ELLIPTIC CURVES

J.S. MILNE

August 21, 1996; v1.01

ABSTRACT. These are the notes for Math 679, University of Michigan, Winter 1996, exactly as they were handed out during the course except for some minor corrections.

Please send comments and corrections to me at jmilne@umich.edu using "Math679" as the subject.

CONTENTS

Introduction

An elliptic curve over a field k is a nonsingular complete curve of genus 1 with a distinguished point. If char$k \neq 2,3$, it can be realized as a plane projective curve

$$Y^2Z = X^3 + aXZ^2 + bZ^3, \qquad 4a^3 + 27b^2 \neq 0,$$

and every such equation defines an elliptic curve over k. As we shall see, the arithmetic theory of elliptic curves over $\mathbb{Q}$ (and other algebraic number fields) is a rich a beautiful subject. Many important phenomena first become visible in the study elliptic curves, and elliptic curves have been used solve some very famous problems that, at first sight, appear to have nothing to do with elliptic curves. I mention three such problems.

Fast factorization of integers. There is an algorithm for factoring integers that uses elliptic curves and is in many respects better than previous algorithms. See [K2, VI.4], [ST,IV.4], or [C2, Chapter26]. People have been factoring integers for centuries, but recently the topic has become of practical significance: given an integer n which is the product $n = pq$ of two (large) primes p and q, there is a code for which anyone who knows n can encode a message, but only those who know p, q can decode it. The security of the code depends on no unauthorized person being able to factor n.

Congruent numbers. A natural number n is said to be *congruent* if it occurs as the area of a right triangle whose sides have rational length. If we denote the lengths of the sides by x, y, z, then n will be congruent if and only if the equations

$$x^2 + y^2 = z^2, \qquad n = \frac{1}{2}xy$$

have simultaneous solutions in $\mathbb{Q}$. The problem was of interest to the Greeks, and was discussed systematically by Arab scholars in the tenth century. Fibonacci showed that 5 and 6 are congruent, Fermat that $1, 2, 3,$ are not congruent, and Euler proved that 7 is congruent, but the problem appeared hopeless until in 1983 Tunnell related it to elliptic curves.

Fermat's last theorem. Recently Wiles proved that all elliptic curves over $\mathbb{Q}$ (with a mild restriction) arise in a certain fashion from modular forms. It follows from his theorem, that for an odd prime $p \neq 3$, there does not exist an elliptic curve over $\mathbb{Q}$ whose equation has the form

$$Y^2 = X(X + a)(X - b)$$

with $a, b, a + b$ all p^{th} powers of integers, i.e., there does not exist a nontrivial solution in $\mathbb{Z}$ to the equation

$$X^p + Y^p = Z^p;$$

—Fermat's Last Theorem is proved!

The course will be an introductory survey of the subject—often proofs will only be sketched, but I will try to give precise references for everything.

There are many excellent books on subject—see the Bibliography. Silverman [S1,S2] is becoming the standard reference.

1. Review of Plane Curves

Affine plane curves. Let k be a field. The *affine plane over* k is $\mathbb{A}^2(k) = k^2$.

A nonconstant polynomial $f \in k[X,Y]$, assumed to have no repeated factor in $k^{\mathrm{al}}[X,Y]$, defines a plane affine curve C_f over k whose points with coordinates in any field $K \supset k$ are the zeros of f in K^2:

$$C_f(K) = \{(x,y) \in K^2 \mid F(x,y) = 0\}.$$

The curve C is said to be *irreducible* if f is irreducible, and it is said be *geometrically irreducible* if f remains irreducible over k^{al} (equivalently, over any algebraically closed field containing k).

Since $k[X,Y]$ is a unique factorization domain, we can write any f as above as a product $f = f_1 f_2 \cdots f_r$ of distinct irreducible polynomials, and then

$$C_f = C_{f_1} \cup \cdots \cup C_{f_r}$$

with the C_{f_i} irreducible curves. The C_{f_i} are called the irreducible components of C_f.

Example 1.1. (a) Let $f_1(X,Y)$ be an irreducible polynomial in $\mathbb{Q}[\sqrt{2}][X,Y]$, no constant multiple of which lies $\mathbb{Q}[X,Y]$, and let $\bar{f}_1(X,Y)$ be its conjugate over $\mathbb{Q}$ (i.e., replace each $\sqrt{2}$ with $-\sqrt{2}$). Then $f(X,Y) =_{df} f_1(X,Y)\bar{f}_1(X,Y)$ lies in $\mathbb{Q}[X,Y]$ because it is fixed by the Galois group of $\mathbb{Q}[\sqrt{2}]/\mathbb{Q}$. The curve C_f is irreducible but not geometrically irreducible.

(b) Let k be a field of characteristic p. Assume k is not perfect, so that there exists an $a \in k$, $a \notin k^p$. Consider

$$f(X,Y) = X^p + aY^p.$$

Then f is irreducible in $k[X,Y]$, but in $k^{\mathrm{al}}[X,Y]$ it equals $(X+\alpha Y)^p$ where $\alpha^p = a$ (remember, the binomial theorem takes on a specially simple form for p^{th} powers in characteristic p). Thus f does not define a curve.

We define the partial derivatives of a polynomial by the obvious formulas.

Let $P = (a,b) \in C_f(K)$, some $K \supset k$. If at least one of the partial derivatives $\frac{\partial f}{\partial X}$, $\frac{\partial f}{\partial Y}$ is nonzero at P, then P is said to be *nonsingular*, and the *tangent line* to C at P is

$$\left(\frac{\partial f}{\partial X}\right)_P (X-a) + \left(\frac{\partial f}{\partial Y}\right)_P (Y-b) = 0.$$

A curve C is said to be *nonsingular* if all the points in $C(k^{\mathrm{al}})$ are nonsingular. A curve or point that is not nonsingular said to be *singular.*

Aside 1.2. Let $f(x,y)$ be a real-valued function on $\mathbb{R}^2$. In Math 215 one learns that $\nabla f =_{df} \left(\frac{\partial f}{\partial X}, \frac{\partial f}{\partial Y}\right)$ is a vector field on $\mathbb{R}^2$ that, at any point $P = (a,b) \in \mathbb{R}^2$, points in the direction in which $f(x,y)$ increases most rapidly (i.e., has the most positive directional derivative). Hence $(\nabla f)_P$ is normal to any level curve $f(x,y) = c$ through P, and the line

$$(\nabla f)_P \cdot (X-a, Y-b) = 0$$

passes through P and is normal to the normal to the level curve. It is therefore the tangent line.

Example 1.3. Consider the curve

$$C: \quad Y^2 = X^3 + aX + b.$$

At a singular point of C

$$2Y = 0, \quad 3X^2 + a = 0, \quad Y^2 = X^3 + aX + b.$$

Assume char $k \neq 2$. Hence $Y = 0$ and X is a common root of $X^3 + aX + b$ and its derivative, i.e., a double root of $X^3 + aX + b$. Thus C is nonsingular $\iff$ $X^3 + aX + b$ has no multiple root (in k^{al}) $\iff$ its discriminant $4a^3 + 27b^2$ is nonzero.

Assume char $k = 2$. Then C always has a singular point (possibly in some extension field of k), namely, (α, β) where $\alpha^2 + a = 0$ and $\beta^2 = \alpha^3 + a\alpha + b$.

Let $P = (a, b) \in C_f(K)$. We can write f as a polynomial in $X - a$ and $Y - b$ with coefficients in K, say,

$$f(X, Y) = f_1(X - a, Y - b) + \cdots + f_n(X - a, Y - b)$$

where f_i is homogeneous of degree i in $X - a$ and $Y - b$ (this the Taylor expansion of f!). The point P is nonsingular if and only if $f_1 \neq 0$, in which case the tangent line to C_f at P has equation $f_1 = 0$.

Suppose that P is singular, so that

$$f(X, Y) = f_m(X - a, Y - b) + \text{terms of higher degree},$$

where $f_m \neq 0$, $m \geq 2$. Then P is said to have *multiplicity* m on C, denoted $m_P(C)$. If $m = 2$, then P is called a *double point.* For simplicity, take $(a, b) = (0, 0)$. Then (over k^{al})

$$f_m(X, Y) = \prod L_i^{r_i}$$

where each L_i is a homogeneous polynomial $c_iX + d_iY$ of degree one with coefficients in k^{al}. The lines $L_i = 0$ are called the *tangent lines* to C_f at P, and r_i is called *multiplicity* of L_i. The point P is said to be an *ordinary singularity* if the tangent lines are all distinct, i.e., $r_i = 1$ for all i. An ordinary double point is called a *node.*

Example 1.4. The curve $Y^2 = X^3 + aX^2$ has a singularity at $(0, 0)$. If $a \neq 0$, it is a node, and the tangent lines at $(0, 0)$ are $Y = \pm\sqrt{a}X$. They are defined over k if and only if a is a square in k.

If $a = 0$, the singularity is a cusp. (A double point P on a curve C is called a *cusp* if there is only one tangent line L to C at P, and, with the notation defined below, $I(P, L \cap C) = 3$.)

Consider two curves C_f and C_g in $\mathbb{A}^2(k)$, and let $P \in C_f(K) \cap C_g(K)$, some $K \supset k$. Assume that P is an *isolated* point of $C_f \cap C_g$, i.e., C_f and C_g do not have a common irreducible component passing through P. We define the *intersection number* of C_f and C_g at P to be

$$I(P, C_f \cap C_g) = \dim_K K[X, Y]_{(X-a, Y-b)}/(f, g)$$

(dimension as K-vector spaces).

Remark 1.5. If C_f and C_g have no common component, then

$$\sum_{P \in C(k^{\text{al}}) \cap C(k^{\text{al}})} I(P, C_f \cap C_g) = \dim_{k^{\text{al}}} k[X, Y]/(f, g).$$

This is particularly useful when C_f and C_g intersect at a single point.

Example 1.6. Let C be the curve $Y^2 = X^3$, and let $L: \quad Y = 0$ be its tangent line at $P = (0,0)$. Then

$$I(P, L \cap C) = \dim_k k[X,Y]/(Y, Y^2 - X^3) = \dim_k k[X]/(X^3) = 3.$$

Remark 1.7. (a) The intersection number doesn't depend on which field K the coordinates of P are considered to lie in.

(b) As expected, $I(P, C \cap D) = 1$ if and only if P is nonsingular on both C and D, and the tangent lines to C and D at P are distinct. More generally, $I(P, C \cap D) \geq m_P(C) \cdot m_P(D)$, with equality if and only if C and D have no tangent line in common at P.

Projective plane curves. The *projective plane* over k is

$$\mathbb{P}^2(k) = \{(x,y,z) \in k^3 \mid (x,y,z) \neq (0,0,0)\}/\sim$$

where $(x,y,z) \sim (x',y',z')$ if and only if there exists a $c \neq 0$ such that $(x',y',z') = (cx,cy,cz)$. We write $(x:y:z)$ for the equivalence class[1] of (x,y,z). Let $P \in \mathbb{P}^2(k)$; the triples (x,y,z) representing P lie on a single line $L(P)$ through the origin in k^3, and $P \mapsto L(P)$ is a bijection from $\mathbb{P}^2(k)$ to the set of all such lines.

Projective n-space $\mathbb{P}^n(k)$ can be defined similarly for any $n \geq 0$.

Let $U_0 = \{(x:y:z\} \mid z \neq 0\}$, and let $L_\infty(k) = \{(x:y:z) \mid z = 0\}$. Then

$$(x,y) \mapsto (x:y:1) : \mathbb{A}^2(k) \to U_0$$

is a bijection, and

$$(x:y) \mapsto (x:y:0) : \mathbb{P}^1(k) \to L_\infty(k)$$

is a bijection. Moreover, $\mathbb{P}^2(k)$ is the disjoint union

$$\mathbb{P}^2(k) = U_0 \sqcup L_\infty(k)$$

of the "affine plane" U_0 with the "line at infinity" L_∞. A line

$$aX + bY + cZ = 0$$

meets L_∞ at the point $(-b:a:0) = (1:-\frac{a}{b},0)$. Thus we can think of $\mathbb{P}^2(k)$ as being the affine plane with exactly one point added for each family of parallel lines.

A nonconstant homogeneous polynomial $F \in k[X,Y,Z]$, assumed to have no repeated factor in k^{al}, defines a *projective plane curve* C_F over k whose points in any field $K \supset k$ are the zeros of F in $\mathbb{P}^2(K)$:

$$C_F(K) = \{(x:y:z) \mid F(x,y,z) = 0\}.$$

Note that, because F is homogeneous,

$$F(cx,cy,cz) = c^{\deg F} F(x,y,z),$$

and so, although it doesn't make sense to speak of the value of F at a point of $\mathbb{P}^2$, it does make sense to say whether or not F is zero at P. Again, the degree of F is called the *degree* of the curve C, and a plane projective curve is (uniquely) a union of irreducible plane projective curves.

The curve

$$Y^2Z = X^3 + aXZ^2 + bZ^3$$

[1] The colon is meant to suggest that only the ratios matter.

intersects the line at infinity at the point $(0:1:0)$, i.e., at the same point as all the vertical lines do. This is plausible geometrically, because, as you go out the affine curve

$$Y^2 = X^3 + aX + b$$

with increasing x and y, the slope of the tangent line tends to ∞.

Let $U_1 = \{(x:y:z\} \mid y \neq 0\}$, and let $U_2 = \{(x:y:z)|x \neq 0\}$. Then U_1 and U_2 are again, in a natural way, affine planes; for example, we can identify U_1 with $\mathbb{A}^2(k)$ via

$$(x:1:z) \leftrightarrow (x,z).$$

Since at least one of x, y, or z is nonzero,

$$\mathbb{P}^2(k) = U_0 \cup U_1 \cup U_2.$$

A plane projective curve $C = C_F$ is the union of three curves,

$$C = C_0 \cup C_1 \cup C_2, \quad C_i = C \cap U_i.$$

When we identify each U_i with $\mathbb{A}^2(k)$ in the natural way, then C_0, C_1, and C_2 become identified with the affine curves defined by the polynomials $F(X,Y,1)$, $F(X,1,Z)$, and $F(1,Y,Z)$ respectively.

The curve

$$C: \quad Y^2Z = X^3 + aXZ^2 + bZ^3$$

is unusual, in that it is covered by two (rather than 3) affine curves

$$C_0: \quad Y^2 = X^3 + aX + b$$

and

$$C_1: \quad Z = X^3 + aXZ^2 + bZ^3.$$

The notions of tangent line, multiplicity, etc. can be extended to projective curves by noting that each point P of a projective curve C will lie on at least one of the affine curves C_i.

Exercise 1.8. Let P be a point on a plane projective curve $C = C_F$. Show that P is singular, i.e., it is singular on the plane affine curve C_i for one (hence all) i if and only if $F(P) = 0 = \left(\frac{\partial F}{\partial X}\right)_P = \left(\frac{\partial F}{\partial Y}\right)_P = \left(\frac{\partial F}{\partial Z}\right)_P$. If P is nonsingular, show that the plane projective line

$$L: \quad \left(\frac{\partial F}{\partial X}\right)_P X + \left(\frac{\partial F}{\partial Y}\right)_P Y \quad \left(\frac{\partial F}{\partial Z}\right)_P Z = 0$$

has the property that $L \cap U_i$ is the tangent line to the affine curve $C \cap U_i$ for $i = 0,1,2$.

Theorem 1.9 (Bezout). *Let C and D be plane projective of degrees m and n respectively over k, and assume that they have no irreducible component in common. Then they intersect over k^{al} in exactly mn points, counting multiplicities, i.e.,*

$$\sum_{P \in C(k^{al}) \cap D(k^{al})} I(P, C \cap D) = mn.$$

Proof. See [F] p112, or many other books. □

For example, a curve of degree m will meet the line at infinity in exactly m points, counting multiplicities. Our favourite curve

$$C: \quad Y^2Z = X^3 + aXZ^2 + bZ^3$$

meets L_∞ at a single point $P = (0:1:0)$, but $I(P, L_\infty \cap C) = 3$. [Exercise: Prove this!] In general, a nonsingular point P on a curve C is called a *point of inflection* if the intersection multiplicity of the tangent line and C at P is ≥ 3.

Suppose k is perfect. Then all the points of $C(k^{\mathrm{al}}) \cap D(k^{\mathrm{al}})$ will have coordinates in some finite Galois extension K of k, and the set

$$C(K) \cap D(K) \subset \mathbb{P}^2(K)$$

is stable under the action of $\mathrm{Gal}(K/k)$.

