **183** (1990), 7–18.

[19] Jerrold Tunnell, *Artin's conjecture for representations of octahedral type*, Bulletin of the American Mathematical Society **5** (1981), 173–175.

[20] André Weil, *Über die Bestimmung Dirichletscher Reihen durch Funktionalgleichungen*, Mathematische Annalen **168** (1967), 149–156.

[21] Richard Taylor and Andrew Wiles, *Ring-theoretic properties of certain Hecke algebras*, Annals of Mathematics **141** (1995), 553–572.

[22] Lawrence Washington, *Introduction to cyclotomic fields*, Springer, 1997.

[23] Andrew Wiles, *Modular elliptic curves and Fermat's last theorem*, Annals of Mathematics **141** (1995), 443–551.

18.783 / 18.7831 Elliptic Curves
Fall 2025