Remark 1.10. (For the experts.) Essentially, we have defined an affine (resp. projective) curve to be a geometrically reduced closed subscheme of $\mathbb{A}^2_k$ (resp. $\mathbb{P}^2_k$) of dimension 1. Such a scheme corresponds to an ideal of height one, which is principal, because polynomial rings are unique factorization domains. The polynomial generating the ideal of the scheme is uniquely determined by the scheme up to multiplication by a nonzero constant. The other definitions in this section are standard.

References: The best reference for what little we need from algebraic geometry is [F].

2. Rational Points on Plane Curves.

Let C be a plane projective curve over $\mathbb{Q}$ (or some other field with an interesting arithmetic), defined by a homogeneous polynomial $F(X, Y, Z)$. The two fundamental questions in diophantine geometry then are:

Question 2.1. (a) Does C have a point in $\mathbb{Q}$, that is, does $F(X, Y, Z)$ have a nontrivial zero in $\mathbb{Q}$?

(b) If the answer to (a) is yes, can we describe the set of common zeros?

There is also the question of whether there is an algorithm to answer these questions. For example, we may know that a curve has only finitely many points without having an algorithm to actually find the points.

For simplicity, in the remainder of this section, I'll assume that C is absolutely irreducible, i.e., that $F(X, Y, Z)$ is irreducible over $\mathbb{Q}^{\mathrm{al}}$.

Here is one observation that we shall use frequently. Let K be a finite (of even infinite) Galois extension of $\mathbb{Q}$, and let

$$f(X, Y) = \sum a_{ij} X^i Y^j \in \mathbb{Q}[X, Y].$$

If $(a, b) \in K^2$ is a zero of $f(X, Y)$, then so also is $(\sigma a, \sigma b)$ for any $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$, because

$$0 = \sigma f(a, b) = \sigma(\sum a_{ij} a^i b^j) = \sum a_{ij} (\sigma a)^i (\sigma b)^j = f(\sigma a, \sigma b).$$

Thus $\mathrm{Gal}(K/\mathbb{Q})$ acts on $C(K)$, where C is the affine curve defined by $f(X, Y)$. More generally, if $C_1, C_2, \ldots$ are affine curves over $\mathbb{Q}$, then $\mathrm{Gal}(K/\mathbb{Q})$ stabilizes the set $C_1(K) \cap C_2(K) \cdots$. On applying this remark to the curves $f(X, Y) = 0$, $\frac{\partial f}{\partial X}(X, Y) = 0$, $\frac{\partial f}{\partial Y}(X, Y) = 0$, we see that $\mathrm{Gal}(K/\mathbb{Q})$ stabilizes the set of singular points in $C(K)$. Similar remarks apply to projective curves.

Curves of degree one. First consider a curve of degree one, i.e., a line,

$$C : aX + bY + cZ = 0, \quad a, b, c \text{ in } \mathbb{Q} \text{ and not all zero.}$$

It always has points, and it is possible to parameterize the points: if, for example, $c \neq 0$, the map

$$(s : t) \mapsto (s : t : -\frac{a}{c}s - \frac{b}{c}t)$$

is a bijection from $\mathbb{P}^1(k)$ onto $C(k)$.

Curves of degree two. In this case $F(X, Y, Z)$ is a quadratic form in 3 variables, and C is a conic. Note that C can't be singular: if P has multiplicity m, then (according to (1.7b)) a line L through P and a second point Q on the curve will have

$$I(P, L \cap C) + I(Q, L \cap C) \geq 2 + 1 = 3,$$

which violates Bezout's theorem.

Sometimes it is easy to see that $C(\mathbb{Q}) = \emptyset$. For example,

$$X^2 + Y^2 + Z^2$$

has no nontrivial zero because it has no nontrivial real zero. Similarly,

$$X^2 + Y^2 - 3Z^2$$

has no nontrivial zero, because if it did it would have a zero (x, y, z) with $x, y, z \in \mathbb{Z}$ and $gcd(x, y, z) = 1$. The only squares in $\mathbb{Z}/3\mathbb{Z}$ are 0 and 1, and so

$$x^2 + y^2 \equiv 0 \mod 3$$

implies that $x \equiv 0 \equiv y \mod 3$. But then 3 must divide z, which contradicts our assumption that $gcd(x, y, z) = 1$. This argument shows, in fact, that $X^2 + Y^2 - 3Z^2$ does not have a nontrivial zero in the field $\mathbb{Q}_3$ of 3-adic numbers.

These examples illustrate the usefulness of the following statement: a necessary condition for C to have a point with coordinates in $\mathbb{Q}$ is that it have a point with coordinates in $\mathbb{R}$ and in $\mathbb{Q}_p$ for all p. A theorem of Legendre says that the condition is also sufficient:

Theorem 2.2 (Legendre). *A quadratic form $F(X, Y, Z)$ with coefficients in $\mathbb{Q}$ has a nontrivial zero in $\mathbb{Q}$ if and only if it has a nontrivial zero in $\mathbb{R}$ and in $\mathbb{Q}_p$ for all p.*

Remark 2.3. (a) This is not quite how Legendre (1752–1833) stated it, since p-adic numbers are less than 100 years old

(b) The theorem does in fact give a practical algorithm for showing that a quadratic form does have a nontrivial rational zero—see (2.11) below.

(c) The theorem is true for quadratic forms $F(X_0, X_2, \dots, X_n)$ in any number of variables over any number field K (Hasse-Minkowski theorem). There is a very down-to-earth proof of the original case of the theorem in [C2]—it takes three lectures. A good exposition of the proof for forms over $\mathbb{Q}$ in any number of variables is to be found in Serre, Course on Arithmetic. The key cases are 3 and 4 variables (2 is trivial, and for ≥ 5 variables, one uses induction on n), and the key result needed for its proof is the quadratic reciprocity law. For number fields K other $\mathbb{Q}$, the proof requires the Hilbert reciprocity law, which is best derived as part of class field theory (see Math 776 for class field theory), but there is a more direct proof of Hilbert's reciprocity law in Chapter 7 of O'Meara, Introduction to Quadratic Forms (which proves the Hasse-Minkowski theorem in full generality).

(d) If for a class of polynomials (better algebraic varieties), it is known that each polynomial (or variety) has a zero in $\mathbb{Q}$ if and only if it has zeros in $\mathbb{R}$ and all $\mathbb{Q}_p$, then one says that the *Hasse*, or *local-global, principle* holds for the class.

Now suppose C has a point P_0 with coordinates in $\mathbb{Q}$. Can we describe all the points? Yes, because each line through P_0 will (by Bezout's theorem, or more elementary arguments) meet the curve in exactly one other point, except for the tangent line. Since the lines through P_0 in $\mathbb{P}^2$ form a "$\mathbb{P}^1$", we obtain in this way a bijection between $C(\mathbb{Q})$ and $\mathbb{P}^1(\mathbb{Q})$. For example, take P_0 to be the point $(-1:0:1)$ on the curve $C: X^2+Y^2=Z^2$. The line $bX-aY+bZ$, $a,b\in\mathbb{Q}$, of slope $\frac{b}{a}$ through P_0 meets C at the point $(a^2-b^2:2ab:a^2+b^2)$. In this way, we obtain a parametrization $(a:b)\mapsto(a^2-b^2:2ab:a^2+b^2)$ of the points of C with coordinates in $\mathbb{Q}$.

Curves of degree 3. Let $C: F(X,Y,Z)=0$ be a plane projective curve over $\mathbb{Q}$ of degree 3. If it has a singular point, then Bezout's theorem shows that it has only one, and that it is a double point. A priori the singular point P_0 may have coordinates in some finite[2] extension K of $\mathbb{Q}$, which we may take to be Galois over $\mathbb{Q}$, but $\mathrm{Gal}(K/\mathbb{Q})$ stabilizes the set of singular points in $C(K)$, hence fixes P_0, and so $P_0\in C(\mathbb{Q})$. Now a line through P_0 will meet the curve in exactly one other point (unless it is a tangent line), and so we again get a parametrization of the points (with finitely many exceptions).

Nonsingular cubics will be the topic of the rest of the course. We shall see that the Hasse principle fails for nonsingular cubic curves. For example,

$$3X^3+4Y^3=5Y^3$$

has points in $\mathbb{R}$ and $\mathbb{Q}_p$ for all p, but not in $\mathbb{Q}$. However, it is conjectured that the Hasse principle fails only by a "finite amount", and that the failure is "measured" by a certain group, called the *Tate-Shafarevich* group.

Let C be a nonsingular cubic curve over $\mathbb{Q}$. From two points $P_1,P_2\in C(\mathbb{Q})$ we can construct[3] a third as the point of intersection of $C(\mathbb{Q})$ with the chord through P_1 and P_2—by Bezout's theorem, there exists exactly one such point, perhaps with coordinates in a Galois extension K of $\mathbb{Q}$, but by the observation at the start of this section, it must be fixed by $\mathrm{Gal}(K/\mathbb{Q})$ and therefore lie in $C(\mathbb{Q})$. Similarly, the tangent line at a point $P\in C(\mathbb{Q})$ will meet C at exactly one other point (unless P is a point of inflection), which[4] will lie in $C(\mathbb{Q})$.

In a famous paper, published in 1922/23, Mordell proved the following theorem:

Theorem 2.4 (Finite basis theorem). *Let C be a nonsingular cubic curve over $\mathbb{Q}$. Then there exists a finite set of points on C with coordinates in $\mathbb{Q}$ from which every other such point can be obtained by successive chord and tangent constructions.*

[2] I should explain why the singular point has coordinates in a finite extension of $\mathbb{Q}$. I claim that if $F(X,Y,Z)$ has a singular point with coordinates in some big field Ω, e.g., $\mathbb{C}$, then it has a singular point with coordinates in $\mathbb{Q}^{\mathrm{al}}$ (hence in a finite extension of $\mathbb{Q}$). Pass to an affine piece of the curve, and consider a (nonhomogeneous) cubic $f(X,Y)$. If the curve $f(X,Y)=0$ has a singular point with coordinates in Ω, then $f,\frac{\partial f}{\partial X},\frac{\partial f}{\partial Y}$ have a common zero in Ω^2, and so the ideal they generate is not the whole of $\Omega[X,Y]$. This implies that the ideal they generate in $\mathbb{Q}^{\mathrm{al}}[X,Y]$ is not the whole ring, and the Hilbert Nullstellensatz then implies that they have a common zero in $(\mathbb{Q}^{\mathrm{al}})^2$.

[3] This construction goes back to Diophantus (3rd century A.D)

[4] This observation was first made by Newton (1642–1727).

In fact, $C(\mathbb{Q})$, if nonempty, can be made into an abelian group, and the finite basis theorem says that $C(\mathbb{Q})$ is finitely generated. There is as yet no proven algorithm for finding the rank of the group.

Curves of genus > 1. Mordell conjectured, in his 1922/23 paper, and Faltings proved, that a nonsingular plane projective curve of degree ≥ 4 has only finitely many points coordinates in $\mathbb{Q}$.

More generally, define the *geometric genus* of a plane projective curve C to be

$$p_g(C) = \frac{(d-1)(d-2)}{2} - \sum \delta_P$$

where d is the degree of C, the sum is over the singular points in $C(\mathbb{Q}^{\mathrm{al}})$, and $\delta_P = \frac{m_P(m_P-1)}{2}$ if P is an ordinary singularity of multiplicity m_P. Then $C(\mathbb{Q})$ is finite if C has geometric genus > 1.

Remark 2.5. (a) Let $P \in \mathbb{P}^2(\mathbb{Q})$. Choose a representative $(a : b : c)$ for P with a, b, c integers having no common factor, and define the height $h(P)$ of P to be $\max(|a|, |b|, |c|)$. The biggest remaining problem in the theory of curves of genus > 1 over $\mathbb{Q}$ is that of giving an upper bound $H(C)$, in terms of the polynomial defining C, for the heights of the points $P \in C(\mathbb{Q})$. With such an upper bound $H(C)$, one could find all the points on C with coordinates in $\mathbb{Q}$ by a finite search.

(b) There is a heuristic explanation for Mordell's conjecture. Let C be a curve of genus $g \geq 1$ over $\mathbb{Q}$, and assume that $C(\mathbb{Q}) \neq \emptyset$. It is possible to embed C into another projective variety J of dimension g (its *Jacobian variety*). The Jacobian variety J is an abelian variety, i.e., it has a group structure, and a generalization of Mordell's theorem (due to Weil) says that $J(\mathbb{Q})$ is finitely generated. Hence, inside the g-dimensional set $J(\mathbb{C})$ we have the countable set $J(\mathbb{Q})$ and the (apparently unrelated) one-dimensional set $C(\mathbb{C})$. If $g > 1$, it would be an extraordinary accident if the second set contained more than a finite number of elements from the first set.

Hensel's lemma.

Lemma 2.6. *Let $f(X_1, \ldots, X_n) \in \mathbb{Z}[X_1, \ldots, X_n]$, and let $\underline{a} \in \mathbb{Z}^n$ have the property that, for some $m \geq 0$,*

$$f(\underline{a}) \equiv 0 \mod p^{2m+1}$$

but, for some i,

$$\left(\frac{\partial f}{\partial X_i}\right)(\underline{a}) \not\equiv 0 \mod p^{m+1}.$$

Then there exists a $\underline{b} \in \mathbb{Z}^n$ such that

$$\underline{b} \equiv \underline{a} \mod p^{m+1} \quad (\Longrightarrow \left(\frac{\partial f}{\partial X_i}\right)(\underline{b}) \not\equiv 0 \mod p^{m+1})$$

and

$$f(\underline{b}) \equiv 0 \mod p^{2m+2}.$$

Proof. Consider the (trivial) Taylor expansion

$$f(X_1, \ldots, X_n) = f(a_1, \ldots, a_n) + \sum_{i=1}^{n} \left(\frac{\partial f}{\partial X_i}\right)_{\underline{a}} (X_i - a_i) + \text{terms of higher degree.}$$

Set $b_i = a_i + h_i p^{m+1}$, $h_i \in \mathbb{Z}$. Then

$$f(b_1, \dots, b_n) = f(a_1, \dots, a_n) + \sum \left(\frac{\partial f}{\partial X_i}\right)_{\underline{a}} h_i p^{m+1} + \text{terms divisible by } p^{2m+2}.$$

We have to choose the h_i so that

$$f(a_1, \dots, a_n) + \sum \left(\frac{\partial f}{\partial X_i}\right)_{\underline{a}} h_i p^{m+1}$$

is divisible by p^{2m+2}. From the assumption, we know that there is a $k \leq m$ such that p^k divides $\left(\frac{\partial f}{\partial X_i}\right)_{\underline{a}}$ for all i but p^{k+1} doesn't divide all of them. Any h_i's satisfying the following equation will suffice:

$$\frac{f(a_1, \dots, a_n)}{p^{k+m+1}} + \sum \frac{\left(\frac{\partial f}{\partial X_i}\right)_{\underline{a}}}{p^k} h_i \equiv 0 \mod p.$$

□

Remark 2.7. If, in the lemma, $\underline{a}$ satisfies the condition

$$f(\underline{a}) \equiv 0 \mod p^{2m+r}$$

for some $r \geq 1$, then the construction in the proof gives a $\underline{b}$ such that

$$\underline{b} \equiv \underline{a} \mod p^{m+r}$$

and

$$f(\underline{b}) \equiv 0 \mod p^{2m+r+1}.$$

Theorem 2.8 (Hensel's Lemma). *Under the hypotheses of the lemma, there exists a* $\underline{b} \in \mathbb{Z}_p^n$ *such that* $f(\underline{b}) = 0$ *and* $\underline{b} \equiv \underline{a} \mod p^{m+1}$.

Proof. On applying the lemma, we obtain an $\underline{a}_{2m+2} \in \mathbb{Z}^n$ such that $\underline{a}_{2m+2} \equiv \underline{a} \mod p^{m+1}$ and $f(\underline{a}_{2m+2}) \equiv 0 \mod p^{2m+2}$. On applying the remark following the lemma, we obtain an $\underline{a}_{2m+3} \in \mathbb{Z}^n$ such that $\underline{a}_{2m+3} \equiv \underline{a}_{2m+2} \mod p^{m+2}$ and $f(\underline{a}_{2m+3}) \equiv 0 \mod p^{2m+3}$. Continuing in this fashion, we obtain a sequence $\underline{a}, \underline{a}_{2m+2}, \underline{a}_{2m+3}, \dots$ of n-tuples of Cauchy sequences. Let $\underline{b}$ be the limit in $\mathbb{Z}_p^n$. The map $f : \mathbb{Z}^n \to \mathbb{Z}$ is continuous for the p-adic topologies, and so

$$f(\underline{b}) = f(\lim_r \underline{a}_{2m+r}) = \lim_r f(\underline{a}_{2m+r}) = 0.$$

□

Example 2.9. Let $f(X) \in \mathbb{Z}[X]$, and let $\overline{f}(X) \in \mathbb{F}_p[X]$ be its reduction mod p. Here $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. Let $a \in \mathbb{Z}$ be such that $\overline{a} \in \mathbb{F}_p$ is a simple root of $\overline{f}(X)$. Then $\frac{d\overline{f}}{dX}(\overline{a}) \neq 0$, and so the theorem shows that $\overline{a}$ lifts to a root of $f(X)$ in $\mathbb{Z}_p$.

Example 2.10. Let $f(X, Y, Z)$ be a homogeneous polynomial in $\mathbb{Z}[X, Y, Z]$, and let $(a, b, c) \in \mathbb{Z}^3$ be such that $(\overline{a}, \overline{b}, \overline{c}) \in \mathbb{F}_p^3$ is a nonsingular point of the curve $\overline{C} : \overline{f}(X, Y, Z) = 0$ over $\mathbb{F}_p$. Then, as in the previous example, $(\overline{a}, \overline{b}, \overline{c})$ lifts to a point on the curve $C : f(X, Y, Z) = 0$ with coordinates in $\mathbb{Z}_p$.

Example 2.11. Let $f(X,Y,Z)$ be a quadratic form with coefficients in $\mathbb{Z}$, and let $D \neq 0$ be its discriminant. If p does not divide D, then $\overline{f}(X,Y,Z)$ is a nondegenerate quadratic form over $\mathbb{F}_p$, and it is known that it has a nontrivial zero in $\mathbb{F}_p$. Therefore $f(X,Y,Z)$ has a nontrivial zero in $\mathbb{Q}_p$ for all such p. If p divides D, then Hensel's lemma shows that $f(X,Y,Z)$ will have a nontrivial zero in $\mathbb{Q}_p$ if and only if it has an "approximate" zero.

A brief introduction to the p-adic numbers. Let p be a prime number. Any nonzero rational number a can be expressed $a = p^r \frac{m}{n}$ with $m, n \in \mathbb{Z}$ and not divisible by p. We then write $\mathrm{ord}_p(a) = r$, and $|a|_p = \frac{1}{p^r}$. We define $|0|_p = 0$. Then:

(a) $|a|_p = 0$ if and only if $a = 0$.
(b) $|ab|_p = |a|_p|b|_p$.
(c) $|a+b|_p \leq \max\{|a|_p, |b|_p\} \quad (\leq |a|_p + |b_p|)$.

These conditions imply that

$$d_p(a,b) =_{df} |a-b|_p$$

is a translation-invariant metric on $\mathbb{Q}$. Note that, according to this definition, to say that a and b are close means that their difference is divisible by a high power of p. The field $\mathbb{Q}_p$ of p-adic numbers is the completion of $\mathbb{Q}$ for this metric. We now explain what this means.

A sequence (a_n) is said to be a *Cauchy sequence* (for the p-adic metric) if, for any $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ such that

$$|a_m - a_n|_p < \varepsilon \quad \text{whenever} \quad m, n > N(\varepsilon).$$

The sequence (a_n) *converges* to a if for any $\varepsilon > 0$, there exists an $N(\varepsilon)$ such that

$$|a_n - a|_p < \varepsilon \quad \text{whenever } n > N(\varepsilon).$$

Let R be the set of all Cauchy sequences in $\mathbb{Q}$ (for the p-adic metric). It becomes a ring with the obvious operations. An element of R is said to be a *null sequence* if it converges to zero. The set of null sequences is an ideal I in R, and $\mathbb{Q}_p$ is defined to be the quotient R/I.

If $\alpha = (a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, then one shows that $|a_n|_p$ becomes constant for large n, and we set this constant value equal to $|\alpha|_p$. The map $\alpha \mapsto |\alpha|_p : R \to \mathbb{Q}$ factors through $\mathbb{Q}_p$, and has the properties (a), (b), (c) listed above. We can therefore talk about Cauchy sequences etc. in $\mathbb{Q}_p$.

Theorem 2.12. *(a) $\mathbb{Q}_p$ is a field, and it is complete, i.e., every Cauchy sequence in $\mathbb{Q}_p$ has a unique limit in $\mathbb{Q}_p$.*

(b) The map sending $a \in \mathbb{Q}$ to the equivalence class of the constant Cauchy sequence $\alpha(a) = a, a, a, \ldots$ is an injective homomorphism $\mathbb{Q} \hookrightarrow \mathbb{Q}_p$, and every element of $\mathbb{Q}_p$ is a limit of a sequence in $\mathbb{Q}$.

Remark 2.13. (a) The same construction as above, but with $|\cdot|_p$ replaced with the usual absolute value, yields $\mathbb{R}$ instead of $\mathbb{Q}_p$.

(b) Just as real numbers can be represented by decimals, p-adic numbers can be represented by infinite series of the form

$$a_{-n}p^{-n} + \cdots + a_0 + a_1 p + \cdots + a_m p^m + \cdots \qquad 0 \leq a_i \leq p-1.$$

The ring of p-adic integers $\mathbb{Z}_p$ can be variously defined as:

(a) the closure of $\mathbb{Z}$ in $\mathbb{Q}_p$;

(b) the set of elements $\alpha \in \mathbb{Q}_p$ with $|\alpha|_p \leq 1$;

(c) the set of elements of $\mathbb{Q}_p$ that can be represented in the form

$$a_0 + a_1 p + \cdots + a_m p^m + \cdots \quad 0 \leq a_i \leq p-1.$$

(d) the inverse limit $\varprojlim \mathbb{Z}/p^m\mathbb{Z}$.

Some history. Hilbert and Hurwitz showed (in 1890) that, if a curve of genus zero has one rational point, then it has infinitely many, all given by rational values of parameter. Poincaré wrote a long article on the rational points on curves in 1901, and is usually credited with introducing the group law on $E(\mathbb{Q})$ and with conjecturing that $E(\mathbb{Q})$ is finitely generated (the finite basis theorem). According to Schappacher (MathReviews 92c:14001), neither is true: although Poincaré was familiar with the use of chords and tangents to construct new rational points from old, he did not define the group law, and he didn't conjecture the finite basis theorem, he simply assumed it was true. In his remarkable 1922/23 paper, Mordell proved the finite basis theorem, and, in a rather off-handed way, conjectured that all curves of genus > 1 over $\mathbb{Q}$ have only finitely many rational points (the Mordell conjecture). He did this without realizing $E(\mathbb{Q})$ is a group, which complicates his proof, since he has to prove that $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite. The Mordell conjecture was proved by Faltings in 1983 (that year's "theorem of the century"). For an interesting discussion of Mordell's famous paper, and related history, see Cassels, *Mordell's finite basis theorem revisited,* Math. Proc. Camb. Phil. Soc. (1986), 100, 31–41.

Exercise 2.14. (a) Let

$$F(X,Y,Z) = 5X^2 + 3Y^2 + 8Z^2 + 6(YZ + ZX + XY).$$

Find $(a,b,c) \in \mathbb{Z}^3$, not all divisible by 13, such that

$$F(a,b,c) \equiv 0 \mod 13^2.$$

(b) Consider the plane affine curve $C : Y^2 = X^3 + p$. Prove that the point $(0,0)$ on the reduced curve over $\mathbb{F}_p$ does not lift to $\mathbb{Z}_p^2$. Why doesn't this violate Hensel's lemma?

3. The Group Law on a Cubic Curve

Let C be a nonsingular projective plane curve of degree 3 over a field k, which, for simplicity, we assume to be perfect. We are especially interested in the case $k = \mathbb{Q}$.

As we discussed in Section 2, Bezout's theorem (or more elementary arguments) show that the line through two points P and Q on C with coordinates in k will meet the curve in exactly one other point, which will also have coordinates in k. We write PQ for this third point. In special cases, this has to interpreted appropriately: if $P = Q$, then PQ is the point of intersection of the tangent line at P with the cubic; if the line through P and Q is tangent to the cubic at Q, then $PQ = Q$; and if P is a point of inflection, then $PP = P$.

If $C(k)$ is empty, then it is not a group. Otherwise we choose a point $O \in C(k)$, which will be the zero element for the group, and for any pair $P, Q \in C(k)$, we define

$$P + Q = O(PQ),$$

i.e., if the line through P and Q intersects C again at R, then $P + Q$ is the third point of intersection with C of the line through O and R.

[[Diagram omitted]]

Theorem 3.1. *The above construction makes $C(k)$ into a commutative group.*

In this section, we sketch an elementary geometric proof of this, which is very beautiful, at least if one ignores the degenerate cases (as we shall). In the next section, we shall give a different proof based on the Riemann-Roch theorem.

First note that the definition doesn't depend on the order of P and Q; thus

$$P+Q=Q+P.$$

Second note that

$$O+P=_{df}O(OP)=P.$$

Given $P\in C(k)$, define $P'=P(OO)$, i.e., if the tangent line at O intersects C at R, then P' is the third point of intersection of the line through P and R. Then $PP'=OO$, and $O(PP')=O(OO)=O$, and so

$$P+P'=O.$$

[[Diagram omitted]]

Thus the law of composition is commutative, has a zero element, and every element has a negative. It remains to check that it is associative, i.e., that

$$(P+Q)+R=P+(Q+R).$$

Consider the following diagram, in which

$$\begin{aligned}\ell_1 &= \text{the line through } P \text{ and } Q;\\ \ell_2 &= \text{the line through } Q \text{ and } R;\\ \ell_3 &= \text{the line through } O \text{ and } PQ;\\ \ell_4 &= \text{the line through } O \text{ and } QR;\\ \ell_5 &= \text{the line through } P+Q \text{ and } R;\\ \ell_6 &= \text{the line through } Q+R \text{ and } P.\end{aligned}$$

[[diagram omitted]]

Let

$$S=(P+Q)R,\quad T=P(Q+R).$$

Then $(P+Q)+R=OS$ and $P+(Q+R)=OT$. We shall show that $S=T$, which implies that $(P+Q)+R=P+(Q+R)$.

We first need a lemma from linear algebra.

Lemma 3.2. *Let $P_1,\dots,P_8$ be 8 points in $\mathbb{P}^2(k)$ in "general position"[5]. Then there exists a ninth point P_9 such that any cubic curve (not necessarily irreducible or nonsingular) passing through $P_1,\dots,P_8$ also passes through P_9.*

Proof. A cubic form

$$F(X,Y,Z)=a_1X^3+a_2X^2Y+\cdots+a_{10}Z^3$$

has 10 coefficients $a_1,\dots,a_{10}$. The condition that $F(P_i)=0$ is a linear condition on $a_1,\dots,a_{10}$, namely, if $P_i=(x_i:y_i:z_i)$, then it is the condition

$$a_1x_i^3+a_2x_i^2y_i+\cdots+a_{10}z_i^3=0.$$

[5] This is the old geometers way of saying "and satisfying whatever additional conditions are needed to make the proof work".

If the vectors $(x_i^3, x_i^2y_i, \dots, z_i^3)$, $i = 1, \dots, 8$ are linearly independent, then the cubic forms having the P_i as zeros form a 2-dimensional space, and so there exist two such forms F and G such that any other such form can be written

$$\lambda F + \mu G, \quad \lambda, \mu \in k.$$

Now F and G have a ninth zero in common (by Bezout), and every form $\lambda F + \mu G$ passes through it. □

Remark 3.3. In order for the proof to work, we need that the points $P_1, \dots, P_8$ impose linearly independent conditions on the coefficients of the cubic forms. According to [C2], this means that no 7 lie on a conic, and no 4 on a line.

We now complete the proof of the theorem. Consider the cubic curves:

$$C, \quad \ell_1\ell_4\ell_5 = 0, \quad \ell_2\ell_3\ell_6 = 0.$$

All three pass through the 8 marked intersection points in the above diagram, and the last two also pass through the unmarked intersection point. Therefore, if the 8 marked intersection points are in "general position", then the unmarked intersection point is the 9th point through which all cubics must pass. In particular, C passes through the unmarked intersection point, which implies that $S = T$, as required.

Remark 3.4. To give a complete proof, one needs to consider the case that the 8 points are not in general position. There is a detailed elementary proof of this along the lines of the above proof in [Kn], pp67–74.

Alternatively, those who know a little algebraic geometry will be able to complete the proof as follows. We have two regular maps of projective varieties

$$C \times C \times C \to C,$$

namely,

$$(P, Q, R) \mapsto (P + Q) + R \text{ and } (P, Q, R) \mapsto P + (Q + R).$$

The above argument shows that they agree on an open subset of $C \times C \times C$, which is nonempty because it contains (O, O, O). Because C is separated, the set where the two maps is closed, and so is all of $C \times C \times C$.

Remark 3.5. The above construction of the group law makes it obvious that the coordinates of $P + Q$ can be expressed as polynomials in the coordinates of P and Q. For special cubics, we shall find these polynomials later. Also, it is clear that we get the same $P + Q$ whether we consider P and Q as elements of $C(k)$ or of $C(K)$ for some $K \supset k$.

4. Functions on Algebraic Curves and the Riemann-Roch Theorem

Assume (initially) that k is algebraically closed.

Regular functions on affine curves. Let C be an affine plane curve over k defined by an irreducible polynomial $f(X,Y)$. A polynomial $g(X,Y)\in k[X,Y]$ defines a function

$$(a,b)\mapsto g(a,b):C(k)\to k$$

and the functions arising in this manner are called the *regular functions* on C.

Clearly, any multiple of $f(X,Y)$ in $k[X,Y]$ defines the zero function on C, and the Hilbert's Nullstellensatz ([F] p21) shows that the converse is true (using that f irreducible implies $(f(X))$ is a prime ideal). Therefore the map sending g to the function $(a,b)\mapsto g(a,b)$ on C defines an isomorphism

$$k[X,Y]/(f(X))\xrightarrow{\approx}\{\text{ring of regular functions on } C\}.$$

Write

$$k[C]=k[X,Y]/(f(X))=k[x,y].$$

Then x and y can be interpreted as the coordinate functions $P\mapsto x(P)$, $P\mapsto y(P)$ on C, and $k[C]$ is the ring of polynomials in x and y. Note that a nonzero regular function on C will have only finitely many zeros on C, because a curve $g(X,Y)=0$ will intersect C in only finitely many points unless $f(X,Y)|g(X,Y)$.

Because $(f(X))$ is irreducible, $k[x,y]$ is an integral domain, and we let $k(C)=k(x,y)$ be its field of fractions. An element $\varphi=\frac{g}{h}\in k(x,y)$ defines a function

$$(a,b)\mapsto\frac{g(a,b)}{h(a,b)}:C(k)\setminus\{\text{zeros of } h\}\to k.$$

We call φ a *meromorphic function* [6] on C, *regular on* $C\setminus\{\text{zeros of } h\}$.

Example 4.1. (a) Let C be the X-axis, i.e., the affine curve defined by the equation $Y=0$. Then $k[C]=k[X,Y]/(Y)=k[X]$ and $k(C)=k(X)$. The meromorphic functions on C are just the rational functions $\frac{g(X)}{h(X)}$, and such a function is regular outside the finite set of zeros of $h(X)$.

(b) Let C be the curve defined by the equation

$$Y^2=X^3+aX+b.$$

Then $k[C]=k[x,y]=k[X,Y]/(Y^2-X^3-aX-b)$. Thus the regular functions on C are the polynomials in the coordinate functions x and y, which satisfy the relation

$$y^2=x^3+ax+b.$$

Regular functions on projective curves. Let C be a plane projective curve over k defined by an irreducible homogeneous polynomial $F(X,Y,Z)$. If $G(X,Y,Z)$ and $H(X,Y,Z)$ are homogeneous polynomials of the same degree and H is not a multiple of F, then

$$(a:b:c)\mapsto\frac{G(a,b,c)}{H(a,b,c)}$$

is a well-defined function on $C(k)\setminus\{\text{zeros of } H\}$. Such a function is called a meromorphic function on C. More precisely, let

$$k[x,y,z]=k[X,Y,Z]/(F(X,Y,Z)).$$

[6] Note that this is an abuse of language since φ is not in fact a function on all of $C(k)$.

Because F is homogeneous, there is a well-defined decomposition

$$k[x,y,z] = \oplus_d k[x,y,z]_d$$

where $k[x,y,z]_d$ consists of the elements of $k[x,y,z]$ having a representative in $k[X,Y,Z]$ that is homogeneous of degree d. Define

$$k(C) = k(x,y,z)_0 = \{\frac{g}{h} \in k(x,y,z) \mid \exists d \text{ such that } g,h \in k[x,y,z]_d\}.$$

It is a subfield of $k(x,y,z)$, and its elements are called the *meromorphic functions* on C. A meromorphic defines a(n honest) function on the complement of a finite set in $C(k)$. Let U be the complement of a finite set in $C(k)$; then a function $\varphi : U \to k$ is said to be *regular* if there exists a meromorphic function without poles in U and agreeing with φ on U.

Remark 4.2. Recall that we have a bijection

$$\begin{array}{ccccc} \mathbb{A}^2(k) & \leftrightarrow & U_0(k) & \subset & \mathbb{P}^2(k) \\ (\frac{a}{c},\frac{a}{c}) & \leftrightarrow & (a:b:c) & & \end{array}$$

To avoid confusion, write $k[X',Y']$ for the polynomial ring associated with $\mathbb{A}^2$ and $k[X,Y,Z]$ for the polynomial ring associated with $\mathbb{P}^2$. A polynomial $g(X',Y')$ defines a function $\mathbb{A}^2(k) \to k$, and the composite

$$U_0 \to \mathbb{A}^2 \xrightarrow{g(X',Y')} k$$

is

$$(a:b:c) \mapsto g(\frac{a}{c},\frac{b}{c}) = \frac{g^*(a,b,c)}{c^{\deg g}}$$

where $g^*(X,Y,Z) = g(\frac{X}{Z},\frac{Y}{Z}) \cdot Z^{\deg g}$ (in other words, $g^*(X,Y,Z)$ is $g(X,Y)$ made homogeneous by using the smallest number of Z's). Thus $g(X',Y')$ as a function on $\mathbb{A}^2 \cong U_0$ agrees with $\frac{g^*(X,Y,Z)}{Z^{\deg g}}$. One see easily that the map

$$\frac{g(X',Y')}{h(X',Y')} \mapsto \frac{g^*(X,Y,Z)}{Z^{\deg g}} \frac{Z^{\deg h}}{h^*(X,Y,Z)} : k(X',Y') \to k(X,Y,Z)$$

is an injection, with image the subfield $k(X,Y,Z)_0$ of $k(X,Y,Z)$ of elements that can be expressed as the quotient of homogeneous polynomials of the same degree.

Now let C be an irreducible curve in $\mathbb{P}^2$, and assume that $C \cap U_0 \neq \emptyset$, i.e., that C is not the "line at infinity" $Z = 0$. Then the map

$$\frac{g(x',y')}{h(x',y')} \mapsto \frac{g^*(x,y,z)}{z^{\deg g}} \frac{z^{\deg h}}{h^*(x,y,z)} : k(x',y') \to k(x,y,z)_0$$

is a bijection from the field of meromorphic functions on the affine curve $U_0 \cap C$ to the field of meromorphic functions on C. Moreover, if $\varphi' \mapsto \varphi$, then $\varphi(a:b:c) = \varphi'(\frac{a}{c},\frac{b}{c})$ for any point $(a:b:c) \in C(k) \cap U_0$ at which φ is defined.

Example 4.3. (a) The meromorphic functions on $\mathbb{P}^1$ are the functions

$$(a:b) \mapsto \frac{G(a,b)}{H(a,b)}$$

where $G(X,Z)$ and $H(X,Z)$ are homogeneous polynomials of the same degree.

(b) Let C be a nonsingular projective curve over $\mathbb{C}$. Then $C(\mathbb{C})$ has the structure of a compact Riemann surface, and the meromorphic functions on $C(\mathbb{C})$ in the sense of complex analysis are exactly the same functions as those defined above. For example, $\mathbb{P}^1(\mathbb{C})$ is the

Riemann sphere, and, written inhomogeneously, the meromorphic functions in both cases are the functions $\frac{g(z)}{h(z)}$ with each of $g(z)$ and $h(z)$ polynomials.

It is not true that the two notions of meromorphic function coincide for an affine curve: every meromorphic function in the above sense is also meromorphic in the sense of complex analysis, but there are more of the latter, for example, e^z.

The Riemann-Roch theorem. Let C be the nonsingular projective curve over a field k (still assumed to be algebraically closed) defined by a polynomial $F(X, Y, Z)$. One tries to understand the meromorphic functions on a C in terms of their zeros and poles.

The *group of divisors* $\mathrm{Div}(C)$ on C is the free abelian group on the set $C(k)$. Thus an element of $\mathrm{Div}(C)$ is a finite sum

$$D = \sum n_P[P], \quad n_P \in \mathbb{Z}, \quad P \in C(k).$$

The *degree* of D is $\sum n_P$.

There is a partial ordering on $\mathrm{Div}(C)$:

$$\sum n_P[P] \geq \sum m_P[P] \iff n_P \geq m_P \text{ for all } P.$$

In particular, $\sum n_P[P] \geq 0$ if and only if all the n_p are nonnegative.

Let φ be a meromorphic function on C. By definition, φ is defined by a quotient $\frac{G(X,Y,Z)}{H(X,Y,Z)}$ of two polynomials of the same degree, say m, and F doesn't divide H. Assume $\varphi \neq 0$—we may then suppose that F doesn't divide $G(X, Y, Z)$ (recall that $k[X, Y, Z]$ is a unique factorization domain). By Bezout's theorem

$$(\deg F)m = \sum_{F(P)=0=G(P)} I(P, C \cap \{G = 0\}) = \sum_{F(P)=0=H(P)} I(P, C \cap \{H = 0\}).$$

Define the divisor of φ to be

$$\mathrm{div}\varphi = \sum_{G(P)=0=F(P)} I(P, C \cap \{G = 0\})[P] - \sum_{H(P)=0=F(P)} I(P, C \cap \{H = 0\})[P]$$

The $[P]$ occurring in $\mathrm{div}\varphi$ with positive coefficient are called the *zeros* of φ, and those occurring with negative coefficient are its *poles.* Note the $\mathrm{div}\varphi$ has degree zero, and so φ has as many zeros as poles (counting multiplicities). Also, note that only the constant functions will have no zeros or poles.

Given a divisor D, we define

$$L(D) = \{\varphi \mid \mathrm{div}\varphi + D \geq 0\} \cup \{0\}.$$

For example, if $D = [P] + 2[Q]$, then $L(D)$ consists of those meromorphic functions having no poles outside $\{P, Q\}$ and having at worst a single pole at P and a double pole at Q. Each $L(D)$ is a vector space over k, and in fact a finite-dimensional vector space. We denote its dimension by $\ell(D)$.

Theorem 4.4 (Riemann-(Roch)). *There exists an integer g such that for all divisors D*

$$\ell(D) \geq \deg D + 1 - g,$$

with equality for $\deg D$ sufficiently large (in fact, equality for $\deg D > 2g - 2$).

Proof. See [F] Chapter 8. □

The integer defined by the theorem is the genus of C.

Example 4.5. Let $a_1, \dots, a_m \in k = \mathbb{A}^1(k) \subset \mathbb{P}^1(k)$, and let $D = \sum r_i[a_i] \in \mathrm{Div}(\mathbb{P}^1)$, $r_i > 0$. The meromorphic functions φ on $\mathbb{A}^1$ with their poles in $\{a_1, \dots, a_m\}$ and at worst a pole of order r_i at a_i are those of the form

$$\varphi = \frac{f(X)}{(X-a_1)^{r_1}\cdots(X-a_m)^{r_m}}, \quad f(X) \in k[X].$$

The function φ will not have a pole at ∞ if and only if $\deg f \leq \sum r_i = \deg D$. The dimension of $L(D)$ is therefore the dimension of the space of polynomials f of degree $\leq \deg D$, which is $\deg D + 1$. This is as expected, because $\mathbb{P}^1$ has genus 0.

The group law revisited. The divisor of a meromorphic function on C is said to be *principal.* Two divisors D and D' are said to be *linearly equivalent,* $D \sim D'$, if they differ by the divisor of a function.

We have groups

$$\mathrm{Div}(C) \supset \mathrm{Div}^0(C) \supset P(C)$$

where $\mathrm{Div}^0(C)$ is the group of divisors of degree 0 on C, and $P(C)$ is the group of principal divisors. Define *Picard groups:*

$$\mathrm{Pic}(C) = \mathrm{Div}(C)/P(C), \quad \mathrm{Pic}^0(C) = \mathrm{Div}^0(C)/P(C).$$

Remark 4.6. We are interested in these groups when C is a projective curve, but the number theorists may be interested to note that when C is a nonsingular *affine* curve, $k[C]$ is a Dedekind domain, and $\mathrm{Pic}(C)$ is its ideal class group.

Consider now a nonsingular projective curve of genus 1—according to the formula on p9, a nonsingular plane projective curve will have genus 1 if and only if it has degree 3. According to the Riemann-Roch theorem,

$$\ell(D) \;=\; \deg D \text{ if } \deg D \geq 1.$$

Proposition 4.7. *Let C be a nonsingular projective curve of genus 1, and let $O \in C(k)$. The map*

$$P \mapsto [P] - [O] : C(k) \to \mathrm{Pic}^0(C)$$

is bijective.

Proof. We define an inverse. Let D be a divisor of degree 0. Then $D + [O]$ has degree 1, and so there exists a meromorphic function φ, unique up to multiplication by a nonzero constant, such that $\mathrm{div}(\varphi) + D + [O] \geq 0$. The only divisors ≥ 0 of degree 1 are of the form $[P]$. Hence there is a well-defined point P such that $D + [O] \sim [P]$, i.e., such that $D \sim [P] - [O]$. □

Thus we have a canonical bijection $C(k) \to \mathrm{Pic}^0(C)$, from which $C(k)$ inherits the structure of an abelian group. Note that this group structure is determined by the condition: $P + Q = S$ if and only if $[P] + [Q] \sim [S] + [O]$.

I claim that this is the same structure as defined in the last section. Let $P, Q \in C(k)$, and suppose $P + Q = S$ with the law of composition in §3. Let L_1 be the line through P and Q, and let L_2 be the line through O and S. From the definition of S, we know that L_1 and

L_2 have a common point R as their third points of intersection. Regard L_1 and L_2 as linear forms in X, Y, Z, and let $\varphi = \frac{L_1}{L_2}$. Then φ has zeros at P, Q, R and poles at O, S, R, and so

$$\operatorname{div}(\varphi) = [P] + [Q] + [R] - [O] - [S] - [R] = [P] + [Q] - [S] - [O].$$

Henc $[P] + [Q] \sim [S] + [O]$, and $P + Q = S$ according to the group structure defined by the bijection.

Perfect base fields. Let C be a nonsingular absolutely irreducible plane projective curve over a perfect field k (e.g., a field of characteristic zero or a finite field), and let $F(X, Y, Z)$ be the polynomial defining C. We can again form $k[x, y, z] = k[X, Y, Z]/(F(X, Y, Z))$—it is an integral domain, and remains so even when tensored with k^{al}—and the field $k(x, y, z)_0 \subset k(x, y, z)$. We define $k(x, y, z)_0$ to be the field of meromorphic functions of C. We can no longer identify its elements with functions on $C(k) \setminus \{\text{finite set}\}$, because, for example, $C(k)$ may be empty. However, we can identify its elements with functions on $C(k^{\mathrm{al}}) \setminus \{\text{finite set}\}$. From another perspective, we can say that a meromorphic function φ on $C(k^{\mathrm{al}})$, i.e., an element of $k^{\mathrm{al}}(x, y, z)_0$, is *defined over (or rational over)* k if it lies in the subfield $k(x, y, z)_0$ of $k^{\mathrm{al}}(x, y, z)_0$.

The Galois group $\Gamma = \operatorname{Gal}(k^{\mathrm{al}}/k)$ acts on $C(k^{\mathrm{al}})$. Its orbits are finite because every $P \in C(k^{\mathrm{al}})$ has coordinates in some finite extension of k. We deduce an action of Γ on $\operatorname{Div}(C(k^{\mathrm{al}}))$: $\tau(\sum n_P[P]) = \sum n_p[\tau P]$. A divisor D is said to be *defined over (or rational over)* k if it is fixed by this action. Thus D is rational over k if and only if all P in each Γ-orbit have the same coefficients n_P in D.

For a divisor D on C rational over k, we can define $L(D)$ to be the set of all meromorphic functions on C rational over k such that $\operatorname{div}\varphi + D \geq 0$ (together with 0). Then the Riemann-Roch theorem continues to hold, and there is a bijection

$$P \mapsto [P] - [O] : C(k) \to \operatorname{Pic}^0(C)$$

where $\operatorname{Pic}^0(C) = (\operatorname{Pic}^0(C(k^{\mathrm{al}}))^{\Gamma}$, i.e., it is the group of divisor classes of degree zero (over k^{al}) fixed by the action of Γ. Unfortunately, such a class need not be represented by a divisor fixed by the action of Γ.

Exercise 4.8. Find a necessary and sufficient condition for the line $L : Y = cX + d$ to be an inflectional tangent to the affine curve $C : Y^2 = X^3 + aX + b$, i.e., to meet C at a point P with $I(P, L \cap C) = 3$. Hence find a general formula for the elliptic curves in canonical form having a rational point of order 3.

5. Definition of an Elliptic Curve

Definition 5.1. Let k be a field. An *elliptic curve* over k can be defined, according to taste, as:

(a) A complete nonsingular curve E of genus 1 over k together with a point $O \in E(k)$.
(b) A nonsingular plane projective E of degree 3 together with a point $O \in E(k)$.
(c) A nonsingular plane projective curve E of the form

$$Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3$$

The relation between these definitions is as follows. Let E be as in (c). Then $E(k)$ contains a canonical element $O=(0:1:0)$, and the pair (E,O) satisfies the other two definitions. This is obvious for (b), and (a) follows from the formula on p9.

Let (E,O) be as in (a). Then (see below) there is an isomorphism from E onto a curve as in (c) sending O to $(0:1:0)$.

Let (E,O) be as in (b). Then (see below) there is a change of variables transforming E into a curve as in (c) and O into $(0:1:0)$ (and if O is a point of inflection, the change of variables can be taken to be linear).

Plane projective cubic curves with a rational inflection point. Let (E,O) be a nonsingular cubic curve in $\mathbb{P}^2(k)$. In this subsection, I assume that O is a point of inflection, and I show that:

(a) after a linear change of variables (with coefficients in k), the point O will be $(0:1:0)$ and its tangent line will be $L_\infty : Z=0$;
(b) if $(0:1:0)\in E(k)$ and the tangent line to E at $(0:1:0)$ is $L_\infty : Z=0$, then the equation of E has the form (5.1c).

Proof of (a). Let $(a:b:c)\in\mathbb{P}^2(k)$, and assume $b\neq 0$. The map

$$(x:y:z)\to(bx-ay:by:bz-cy)$$

is well-defined for all $(x:y:z)\in\mathbb{P}^2(k)$ and sends $(a:b:c)$ to $(0:b^2:0)=(0:1:0)$. If $b=0$, but $c\neq 0$, we first interchange the y and z coordinates.

Consider a line

$$L: aX+bY+cZ=0,\quad a,b,c\in k,\quad \text{not all } a,b,c \text{ zero.}$$

Choose $A=(a_{ij})$ to be an invertible 3×3 matrix whose first two columns are orthogonal to (a,b,c), and define a change of variables

$$A\begin{pmatrix}X'\\Y'\\Z'\end{pmatrix}=\begin{pmatrix}X\\Y\\Z\end{pmatrix}.$$

With respect to the variables X',Y',Z', the equation of the line L becomes

$$0=(a,b,c)\begin{pmatrix}X\\Y\\Z\end{pmatrix}=(a,b,c)A\begin{pmatrix}X'\\Y'\\Z'\end{pmatrix}=(0,0,d)\begin{pmatrix}X'\\Y'\\Z'\end{pmatrix}=dZ'.$$

Moreover, $d\neq 0$, and so we may take the equation of the line to be $Z'=0$.

This completes the proof of (a).

Proof of (b). The general cubic form is $F(X,Y,Z)$:

$$c_1X^3+c_2X^2Y+c_3X^2Z+c_4XY^2+c_5XYZ+c_6XZ^2+c_7Y^3+c_8Y^2Z+c_9YZ^2+c_{10}Z^3.$$

Let E be the curve $F(X,Y,Z)=0$. We assume that E is nonsingular; in particular, this implies that F is absolutely irreducible.

If $O=(0:1:0)$ lies on E, then $\boxed{c_7=0}$.

Recall that $U_1=\{(x:y:z)\mid y=1\}$, and we identify U_1 with $\mathbb{A}^2$ via $(x:1:z)\leftrightarrow(x:z)$. The curve $C\cap U_1$ is an affine curve with equation $F(X,1,Z)$:

$$c_1X^3+c_2X^2+c_3X^2Z+c_4X+c_5XZ+c_6XZ^2+c_8Z+c_9Z^2+c_{10}Z^3.$$

The tangent line at $(0:1:0) \leftrightarrow (0,0)$ is

$$c_4 X + c_8 Z = 0.$$

If this is $L_\infty : Z = 0$, then $\boxed{c_4 = 0}$. Since E is nonsingular, we must have $\boxed{c_8 \neq 0}$.

According to (1.5), the intersection number

$$I(O, L_\infty \cap E) \leq \dim_k k[X,Z]/(Z, F(X,1,Z)).$$

But

$$k[X,Z]/(Z, F(X,1,Z)) \approx k[X]/(c_2 X^2 + c_1 X^3).$$

If O is a point of inflection, then $I(O, L_\infty \cap E) \geq 3$, and so $\boxed{c_2 = 0}$.

On combining the boxed statements, we find that our equation has become

$$c_1 X^3 + c_3 X^2 Z + c_5 XYZ + c_6 XZ^2 + c_8 Y^2 Z + c_9 YZ^2 + c_{10} Z^3, \quad c_8 \neq 0.$$

Moreover, $c_1 \neq 0$ because otherwise the polynomial is divisible by Z. Finally, after dividing through by c_1 and replacing Z with $\frac{Z}{-c_8}$, we obtain an equation of the same form as that in (5.1c).

Remark 5.2 (Remedial math.). [7] The *Hessian* of a homogeneous polynomial $F(X,Y,Z)$ is the matrix

$$H(X,Y,Z) = \begin{pmatrix} \frac{\partial^2 F}{\partial X^2} & \frac{\partial^2 F}{\partial X \partial Y} & \frac{\partial^2 F}{\partial X \partial Z} \\ \frac{\partial^2 F}{\partial X \partial Y} & \frac{\partial^2 F}{\partial Y^2} & \frac{\partial^2 F}{\partial Y \partial Z} \\ \frac{\partial^2 F}{\partial X \partial Z} & \frac{\partial^2 F}{\partial Y \partial Z} & \frac{\partial^2 F}{\partial Z^2} \end{pmatrix}.$$

Assume char $k \neq 2$. Then a nonsingular point $(a:b:c)$ on the curve $C : F = 0$ is a point of inflection if and only if $\det H(a,b,c) = 0$ ([Kn] p38). This fact can be used to find a point of inflection over k on a cubic curve (when it exists), and once one has such a point, the above procedure allows one to find an equation for the curve in the form (5.1c) (ibid. 3.1).

If F has degree d, then $\det H$ has degree $3(d-2)$. Thus, an irreducible cubic curve has at least one point of inflection over k^{al}, and at most 3. Unfortunately, it may have no point of inflection with coordinates in k.

General plane projective curves. As we just noted, a plane projective cubic curve may not have a point of inflection with coordinates in k. An invertible linear change of variables will not change this (it will only multiply the Hessian by a nonzero constant). However, if the curve has some point with coordinates in k, then there is a method (due to Nagell, 1928/29) that transforms the equation by a nonlinear change of variables so that the point becomes a point of inflection, still with coordinates in k ([C2], p34).

Complete nonsingular curves of genus 1. Here we assume some algebraic geometry. Let E be a complete nonsingular curve of genus 1 over a field k and let $O \in E(k)$. According to the Riemann-Roch theorem, the meromorphic functions on E having no poles except at O and having at worst a pole of order $m \geq 1$ at O, form a vector space of dimension m over k, i.e., $L(m[O])$ has dimension m for $m \geq 1$.

The constant functions lie in $L([O])$, and according to the Riemann-Roch theorem, there are no other. Thus 1 is a basis for $L([O])$.

Choose x so that $\{1, x\}$ is a basis for $L(2[O])$.

[7] These are topics that were once taught in high school, but are no longer taught anywhere, well, hardly anywhere.

Choose y so that $\{1, x, y\}$ is a basis for $L(3[O])$.

Then $\{1, x, y, x^2\}$ is a basis for $L(4[O])$. (If it were linearly dependent, then x^2 would have to be a linear combination of $1, x, y$, but then it couldn't have a quadruple pole at O.)

And $\{1, x, y, x^2, xy\}$ is a basis for $L(5[O])$.

The subset $\{1, x, y, x^2, xy, x^3, y^2\}$ of $L(6[O])$ contains 7 elements, and so must be linearly dependent: there exist constants a_i such that

$$a_0 y^2 + a_1 xy + a_3 y = a_0' x^3 + a_2 X^2 + a_4 X + a_6$$

(as regular functions on $E \setminus \{O\}$). Moreover, a_0 and a_0' must be nonzero, because the set with either x^3 or y^2 omitted is linearly independent, and so, after multiplying through by a constant and making a change of variables, we can suppose both equal 1. The map $P \mapsto (x(P), y(P))$ sends $E \setminus \{O\}$ onto the plane affine curve

$$C : Y^2 + a_1 XY + a_3 Y = X^3 + a_2 X^2 + a_4 X + a_6.$$

The function x has a double pole at O and no other pole, and so it has only two other zeros. Therefore, the composite

$$E \setminus \{O\} \to C \to \mathbb{A}^1, \quad P \mapsto (x(P), y(P)) \mapsto x(P)$$

has degree 2, i.e., it is $2:1$ on points with coordinates in k^{al} (at least, if the characteristic is zero). Similarly, the composite

$$E \setminus \{O\} \to C \to \mathbb{A}^1, \quad P \mapsto (x(P), y(P)) \mapsto y(P)$$

has degree 3. The degree of $E \setminus \{O\} \to C$ divides both 2 and 3, and therefore is 1. In fact, it is an isomorphism, and it extends to an isomorphism of E onto the Zariski closure of C in $\mathbb{P}^2$, i.e., onto the curve given by an equation of the form (5.1c).

The canonical form of the equation. Thus, however we define it, an elliptic curve is isomorphic to a curve of the form

$$E : Y^2 Z + a_1 XYZ + a_3 YZ^2 = X^3 + a_2 X^2 Z + a_4 XZ^2 + a_6 Z^3,$$

and, conversely, every nonsingular such E is an elliptic curve. This is usually referred to as the *canonical* or *Weierstrass* equation of the curve, but how canonical is it? One can show that it is canonical up to a change of variables of the form:

$$\begin{aligned} X &= u^2 X' + r \\ Y &= u^3 Y' + s u^2 X' + t \end{aligned}$$

with $u, r, s, t \in k$ and $u \neq 0$.

Everything becomes simpler if we assume that char$k \neq 2, 3$. A change of variables

$$X' = X, \quad Y' = Y + \frac{a_1}{2} X, \quad Z' = Z$$

will eliminate the XYZ term, and a change of variables

$$X' = X + \frac{a_2}{3}, \quad Y' = Y + \frac{a_3}{2}, \quad Z' = Z$$

will then eliminate the X^2 and Y terms. Thus we arrive at the equation:

$$Y^2 Z = X^3 + aXZ^2 + bZ^3.$$

Theorem 5.3. *Assume char $k \neq 2, 3$.*

(a) The curve

$$E(a,b) : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in k,$$

is nonsingular, and hence (together with $O = (0:1:0)$) defines an elliptic curve over k, if and only if $4a^3 + 27b^2 \neq 0$.

(b) Every elliptic curve over k is isomorphic to one of the form $E(a,b)$.

(c) Two elliptic curves $E(a,b)$ and $E(a',b')$ are isomorphic if and only if there exists a $c \in k^\times$ such $a' = c^4a$, $b' = c^6b$; the isomorphism is then

$$(x:y:z) \mapsto (c^2x : c^3y : z).$$

Proof. (a) We proved in (1.3) that the affine curve

$$Y^2 = X^3 + aX + b$$

is nonsingular if and only if $4a^3 + 27b^2 \neq 0$. The point $(0:1:0)$ is always nonsingular.

(b) This was discussed above.

(c) The "if" is obvious. We omit the "only if" (see, for example, [S1] pp64–65). □

Remark 5.4. For an elliptic curve E, define

$$j(E) = \frac{1728(4a^3)}{4a^3 + 27b^2}$$

if $E \approx E(a,b)$. Since the expression on the right is unchanged when (a,b) is replaced by (c^4a, c^6b), this is well-defined, and $E \approx E' \implies j(E) = j(E')$. Conversely, $j(E) = j(E') \implies E \approx E'$ when k is algebraically closed (see later), but not otherwise. For example, if c is not a square in k, then

$$Y^2Z = X^3 + ac^2XZ^2 + bc^3Z^3$$

has the same j invariant as $E(a,b)$, but it is not isomorphic to it.

The group law for the canonical form. The point at ∞ is the zero for the group law. The group law is determined by:

$$P + Q + R = O \iff P, Q, R \text{ lie on a straight line;}$$

$$\text{if } P = (x:y:z), \text{ then } -P = (x:-y:z).$$

In particular, $-P = P$, i.e., P has order 2 if and only if $y = 0$. The points of order two are the points $(x:0:1)$ where x is a root of $X^3 + aX + b$.

6. Reduction of an Elliptic Curve Modulo p

Consider an elliptic curve

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in \mathbb{Q}, \quad \Delta = 4a^3 + 27b^2 \neq 0.$$

After a change of variables we may suppose $a, b \in \mathbb{Z}$, and so we may look at them modulo p to get a curve over $\mathbb{F}_p =_{df} \mathbb{Z}/p\mathbb{Z}$. In this section, we examine what curves we get in this fashion.

Algebraic groups of dimension 1. Let k be an arbitrary perfect field. The following is a complete list of connected algebraic curves over k having group structures defined by polynomial maps.

Elliptic curves. These are the only projective curves having a group structure defined by polynomial maps.

The additive group. The affine line $\mathbb{A}^1$ is a group under addition:

$$\mathbb{A}^1(k) = k, \quad (x, y) \mapsto x + y : k \times k \to k.$$

We sometimes write $\mathbb{G}_a$ for $\mathbb{A}^1$ endowed with this group structure.

The multiplicative group. The affine line with the origin removed is a group under multiplication:

$$\mathbb{A}^1(k) \setminus \{(0)\} = k^\times, \quad (x, y) \mapsto xy : k^\times \times k^\times \to k^\times.$$

We sometimes write $\mathbb{G}_m$ for $\mathbb{A}^1 \setminus \{(0)\}$ endowed with this group structure. Note that the map $x \mapsto (x, x^{-1})$ identifies $\mathbb{G}_m$ with the plane affine curve $XY = 1$.

Twisted multiplicative groups. Let $a \in k \setminus k^2$, and let $L = k[\sqrt{a}]$. There is an algebraic group G over k such that

$$G(k) = \{\gamma \in L^\times \mid \mathrm{Nm}_{L/k}\, \gamma = 1\}.$$

Let $\alpha = \sqrt{a}$, so that $\{1, \alpha\}$ is a basis for L as a k-vector space. Then $\mathrm{Nm}(x + \alpha y) = x^2 - ay^2$ and $(x + \alpha y)(x' + \alpha y') = xx' + ayy' + \alpha(xy' + x'y)$. We define G to be the plane affine curve $X^2 - aY^2 = 1$, with the group structure

$$(x, y) \times (x', y') = (xx' + ayy', xy' + x'y).$$

We denote this group by $\mathbb{G}_m[a]$. For example, when $k = \mathbb{R}$ and $a = -1$, we get the circle group $X^2 + Y^2 = 1$.

Note that a change of variables transforms $\mathbb{G}_m[a]$ into $\mathbb{G}_m[ac^2]$, any $c \neq 0$, and so, up to a change of variables, $\mathbb{G}_m[a]$ depends only on the field $k[\sqrt{a}]$. The equations defining $\mathbb{G}_m[a]$ still define an algebraic group when a is a square in k, say $a = \alpha^2$, but then $X^2 - aY^2 = (X + \alpha Y)(X - \alpha Y)$, and so the change of variables $X' = X + \alpha Y$, $Y' = X - \alpha Y$ transforms the group into $\mathbb{G}_m$. In particular, this shows that $\mathbb{G}_m[a]$ becomes isomorphic to $\mathbb{G}_m$ over $k[\sqrt{a}]$, and so it can be thought of a "twist" of $\mathbb{G}_m$.

Remark 6.1. Let $k = \mathbb{F}_q$, the field with q-elements. Then $\mathbb{G}_a(k)$ has q-elements, $\mathbb{G}_m(k)$ has $q - 1$ elements, and $\mathbb{G}_m[a](k)$ has $q + 1$ elements (here a is any nonsquare in k). Only the last is not obvious. From the definition of $\mathbb{G}_m[a]$, we know there is an exact sequence

$$0 \to \mathbb{G}_m[a](\mathbb{F}_q) \to (\mathbb{F}_{q^2})^\times \xrightarrow[\mathrm{Nm}]{} \mathbb{F}_q^\times \to 0.$$

The second map is surjective (because a quadratic form in at least three variables over a finite field always has a nontrivial zero (Serre, Course on Arithmetic)), and so

$$\#\mathbb{G}_m[a](\mathbb{F}_q) = (q^2 - 1)/(q - 1) = q + 1.$$

We make a few remarks concerning the proofs of the above statements. We have seen that if a nonsingular projective curve has genus 1, then its has a group structure, but why is the converse true? The simplest explanation (for the case $k = \mathbb{C}$) comes from topology. The Lefschetz fixed point theorem[8] says that, if M is a compact oriented manifold, then for any continuous map $\alpha : M \to M$,

$$(\Delta \cdot \Gamma_\alpha) = \sum (-1)^i \mathrm{Trace}(\alpha | H^i(M, \mathbb{Q})).$$

Here Δ is the diagonal in $M \times M$ and Γ_α is the graph of α, so that $(\Delta \cdot \Gamma_\alpha)$ is the number of "fixed points of α counting multiplicities". Let $L(\alpha)$ be the integer on the right. If M has a group structure, then the translation map $\tau_a = (x \mapsto x + a)$, $a \neq 0$, has no fixed points, and so

$$L(\tau_a) = (\Delta \cdot \Gamma_\alpha) = 0.$$

But $a \mapsto L(\tau_a) : M \to \mathbb{Z}$ is continuous, and hence constant on each connected component. On letting a tend to zero, we find that $L(\tau_0) = 0$. But τ_0 is the identity map, and so

$$L(\tau_0) = \sum (-1)^i \operatorname{Tr}(\mathrm{id} \,| H^i(M, \mathbb{Q})) = \sum (-1)^i \dim_{\mathbb{Q}} H^i(M, \mathbb{Q}).$$

Thus, if M has a group structure, its Euler-Poincaré characteristic must be zero. The Euler-Poincaré characteristic of a complex curve of genus g is $1 - 2g + 1 = 2 - 2g$, and so $g = 1$ the curve has a continuous group structure.

A similar argument works over any field. One proves directly that for the diagonal Δ in $C \times C$,

$$(\Delta \cdot \Delta) = 2 - 2g, \quad (\Delta \cdot \Gamma_{\tau_a}) = 0, \quad a \neq 0,$$

and then "by continuity" that $(\Delta \cdot \Delta) = (\Delta \cdot \Gamma_{\tau_a})$.

The proof that $\mathbb{G}_a$ and $\mathbb{G}_m$ are the only affine algebraic groups of dimension one can be found in most books on algebraic groups when k is algebraically closed (see for example, Borel, Linear Algebraic Groups, 10.9, who notes that the first published proof is in an article of Grothendieck). The extension to nonalgebraically closed fields is an easy exercise in Galois cohomology.

Singular cubic curves. Let E be a singular plane projective curve over a perfect field k of characteristic $\neq 2$. As we noted on p8, it will have exactly one singular point S, and S will have coordinates in k. Assume $E(k)$ contains a point $O \neq S$. It is a curious fact that exactly the same definition as in the nonsingular case turns $E(k) \setminus \{S\}$ into a group. Namely, consider the line through two nonsingular points P and Q. According to Bezout's theorem and (1.7), it will intersect the curve in exactly one additional point R, which can't be singular. Define $P + Q$ to be the third point of intersection of the line through R and O with the cubic. We examine this in the three possible cases.

Cubic curves with a cusp. The plane projective curve

$$E : Y^2 Z = X^3$$

has a cusp at $S = (0 : 0 : 1)$ because the affine curve $Y^2 = X^3$ has a cusp at $(0, 0)$. Note that S is the only point on the projective curve with Y-coordinate zero, and so $E(k) \setminus \{S\}$ is equal to the set of points on the affine curve $E \cap \{Y \neq 0\}$, i.e., on the curve

$$E_1 : Z = X^3.$$

[8] See, for example, Greenberg, Lectures on Algebraic Topology.

The line $Z = \alpha X + \beta$ intersects E_1 at the points $P_1 = (x_1, z_1)$, $P_2 = (x_2, z_2)$, $P_3 = (x_3, z_3)$ with x_1, x_2, x_3 roots of

$$X^3 - \alpha X - \beta.$$

Because the coefficient of X^2 in this polynomial is zero, the sum $x_1 + x_2 + x_3$ of its roots is zero. Therefore the map

$$P \mapsto x(P) : E_1(k) \to k$$

has the property that

$$P_1 + P_2 + P_3 = 0 \implies x(P_1) + x(P_2) + x(P_3) = 0.$$

Since $O = (0, 0)$, the map $P \mapsto -P$ is $(x, z) \mapsto (-x, -z)$, and so $P \mapsto x(P)$ also has the property that

$$x(-P) = -P.$$

These two properties imply that $P \mapsto x(P) : E_1(k) \to k$ is a homomorphism. In fact, it is an isomorphism. In summary: the map $P \mapsto \frac{x(P)}{y(P)} : E(k) \setminus S \to \mathbb{G}_a(k)$ is an isomorphism.

Cubic curves with a node. The curve

$$Y^2Z = X^3 + cX^2Z, \quad c \neq 0,$$

has a node at $(0 : 0 : 1)$ because the affine curve

$$Y^2 = X^3 + cX^2, \quad c \neq 0,$$

has a node at $(0, 0)$. The tangent lines at $(0, 0)$ are given by the equation

$$Y^2 - cX^2 = 0.$$

If c is a square, this factors as

$$(Y - \sqrt{c}X)(Y + \sqrt{c}X) = 0$$

and we get two tangent lines. In this case the tangent lines are said to be *rational* (i.e., defined) over k. When endowed with its group structure, $E \setminus \{\text{singular point}\}$ becomes isomorphic to $\mathbb{G}_m$.

If c is not a square, so the tangent lines are not rational over k, then $E \setminus \{\text{singular point}\} \approx \mathbb{G}_m[c]$. See $[C_2]$, Chapter 9.

Criterion. We now derive a criterion for deciding which of the above cases the curve

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in k, \quad \Delta = 4a^3 + 27b^2 = 0$$

falls into. We assume char $k \neq 2, 3$. Since the point $(0 : 1 : 0)$ is always nonsingular, we only need to study the affine curve

$$E_0 : Y^2 = X^3 + aX + b.$$

We try to find a t such that equation is

$$\begin{aligned} Y^2 &= (X - t)^2(X + 2t) \\ &= X^3 - 3t^2X + 2t^3. \end{aligned}$$

For this, we need to choose t so that

$$t^2 = -\frac{a}{3}, \quad t^3 = \frac{b}{2}.$$

Hence $t = \frac{b/2}{-a/3} = -\frac{3}{2}\frac{b}{a}$. Using that $\Delta = 0$, one checks that this works.

Now, we can rewrite the equation as

$$Y^2 = 3t(X-t)^2 + (X-t)^3.$$

This has a singularity at $(t,0)$, which is a cusp if $3t = 0$, a node with rational tangents if $3t$ is a nonzero square in k, and a node with nonrational tangents if $3t$ is a nonzero nonsquare. Note that

$$-2ab = -2(-3t^2)(2t^3) = (2t^2)^2(3t)$$

and so $3t$ is zero or nonzero, a square or a nonsquare, according as $-2ab$ is.

Reduction of an elliptic curve. Consider an elliptic curve

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in \mathbb{Q}, \quad \Delta = 4a^3 + 27b^2 \neq 0.$$

We make a change a variables $X \mapsto X/c^2$, $Y \mapsto Y/c^3$ with c chosen so that the new a, b are integers and $|\Delta|$ is minimal—such an equation is said to be *minimal.* The equation

$$\bar{E} : Y^2Z = X^3 + \bar{a}XZ^2 + \bar{b}Z^3$$

where $\bar{a}$ and $\bar{b}$ are the images of a and b in $\mathbb{F}_p$ is called the *reduction of E modulo p.*

There are three cases to consider (and two subcases).

(a) *Good reduction.* If $p \neq 2$ and p does not divide Δ, then $\bar{E}$ is an elliptic curve over $\mathbb{F}_p$. For a point $P = (x : y : z)$ on E, we can choose a representative (x, y, z) for P with $x, y, z \in \mathbb{Z}$ and having no common factor, and then $\bar{P} =_{df} (\bar{x} : \bar{y} : \bar{z})$ will be a well-defined point on $\bar{E}$. Since $(0 : 1 : 0)$ reduces to $(0 : 1 : 0)$ and lines reduce to lines, the map $E(\mathbb{Q}) \to \bar{E}(\mathbb{F}_p)$ is a homomorphism, and Hensel's lemma implies that $E(\mathbb{Q}_p) \to \bar{E}(\mathbb{F}_p)$ is surjective (see 2.10). The Riemann hypothesis (see later) shows that

$$|\#\bar{E}(\mathbb{F}_p) - p - 1| \leq 2\sqrt{p}.$$

(b) *Cuspidal reduction.* Here the reduced curve has a cusp as singularity. For $p \neq 2, 3$, it occurs exactly when $p|4a^3 + 27b^2$ and $p| - 2ab$.

(c) *Nodal reduction.* Here the reduced curve has a node as singularity. For $p \neq 2, 3$, it occurs exactly when $p|4a^3 + 27b^2$ and p does not divide $-2ab$. The tangents at the node are rational over $\mathbb{F}_p$ if and only if $-2ab \mod p$ is a square in $\mathbb{F}_p$.

The following table summarizes our results ($p \neq 2, 3$, N is the number of nonsingular points on $\bar{E}$ with coordinates in $\mathbb{F}_p$).

Reduction	$\Delta \mod p$	$-2ab \mod p$	E^{ns}	N
good	$\neq 0$		$\overline{E}$	$\lvert N - p - 1\rvert \leq 2\sqrt{p}$
cusp	0	0	$\mathbb{G}_a$	p
node; rational tangents	0	$\square$	$\mathbb{G}_m$	$p-1$
node; nonrational tangents	0	$\neq \square$	$\mathbb{G}_m[-2\bar{a}\bar{b}]$	$p+1$

Other names:

cuspidal = additive;

nodal with rational tangents = split multiplicative;

nodal with nonrational tangents = nonsplit multiplicative.

Semistable reduction. If E has good or nodal reduction, then the minimal equation remains minimal after replacing the ground field (here $\mathbb{Q}$) by a larger field. This is not so for cuspidal reduction. Consider, for example, the curve

$$E : Y^2Z = X^3 + pXZ^2 + pZ^3.$$

After passing to a larger extension, in which p is a sixth power, say, $c^6 = p$, we can make a change of variables so that the equation becomes

$$E : Y^2Z = X^3 + c^2XZ^2 + Z^3.$$

This reduces to

$$Y^2Z = X^3 + Z^3,$$

which is nonsingular. In fact, for any curve E with cuspidal reduction at p, there will exist a finite extension of the ground field such that E will have either have good or nodal reduction at the primes over p.

In summary: good and nodal reduction are not changed by a field extension (in fact, the minimal model remains minimal) but cuspidal reduction always becomes good or nodal reduction in an appropriate finite extension (and the minimal model changes). For this reason, a curve is often said to have *semistable* reduction at p if it has good or nodal reduction there.

Reduction modulo 2 and 3. When considering reduction at 2 or 3, one needs to consider the full equation

$$Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3$$

because it may be possible to find an equation of this form that is "more minimal" for 2 or 3 than any of the form

$$Y^2Z = X^3 + aXZ^2 + bZ^3.$$

For example, it may be possible to find one of the first form that gives a nonsingular curve over $\mathbb{F}_2$, whereas all equations of the second form become singular over $\mathbb{F}_2$ (see 1.3).

Other fields. Throughout this section, we can replace $\mathbb{Q}$ and $\mathbb{Z}$ with $\mathbb{Q}_p$ and $\mathbb{Z}_p$, or in fact with any local field and its ring of integers. Also, we can replace $\mathbb{Q}$ and $\mathbb{Z}$ with a number field K and its ring of integers, with the caution that, for a number field K with class number $\neq 1$, it may not be possible to find an equation for the elliptic curve that is minimal for all primes simultaneously.

Exercise 6.2. (a) Find examples of elliptic curves E over $\mathbb{Q}$ such that

(i) $\bar{E}$ has a cusp S which lifts to a point in $E(\mathbb{Q}_p)$;

(ii) $\bar{E}$ has a node S which lifts to a point in $E(\mathbb{Q}_p)$;

(iii) $\bar{E}$ has a node S which does not lift to a point in $E(\mathbb{Q}_p)$.

Here $\bar{E}$ is the reduction of the curve modulo a prime $p \neq 2, 3$. The equation you give for E should be a minimal equation of the standard form $Y^2Z = X^3 + aXZ^2 + bZ^3$.

(b) For the example you gave in (a)(i), decide whether it acquires good or nodal reduction in a finite extension of $\mathbb{Q}$.

7. Elliptic Curves over $\mathbb{Q}_p$

Notation: A nonzero rational number a can be written $a = p^m \frac{r}{s}$ with r and s not divisible by p. We then set $\operatorname{ord}_p(a) = m$. The following rule is obvious:

$$\operatorname{ord}_p(a+b) \geq \min\{\operatorname{ord}_p(a), \operatorname{ord}_p(b)\}, \text{ with equality unless } \operatorname{ord}_p(a) = \operatorname{ord}_p(b).$$

Similarly, for an $a \in \mathbb{Q}_p$, we set $\operatorname{ord}_p(a) = m$ if $a \in p^m\mathbb{Z}_p \setminus p^{m+1}\mathbb{Z}_p$. The same rule holds, and the two definitions of ord_p agree on $\mathbb{Q}$. In both cases, we set $\operatorname{ord}_p(0) = \infty$. Note that ord_p is a homomorphism $\mathbb{Q}_p^\times \to \mathbb{Z}$.

Consider a curve

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in \mathbb{Q}_p, \quad 4a^3 + 27b^2 \neq 0.$$

After a change of variables $X \mapsto X/c^2$, $Y \mapsto Y/c^3$, $Z \mapsto Z$, we may suppose that $a, b \in \mathbb{Z}_p$. As in the last section, we obtain from E a curve $\bar{E}$ over $\mathbb{F}_p$ and a reduction map

$$P \mapsto \bar{P} : E(\mathbb{Q}_p) \to \bar{E}(\mathbb{F}_p).$$

We shall define a filtration

$$E(\mathbb{Q}_p) \supset E^0(\mathbb{Q}_p) \supset E^1(\mathbb{Q}_p) \supset \cdots \supset E^n(\mathbb{Q}_p) \supset \cdots$$

and identify the quotients. First, define

$$E^0(\mathbb{Q}_p) = \{P \mid \bar{P} \text{ is nonsingular}\}.$$

It is a subgroup because, as we observed on p26, a line through two nonsingular points on a cubic (or tangent to a nonsingular point), will meet the cubic again at a nonsingular point.

Write $\bar{E}^{\mathrm{ns}}$ for $\bar{E} \setminus \{\text{any singular point}\}$. The reduction map

$$P \mapsto \bar{P} : E^0(\mathbb{Q}_p) \to \bar{E}^{\mathrm{ns}}(\mathbb{F}_p)$$

is a homomorphism, and we define $E^1(\mathbb{Q}_p)$ be its kernel. Thus $E^1(\mathbb{Q}_p)$ consists of the points P that can be represented as $(x : y : z)$ with x and z divisible by p but y not divisible by p. In particular, $P \in E^1(\mathbb{Q}_p) \implies y(P) \neq 0$.

Define

$$E^n(\mathbb{Q}_p) = \{P \in E^1(\mathbb{Q}_p) \mid \frac{x(P)}{y(P)} \in p^n\mathbb{Z}_p\}.$$

Theorem 7.1. *The filtration $E(\mathbb{Q}_p) \supset E^0(\mathbb{Q}_p) \supset E^1(\mathbb{Q}_p) \supset \cdots \supset E^n(\mathbb{Q}_p) \supset \cdots$ has the following properties:*

(a) *the quotient $E(\mathbb{Q}_p)/E^0(\mathbb{Q}_p)$ is finite;*
(b) *the map $P \mapsto \bar{P}$ defines an isomorphism $E^0(\mathbb{Q}_p)/E^1(\mathbb{Q}_p) \to \bar{E}(\mathbb{F}_p)$;*
(c) *for $n \geq 1$, $E^n(\mathbb{Q}_p)$ is a subgroup of $E(\mathbb{Q}_p)$, and the map $P \mapsto p^{-n}\frac{x(P)}{y(P)} \mod p$ is an isomorphism $E^n(\mathbb{Q}_p)/E^{n+1}(\mathbb{Q}_p) \to \mathbb{F}_p$;*
(d) *the filtration is exhaustive, i.e., $\cap_n E^n(\mathbb{Q}_p) = \{0\}$.*

Proof. (a) We prove that $E(\mathbb{Q}_p)$ has a natural topology with respect to which it is compact and $E^0(\mathbb{Q}_p)$ is an open subgroup. Since $E(\mathbb{Q}_p)$ is a union of the cosets of $E^0(\mathbb{Q}_p)$, it will follow that there can only be finitely many of them.

Endow $\mathbb{Q}_p \times \mathbb{Q}_p \times \mathbb{Q}_p$ with the product topology, $\mathbb{Q}_p^3 \setminus \{(0,0,0)\}$ with the subspace topology, and $\mathbb{P}^2(\mathbb{Q}_p)$ with the quotient topology via $\mathbb{Q}_p^3 \setminus \{(0,0,0)\} \to \mathbb{P}^2(\mathbb{Q}_p)$. Then $\mathbb{P}^2(\mathbb{Q}_p)$ is the union of the images of the sets $\mathbb{Z}_p^\times \times \mathbb{Z}_p \times \mathbb{Z}_p$, $\mathbb{Z}_p \times \mathbb{Z}_p^\times \times \mathbb{Z}_p$, $\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p^\times$, each of which is compact and open. Therefore $\mathbb{P}^2(\mathbb{Q}_p)$ is compact. Its subset $E(\mathbb{Q}_p)$ is closed, because it is the zero set of a polynomial. Relative to this topology on $\mathbb{P}^2(\mathbb{Q}_p)$ two points that are close will have the same reduction modulo p. Therefore $E^0(\mathbb{Q}_p)$ is the intersection of $E(\mathbb{Q}_p)$ with an open subset of $\mathbb{P}^2(\mathbb{Q}_p)$.

(b) Hensel's lemma implies that the reduction map $E^0(\mathbb{Q}_p) \to \bar{E}(\mathbb{F}_p)$ is surjective, and we defined $E^1(\mathbb{Q}_p)$ to be its kernel.

(c) We assume (inductively) that $E^n(\mathbb{Q}_p)$ is a subgroup of $E(\mathbb{Q}_p)$. If $P = (x : y : 1)$ lies in $E^1(\mathbb{Q}_p)$, then $y \notin \mathbb{Z}_p$. Set $x = p^{-m}x_0$ and $y = p^{-m'}y_0$ with x_0 and y_0 units in $\mathbb{Z}_p$. Then

$$p^{-2m'}y_0^2 = p^{-3m}x_0^3 + ap^{-m}x_0 + b.$$

On taking ord_p of the two sides, we find that $-2m' = -3m$. Since m' and m are integers, this implies that there is an integer n such $m = 2n$ and $m' = 3n$; in fact, $n = \mathrm{ord}_p(\frac{x}{y})$.

The above discussion shows that if $P = (x : y : z) \in E^n(\mathbb{Q}_p) \setminus E^{n+1}(\mathbb{Q}_p)$, $n \geq 1$, then

$$\begin{cases} \mathrm{ord}_p(x) &= \mathrm{ord}_p(z) - 2n \\ \mathrm{ord}_p(y) &= \mathrm{ord}_p(z) - 3n. \end{cases}$$

Hence P can be expressed $P = (p^n x_0 : y_0 : p^{3n} z_0)$ with $\mathrm{ord}_p(y_0) = 0$ and $x_0, z_0 \in \mathbb{Z}_p$. In fact, this is true for all $P \in E^n(\mathbb{Q}_p)$. Since P lies on E,

$$p^{3n}y_0^2 z_0 = p^{3n}x_0^3 + ap^{7n}x_0 z_0^2 + bp^{9n}z_0^3,$$

and so $P_0 =_{df} (\bar{x}_0 : \bar{y}_0 : \bar{z}_0)$ lies on the curve

$$E_0 : Y^2 Z = X^3.$$

As $\bar{y}_0 \neq 0$, P_0 is not the singular point of E_0. From the description of the group laws in terms of chords and tangents, we see that the map

$$P \mapsto P_0 : E^n(\mathbb{Q}_p) \to E_0(\mathbb{F}_p)$$

is a homomorphism. Its kernel is $E^{n+1}(\mathbb{Q}_p)$, which is therefore a subgroup, and it follows from Hensel's lemma that its image is the set of nonsingular points of $E_0(\mathbb{F}_p)$. We know (see p27) that $Q \mapsto \frac{x(Q)}{y(Q)}$ is an isomorphism $E_0(\mathbb{F}_p) \setminus \{\text{singularity}\} \to \mathbb{F}_p$. The composite $P \mapsto P_0 \mapsto \frac{x(P_0)}{y(P_0)}$ is $P \mapsto \frac{p^{-n}x(P)}{y(P)} \mod p$.

(d) If $P \in \cap E^n(\mathbb{Q}_p)$, then $x(P) = 0$, $y(P) \neq 0$. This implies that either $z(P) = 0$ or $Y^2 = bZ^2$, but the second equation would contradict $P \in E^1(\mathbb{Q}_p)$. Hence $z(P) = 0$ and $P = (0 : 1 : 0)$. □

Remark 7.2. In the above, $\mathbb{Q}_p$ can be replaced with any local field.

Remark 7.3. It is possible to say much more about the structure of $E(\mathbb{Q}_p)$. A *one-parameter commutative formal group* over a (commutative) ring R is a power series $F(X, Y) \in R[[X, Y]]$ satisfying the following conditions:

(a) $F(X, Y) = X + Y + \text{terms of degree } \geq 2$;
(b) $F(X, F(Y, Z)) = F(F(X, Y), Z)$;
(c) $F(X, Y) = F(Y, X)$;
(d) there is a unique power series $i(T) \in R[[T]]$ such that $F(T, i(T)) = 0$.
(e) $F(X, 0) = X$ and $F(0, Y) = Y$.

In fact, (a) and (b) imply (d) and (e). If F is such a formal group over $\mathbb{Z}_p$, then the series $F(a,b)$ converges for $a, b \in p\mathbb{Z}_p$, and so F makes $p\mathbb{Z}_p$ into a group. One can show ([S1] Chapter IV) that an elliptic curve E over $\mathbb{Q}_p$ defines a formal group F over $\mathbb{Z}_p$, and that there are power series $x(T)$ and $y(T)$ such that $t \mapsto (x(t) : y(t) : 1)$ is an isomorphism of $p\mathbb{Z}_p$ (endowed with the group structure provided by F) onto $E^1(\mathbb{Q}_p)$. This is useful because it allows us to derive results about elliptic curves from results about formal groups, which are generally easier to prove.

An algorithm to compute intersection numbers. For $f(X,Y), g(X,Y) \in k[X,Y]$, set $I(f,g) = I(\text{origin}, \{f=0\} \cap \{g=0\})$. We explain how to compute $I(f,g)$ using only the following properties of the symbol: $I(X,Y) = 1$; $I(f,g) = I(g,f)$; $I(f,gh) = I(f,g) + I(f,h)$; $I(f, g+hf) = I(f,g)$ for all h; $I(f,g) = 0 \quad$ if $g(0,0) \neq 0$. Regard $f(X,Y)$ and $g(X,Y)$ as elements of $k[X][Y]$. The theory of resultants allows us to construct polynomials $a(X,Y)$ and $b(X,Y)$ such that $af + bg = r(X)$ with $r(X) \in k[X]$ and $\deg_Y(b) < \deg_Y(f)$, $\deg_Y(a) < \deg_Y(g)$. Now

$$I(f,g) = I(f,bg) - I(f,b) = I(f,r) - I(f,b).$$

Continue in this fashion until Y is eliminated from one of the polynomials, say, from g, so that $g = g(X) \in k[X]$. Write $g(X) = X^m g_0(X)$ where $g_0(0) \neq 0$. Then

$$I(f,g) = mI(f,X).$$

After subtracting a multiple of X from $f(X,Y)$, we can assume that it is a polynomial in Y. Write $f(Y) = Y^n f_0(Y)$ where $f_0(0) \neq 0$. Then

$$I(f,X) = n.$$

This algorithm is practical on a computer, but if the polynomials are monic when regarded as polynomials in Y, the following method is faster. If $\deg_Y(g) \geq \deg_Y(f)$, we can divide f into g (as polynomials in Y) and obtain

$$g = fh + r, \quad \deg_Y r < \deg_Y f \text{ or } r = 0.$$

Moreover,

$$I(f,g) = I(f,r).$$

Continue in this fashion until one of the polynomials has degree 1 in Y, and apply the following lemma.

Lemma 7.4. *If $f(0) = 0$, then $I(Y - f(X), g(X,Y)) = m$ where X^m is the power of X dividing $g(X, f(X))$.*

Proof. We divide $Y - f(X)$ into $g(X,Y)$ (as polynomials in Y) to obtain

$$g(X,Y) = (Y - f(X))h(X,Y) + g(X, f(X)),$$

from which it follows that

$$I(Y - f(X), g(X,Y)) = I(Y - f(X), g(X, f(X)) = mI(Y - f(X), X).$$

Finally, since we are assuming $f(0) = 0$, $f(X) = Xh(X)$, and so

$$I(Y - f(X), X) = I(Y, X) = 1.$$

□

8. Torsion Points

Throughout this section, E will be the elliptic curve

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in \mathbb{Z}, \quad \Delta = 4a^3 + 27b^2 \neq 0,$$

except that in second half of the section, we allow $a, b \in \mathbb{Z}_p$.

Theorem 8.1 (Lutz-Nagell). *If $P = (x : y : 1) \in E(\mathbb{Q})_{\text{tors}}$, then $x, y \in \mathbb{Z}$ and either $y = 0$ or $y|\Delta$.*

Remark 8.2. (a) The theorem provides an algorithm for finding all the torsion points on E: for each pair $(x, y) \in \mathbb{Z}$ with $y = 0$ or $y|\Delta$, check to see whether $(x : y : 1)$ is on E and whether it is a torsion point. It is not essential, but it helps, if the equation of E is chosen so that Δ is minimal among those with integer coefficients.

(b) The converse of the theorem is not true: a point $P = (x : y : 1) \in E(\mathbb{Q})$ can satisfy the conditions in the theorem without being a torsion point.

(c) The theorem can often be used to prove that a point $P \in E(\mathbb{Q})$ is of infinite order: compute multiples nP of P until you arrive at one whose coordinates are not integers, or better, just compute the x-coordinates of $2P$, $4P$, $8P$, using the duplication formula (see the end of this section).

The theorem will follow from the next two results: the first says that if P and $2P$ have integer coordinates (when we set $z = 1$), then either $y = 0$ or $y|\Delta$; the second implies that torsion points (hence also their multiples) have integer coordinates.

Lemma 8.3. *Let $P = (x_1 : y_1 : 1) \in E(\mathbb{Q})$. If P and $2P$ have integer coordinates (when we set $z = 1$), then either $y_1 = 0$ or $y_1|\Delta$.*

Proof. Assume $y_1 \neq 0$, and set $2P = (x_2 : y_2 : 1)$. Then $2P$ is the second point of intersection of the tangent at P to the affine curve

$$Y^2 = f(X), \quad f(X) = X^3 + aX + b.$$

The tangent line at P is

$$Y = \alpha X + \beta, \text{ where } \alpha = \left(\frac{dY}{dX}\right)_P = \frac{f'(x_1)}{2y_1}.$$

To find where this line intersects the affine curve, substitute for Y in the equation of the curve to obtain:

$$(\alpha X + \beta)^2 = X^3 + aX + b.$$

Thus the X-coordinates of the points of intersection are the roots of the cubic:

$$X^3 + aX + b - (\alpha X + \beta)^2 = X^3 - \alpha^2 X^2 + \cdots .$$

But we know that the X-coordinates of these points are x_1, x_1, x_2, and so

$$x_1 + x_1 + x_2 = \alpha^2.$$

Since x_1 and x_2 are integers, so also are α^2 and $\alpha = \frac{f'(x_1)}{2y_1}$. Thus $y_1|f'(x_1)$, and directly from the equation $y_1^2 = f(x_1)$ we see that $y_1|f(x_1)$. Hence y_1 divides both $f(x_1)$ and $f'(x_1)$. The theory of resultants (see $[C_2]$, Chapter on Remedial Mathematics) shows that

$$\Delta = r(X)f(X) + s(X)f'(X), \quad r(X), s(X) \in \mathbb{Z}[X],$$

and so this implies that $y_1|\Delta$. [In our case, $r(X) = -27(X^3 + aX - b)$ and $s(X) = (3X^2 + 4a)(3X^2 + a)$.] □

Proposition 8.4. *The group $E^1(\mathbb{Q}_p)$ is torsion-free.*

Before proving the proposition, we derive some consequences.

Corollary 8.5. *If $P = (x : y : 1) \in E(\mathbb{Q}_p)_{\mathrm{tors}}$, then $x, y \in \mathbb{Z}_p$.*

Proof. Recall that $\bar{P}$ is obtained from P by choosing primitive coordinates $(x : y : z)$ for P (i.e., coordinates such that $x, y, z \in \mathbb{Z}_p$ but not all of $x, y, z \in p\mathbb{Z}_p$), and setting $\bar{P} = (\bar{x} : \bar{y} : \bar{z})$, and that $E^1(\mathbb{Q}_p) = \{P \in E(\mathbb{Q}_p) \mid \bar{P} = (0 : 1 : 0)\}$. If $P = (x : y : 1)$ with x or y not in $\mathbb{Z}_p$, then any primitive coordinates $(x' : y' : z')$ for P will have $z' \in p\mathbb{Z}_p$. Hence $z(\bar{P}) = 0$, which implies $\bar{P} = (0 : 1 : 0)$, and so $P \in E^1(\mathbb{Q}_p)$. We have proved (the contrapositive of) the statement:

$$\text{if } P = (x : y : 1) \notin E^1(\mathbb{Q}_p), \text{ then } x, y \in \mathbb{Z}_p.$$

The proposition shows that if P is a nonzero torsion point, then $P \notin E^1(\mathbb{Q}_p)$. □

Corollary 8.6. *If $P = (x : y : 1) \in E(\mathbb{Q})_{\mathrm{tors}}$, then $x, y \in \mathbb{Z}$.*

Proof. This follows from the previous corollary, because if a rational number r is not an integer, then $\mathrm{ord}_p(r) < 0$ for some p, and so $r \notin \mathbb{Z}_p$. □

Corollary 8.7. *If E has good reduction at p (i.e., $p \neq 2$ and p does not divide Δ), then the reduction map*

$$E(\mathbb{Q})_{\mathrm{tors}} \to \bar{E}(\mathbb{F}_p)$$

is injective.

Proof. Because E has good reduction, $E^0(\mathbb{Q}_p) = E(\mathbb{Q}_p)$. The reduction map $E(\mathbb{Q}_p) \to \bar{E}(\mathbb{Q}_p)$ has kernel $E^1(\mathbb{Q}_p)$, which intersects $E(\mathbb{Q})_{\mathrm{tors}}$ in $\{O\}$. □

Remark 8.8. This puts a very serious restriction on the size of $E(\mathbb{Q})_{\mathrm{tors}}$. For example, if E has good reduction at 5, then, according to the Riemann hypotheses, $\bar{E}$ will have at most $5 + 1 + 2\sqrt{5}$ points with coordinates in $\mathbb{F}_5$, and so E will have at most 10 torsion points with coordinates in $\mathbb{Q}$.

We now prove Proposition 8.4. In one case this follows directly from the results of Section 7. Let $P \in E^1(\mathbb{Q}_p)$ be a torsion point of order m not divisible by p. If $P \neq 0$, then[9] $P \in E^n(\mathbb{Q}_p) \setminus E^{n+1}(\mathbb{Q}_p)$ for some n. But we have an isomorphism (of abelian groups)

$$P \mapsto p^{-n}\frac{x(P)}{y(P)} : E^n(\mathbb{Q}_p)/E^{n+1}(\mathbb{Q}_p) \xrightarrow{\approx} \mathbb{Z}/p\mathbb{Z}$$

(Theorem 7.1c). By assumption, the image of P under this map is nonzero, which implies that m times the image will also be nonzero. This contradicts the fact that $mP = 0$.

To prove the general case, where p may divide the order of P, we have to analyze the filtration more carefully.

For $P \in E^1(\mathbb{Q}_p)$, we have $y(P) \neq 0$, which suggests that we look at the affine curve $E \cap \{(x : y : z\} \mid y \neq 0\}$:

$$E_1 : Z = X^3 + aXZ^2 + bZ^3.$$

[9] "\" is "setminus", so this means $P \in E^n(\mathbb{Q}_p)$, $P \notin E^{n+1}(\mathbb{Q}_p)$.

A point $P = (x : y : z)$ on E has coordinates $x'(P) =_{df} \frac{x(P)}{y(P)}$, $z'(P) =_{df} \frac{z(P)}{y(P)}$ on E_1. For example, $O = (0 : 1 : 0)$ becomes the origin on E_1, and so $P \mapsto -P$ becomes reflection in the origin $(x', z') \mapsto (-x', -z')$. Just as on E, $P + Q + R = 0$ if and only if P, Q, R lie on a line.

In terms of our new picture,

$$E^n(\mathbb{Q}_p) = \{P \in E^1(\mathbb{Q}_p) \mid x'(P) \in p^n\mathbb{Z}_p\}.$$

Thus the $E^n(\mathbb{Q}_p)$'s form a fundamental system of neighbourhoods of the origin in $E_1(\mathbb{Q}_p)$. The key lemma is following:

Lemma 8.9. *Let $P_1, P_2, P_3 \in E(\mathbb{Q}_p)$ be such that $P_1 + P_2 + P_3 = O$. If $P_1, P_2 \in E^n(\mathbb{Q}_p)$, then $P_3 \in E^n(\mathbb{Q}_p)$, and*

$$x'(P_1) + x'(P_2) + x'(P_3) \in p^{5n}\mathbb{Z}_p.$$

Before proving the lemma, we explain why it implies the proposition. For $P \in E^n(\mathbb{Q}_p)$, let $\bar{x}(P) = x'(P) \mod p^{5n}\mathbb{Z}_p$. The lemma shows that the map

$$P \mapsto \bar{x}(P) : E^n(\mathbb{Q}_p) \to p^n\mathbb{Z}_p/p^{5n}\mathbb{Z}_p$$

has the property:

$$P_1 + P_2 + P_3 = 0 \implies \bar{x}(P_1) + \bar{x}(P_2) + \bar{x}(P_3) = 0.$$

Since $\bar{x}(-P) = -\bar{x}(P)$, it is therefore a homomorphism of abelian groups. Suppose that $P \in E^1(\mathbb{Q}_p)$ has order m divisible by p. Then $Q =_{df} \frac{m}{p}P$ will also lie in $E^1(\mathbb{Q}_p)$ and will have order p. Since $Q \neq 0$, for some n, $Q \in E^n(\mathbb{Q}_p) \setminus E^{n+1}(\mathbb{Q}_p)$. Then $\bar{x}(Q) \in p^n\mathbb{Z}_p \setminus p^{n+1}\mathbb{Z}_p \mod p^{5n}\mathbb{Z}_p$, and so

$$\bar{x}(pQ) = p\bar{x}(Q) \in p^{n+1}\mathbb{Z}_p \setminus p^{n+2}\mathbb{Z}_p \mod p^{5n}\mathbb{Z}_p.$$

This contradicts the fact that $pQ = 0$.

We now prove the lemma. We saw in Section 7 that if $P = (x : y : 1) \in E^n(\mathbb{Q}_p) \setminus E^{n+1}(\mathbb{Q}_p)$, then $\text{ord}_p(x) = -2n$, $\text{ord}_p(y) = -3n$. In terms of homogeneous coordinates $P = (x : y : z)$, this means that

$$P \in E^n(\mathbb{Q}_p) \setminus E^{n+1}(\mathbb{Q}_p) \implies \left\{ \begin{array}{lcl} \text{ord}_p \frac{x(P)}{z(P)} & = & -2n \\ \text{ord}_p \frac{y(P)}{z(P)} & = & -3n \end{array} \right. \implies \left\{ \begin{array}{lcl} \text{ord}_p \frac{x(P)}{y(P)} & = & n \\ \text{ord}_p \frac{z(P)}{y(P)} & = & 3n \end{array} \right. .$$

Thus

$$P \in E^n(\mathbb{Q}_p) \implies x'(P) \in p^n\mathbb{Z}_p, \quad z'(P) \in p^{3n}\mathbb{Z}_p.$$

Let $P_i = (x'_i, z'_i)$, $i = 1, 2, 3$. The line through P_1, P_2 (assumed distinct) is $Z = \alpha X + \beta$ where

$$\alpha = \frac{z'_2 - z'_1}{x'_2 - x'_1} = \ldots = \frac{{x'_2}^2 + x'_1 x'_2 + {x'_1}^2 + a{z'_2}^2}{1 - a x'_1(z'_2 + z'_1) - b({z'_2}^2 + z'_1 z_2 + {z'_1}^2)} \in p^{2n}\mathbb{Z}_p.$$

Moreover

$$\beta = z'_1 - \alpha x'_1 \in p^{3n}\mathbb{Z}_p.$$

On substituting $\alpha X + \beta$ for Z in the equation for E_1, we obtain the equation

$$\alpha X + \beta = X^3 + aX(\alpha X + \beta)^2 + b(\alpha X + \beta)^3.$$

We know that the solutions of this equation are x_1', x_2', x_3', and so

$$x_1' + x_2' + x_3' = \frac{2a\alpha\beta + 3b\alpha^2\beta}{1 + a\alpha^2 + b\alpha^3} \in p^{5n}\mathbb{Z}_p.$$

The proof when $P_1 = P_2$ is similar. For full details, including the elementary calculation omitted for α, see [ST] p50–54.

Remark 8.10. When $\mathbb{Q}$ is replaced by a number field K, the above argument may fail to show that torsion elements of $E(K)$ have coordinates that are algebraic integers (when z is taken to be 1). Let π be a prime element in K_v. The same argument as above shows that there is an isomorphism

$$E^n(K_v)/E^{5n}(K_v) \to \pi^n\mathcal{O}_v/\pi^{5n}\mathcal{O}_v.$$

However, if p is a high power of π (i.e., the extension $K/\mathbb{Q}$ is highly ramified v) and n is small, this no longer excludes the possibility that $E^n(K_v)$ may contain an element of order p.

Formulas. We give formulas for the addition and doubling of points on the curve

$$E : Y^2 = X^3 + aX + b, \quad a, b \in k \quad \Delta = 4a^3 + 27b^2 \neq 0.$$

As above, the strategy for deriving the formulas is to first find the x-coordinate of the point sought by using that the sum of the roots of a polynomial $f(X)$ is $-$(coefficient of $X^{\deg f - 1}$).

Addition formula. Let $P = (x, y)$ be the sum of $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$. If $P_2 = -P_1$, then $P = O$, and if $P_1 = P_2$, we can apply the duplication formula. Otherwise, $x_1 \neq x_2$, and (x, y) is determined by the following formulas:

$$x(x_1 - x_2)^2 = x_1x_2^2 + x_1^2x_2 - 2y_1y_2 + a(x_1 + x_2) + 2b$$

and

$$y(x_1 - x_2)^3 = W_2y_2 - W_1y_1$$

where

$$\begin{aligned} W_1 &= 3x_1x_2^2 + x_2^3 + a(x_1 + 3x_2) + 4b \\ W_2 &= 3x_1^2x_2 + x_1^3 + a(3x_1 + x_2) + 4b. \end{aligned}$$

Duplication formula. Let $P = (x, y)$ and $2P = (x_2, y_2)$. If $y = 0$, then $2P = 0$. Otherwise $y \neq 0$, and (x_2, y_2) is determined by the following formulas:

$$\begin{aligned} x_2 &= \frac{(3x^2 + a)^2 - 8xy^2}{4y^2} = \frac{x^4 - 2ax^2 - 8bx + a^2}{4(x^3 + ax + b)} \\ y_2 &= \frac{x^6 + 5ax^4 + 20bx^3 - 5a^2x^2 - 4abx - a^3 - 8b^2}{(2y)^3}. \end{aligned}$$

Exercise 8.11. For four of the following elliptic curves (including at least one of the last four), compute the torsion subgroups of $E(\mathbb{Q})$. (Include only enough details to convince the grader that you really did work it out.)

$$\begin{array}{rcl}
Y^2 &=& X^3+2\\
Y^2 &=& X^3+X\\
Y^2 &=& X^3+4\\
Y^2 &=& X^3+4X\\
Y^2+Y &=& X^3-X^2\\
Y^2 &=& X^3+1\\
Y^2-XY+2Y &=& X^3+2X^2\\
Y^2+7XY-6Y &=& X^3-6X^2\\
Y^2+3XY+6Y &=& X^3+6X^2\\
Y^2-7XY-36Y &=& X^3-18X^2\\
Y^2+43XY-210Y &=& X^3-210X^2\\
Y^2 &=& X^3-X\\
Y^2 &=& X^3+5X^2+4X\\
Y^2+5XY-6Y &=& X^3-3X^2\\
Y^2 &=& X^3+337X^2+20736X
\end{array}$$

Solution to Exercise 4.8. No vertical line is an inflectional tangent, and so we may assume $c \neq 0$. The line $L : Y = cX + d$ intersects the curve at the points whose X-coordinates satisfy

$$(cX+d)^2 = X^3 + aX + b.$$

By Bezout's theorem, L be an inflectional tangent to E if and only if it meets the projective curve in a single point. This will be so if and only if

$$X^3 - c^2X^2 + (a-2cd)X + b - d^2$$

has a triple root (which will automatically lie in $\mathbb{Q}$). Hence there must exist an $r \in \mathbb{Q}$ such that

$$-3r = -c^2, \quad 3r^2 = a - 2cd, \quad -r^3 = b - d^2.$$

When we use the first equation to eliminate r from the remaining two, we find that

$$a = 2cd + \frac{c^4}{3}, \quad b = d^2 - \frac{c^6}{27}.$$

Conversely, if these equations hold, then $r = c^2/3$ is a triple root of the above polynomial, and so L is an inflectional tangent.

Note that $3P = 0$ if and only if $2P = -P$, i.e., if and only if the tangent line at P is an inflectional tangent. Only the line $L_\infty : Z = 0$ is an inflectional tangent at O. Thus E will have a rational point of order 3 if and only if a, b can be expressed as above in terms of two rational numbers c, d. Therefore the general form of an elliptic curve having a rational point of order 3 is

$$Y^2Z = X^3 + (2cd + \frac{c^4}{3})X + (d^2 - \frac{c^6}{27}), \quad 4(2cd + \frac{c^4}{3})^3 + 27(d^2 - \frac{c^6}{27}) \neq 0.$$

9. Néron Models

Consider an elliptic curve over $\mathbb{Q}_p$

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in \mathbb{Q}_p, \quad \Delta = 4a^3 + 27b^2 \neq 0.$$

After making a change of variables $X \mapsto X/c^2$, $Y \mapsto Y/c^3$, $Z \mapsto Z$, we can suppose that $a, b \in \mathbb{Z}_p$ and $\mathrm{ord}_p(\Delta)$ is minimal. We can think of E as defining a curve over $\mathbb{Z}_p$, which will be the best "model" of E over $\mathbb{Z}_p$ among plane projective curves when $p \neq 2, 3$. However, when $p = 2$ or 3 we may be able to get a better model of E over $\mathbb{Z}_p$ by allowing a more complicated equation. Moreover, Néron showed that if we allow our models to be curves over $\mathbb{Z}_p$ that are not embeddable in $\mathbb{P}^2$, then we obtain models that are better in some senses than any plane model. I'll attempt to explain what these Néron models are in this section. Unfortunately, this is a difficult topic, which requires the theory of schemes for a satisfactory explanation[10] and so I'll have to be very superficial. The only good treatment of Néron models is in Chapter IV of [S2].

Weierstrass minimal models. As we noted in (1.3), a curve of the form

$$Y^2 = X^3 + aX + b$$

is always singular in characteristic 2. However, the curve

$$Y^2 + Y = X^3 - X^2 - 10X - 20$$

has good reduction at 2 (and, in fact, at all primes except 11). In general we should allow equations for E of the form

$$E : Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3,$$

and changes of variables of the form

$$\begin{aligned} X &= u^2X' + r \\ Y &= u^3Y' + su^2X' + t \end{aligned}$$

with $u, r, s, t \in \mathbb{Q}_p$ and $u \neq 0$. One can attach to such a curve a discriminant $\Delta(a_1, a_2, a_3, a_4a_5)$, which is a complicated polynomial in the a_i's, and which is zero if and only if E is singular. Moreover, one can choose a change of variables which makes the $a_i \in \mathbb{Z}_p$ and is such that $\mathrm{ord}_p(\Delta)$ is minimal. The equation (or rather the curve it defines over $\mathbb{Z}_p$) is called the *Weierstrass minimal model* of E. If $p \neq 2, 3$, this agrees with the model defined in the first paragraph above.

The work of Kodaira. Before considering Néron models, we look at an analogous situation, which was a precursor.

Consider an equation

$$Y^2Z = X^3 + a(T)XZ^2 + b(T)Z^3, \quad a(T), b(T) \in \mathbb{C}[T], \quad \Delta(T) = 4a(T)^3 + 27b(T)^2 \neq 0.$$

We can view this in three different ways:

(a) as defining an elliptic curve E over the field $\mathbb{C}(T)$;
(b) as defining a surface S in $\mathbb{A}^1(\mathbb{C}) \times \mathbb{P}^2(\mathbb{C})$;
(c) as defining a family of (possibly degenerate) elliptic curves $E(T)$ parametrized by T.

[10] Néron himself didn't use schemes, but rather invented his own private version of algebraic geometry over discrete valuation rings, which makes his papers almost unreadable.

By (c) we mean the following: for each $t_0 \in \mathbb{C}$ we have a curve

$$E(t_0) : Y^2Z = X^3 + a(t_0)XZ^2 + b(t_0)Z^3, \quad a(t_0), b(t_0) \in \mathbb{C},$$

with discriminate $\Delta(t_0)$. This is nonsingular, and hence an elliptic curve, if and only if $\Delta(t_0) \neq 0$. Otherwise, it will have a singularity, and we view it as a degenerate elliptic curve. Note that the projection map $\mathbb{A}^1(\mathbb{C}) \times \mathbb{P}^2(\mathbb{C}) \to \mathbb{A}^1(\mathbb{C})$ induces a map $S \to \mathbb{A}^1(\mathbb{C})$ whose fibres are the curves $E(t)$. We can view S as a "model" of E over $\mathbb{C}[T]$ (or over $\mathbb{A}^1(\mathbb{C})$). We should choose the equation of E so that $\Delta(T)$ has minimum degree and there are as few singular fibres as possible.

For the sake of simplicity, we now drop the Z, and consider the equation

$$Y^2 = X^3 + a(T)X + b(T), \quad a(T), b(T) \in \mathbb{C}[T],$$

—strictly, we should work with the family of projective curves.

Let $P = (x, y, t) \in S(\mathbb{C})$, and let $f(X, Y, T) = X^3 + a(T)X + b(T) - Y^2$. Then P is singular on $E(t)$ if and only if it satisfies the following equations:

$$\begin{aligned} \tfrac{\partial f}{\partial Y} &= -2Y = 0 \\ \tfrac{\partial f}{\partial X} &= 3X^2 + a(T) = 0. \end{aligned}$$

It is singular in S if on addition it satisfies the equation

$$\frac{\partial f}{\partial T} = a'(T)X + b'(T) = 0.$$

Thus, if P is singular in S, then it is singular in its fibre $E(t)$, but the converse is need not be true.

Example 9.1. (a) Consider the equation

$$Y^2 = X^3 - T.$$

The origin is singular (in fact, it is a cusp) when regarded as a point on $E(0) : Y^2 = X^3$, but not when regarded as a point on $S : Y^2 = X^3 - T$. In fact, the tangent plane to S at the origin is the X, Y-plane, $T = 0$.

(b) Consider the equation

$$Y^2 = X^3 - T^2.$$

In this case, the origin is singular when regarded as a point on $E(0)$ *and* when regarded as a point on S.

(c) Consider the equation

$$\begin{aligned} Y^2 &= (X - 1 + T)(X - 1 - T)(X + 2) \\ &= X^3 - (3 + T^2)X + 2 - 2T^2. \end{aligned}$$

The discriminant is

$$\Delta(T) = -324T^2 + 72T^4 - 4T^6.$$

The curve $E(0)$ is

$$Y^2 = X^3 - 3X + 2 = (X - 1)^2(X + 2),$$

which has a node at $(1,0)$. Replace $X-1$ in the original equation with X in order to translate $(1,0,0)$ to the origin. The equation becomes

$$\begin{aligned} Y^2 &= (X+T)(X-T)(X+3) \\ &= (X^2-T^2)(X+3) \\ &= X^3+3X^2-T^2X-3T^2. \end{aligned}$$

This has surface has a singularity at the origin (because the equation has no linear term).

Kodaira showed (Collected Works [51], [52], 1960) that, by blowing up points, and blowing down curves, etc., it is possible to obtain from the surface

$$S: Y^2Z = X^3 + a(T)XZ^2 + b(T)Z^3, \quad a(T), b(T) \in \mathbb{C}[T], \quad \Delta[T] \neq 0$$

a new surface S' endowed with a map $S' \to \mathbb{A}^1$ having the following properties:

(a) S' is nonsingular;
(b) S' regarded as a curve over $\mathbb{C}(T)$ is equal to S regarded as a curve over $\mathbb{C}(T)$ (for the experts, the maps $S \to \mathbb{A}^1$ and $S' \to \mathbb{A}^1$ have the same generic fibres);
(c) the fibres $E'(t_0)$ of S' over $\mathbb{A}^1(\mathbb{C})$ are all projective curves; moreover $E'(t_0) = E(t_0)$ if the points of $E(t_0)$ are nonsingular when regarded as points on S (for example, if $E(t_0)$ itself is nonsingular);
(d) S' has a certain minimality property: if S'' is a second surface with the above properties, then any regular map $S' \to S''$ is an isomorphism.

Moreover, Kodaira showed that S is unique, and he classified the possible fibres of $S' \to \mathbb{A}^1$.

"Blowing up" a point P in a variety V leaves the variety unchanged except that it replaces the point P with the projective space of lines through the origin in the tangent space $Tgt_P(V)$ to V at P. A curve C in V, when regarded as a point in the blown-up variety, meets the projective space at the point corresponding to the tangent line to the curve. Even when $V \subset \mathbb{P}^m$, the blown-up variety doesn't have a natural embedding into a projective space.

Example 9.2. To illustrate the phenomenon of "blowing up", consider the map

$$\sigma : k^2 \to k^2, \quad (x,y) \mapsto (x, xy).$$

Its image omits only the points on the Y-axis where $Y \neq 0$. A point in the image is the image of a unique point in k^2 except for $(0,0)$, which is the image of the whole of the Y-axis. Thus the map is one-to-one, except that the Y-axis has been "blown down" to a point.

The line

$$C: Y = \alpha X$$

has inverse image equal to the union of the Y-axis and the line $Y = \alpha$. The curve

$$Y^2 = X^3 + \alpha X^2$$

has as inverse image the union of the Y-axis and a nonsingular curve that meets the Y-axis at the points $(0, \pm\sqrt{\alpha})$, i.e., at the same points that its tangents do.

In the above map, $(0,0)$ in $\mathbb{A}^2(k)$ was blown up to an affine line. In a true blowing-up, it would be replaced by a projective line, and the description of the map would be more complicated.

The complete Néron model. Néron proved an analogue of Kodaira's result for elliptic curves over $\mathbb{Q}_p$. To explain his result, we need to talk about schemes. For the nonexperts, a *scheme* $\mathcal{E}$ over $\mathbb{Z}_p$ is simply the object defined by a collection of polynomial equations with coefficents in $\mathbb{Z}_p$. The object defined by the same equations regarded as having coefficients in $\mathbb{Q}_p$ is a variety E over $\mathbb{Q}_p$ called the *generic fibre* of $\mathcal{E}/\mathbb{Z}_p$, and the object defined by the equations with the coefficients reduced modulo p is a variety $\bar{E}$ over $\mathbb{F}_p$ called the *special fibre* of $\mathcal{E}/\mathbb{Z}_p$. For example, if $\mathcal{E}$ is the scheme defined by the equation

$$Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3, \quad a_i \in \mathbb{Z}_p,$$

then E is the elliptic curve over $\mathbb{Q}_p$ defined by the same equation, and $\bar{E}$ is the elliptic curve over $\mathbb{F}_p$

$$Y^2Z + \bar{a}_1XYZ + \bar{a}_3YZ^2 = X^3 + \bar{a}_2X^2Z + \bar{a}_4XZ^2 + \bar{a}_6Z^3, \quad \bar{a}_i \in \mathbb{F}_p.$$

Given an elliptic curve $E/\mathbb{Q}_p$, Néron constructs a scheme $\mathcal{E}$ over $\mathbb{Z}_p$ having the following properties:

(a) $\mathcal{E}$ is a regular scheme; this means that all the local rings associated with $\mathcal{E}$ are regular local rings (for a variety over an algebraically closed field, this condition is equivalent to the variety being nonsingular);
(b) the generic fibre of $\mathcal{E}$ is the original curve E;
(c) $\mathcal{E}$ is proper over $\mathbb{Z}_p$; this simply means that both E and $\bar{E}$ are complete curves (this is a compactness condition: affine curves aren't complete; projective curves are).
(d) $\mathcal{E}$ has a certain minimality property sufficient to determine it uniquely: if $\mathcal{E}'$ is a second scheme over $\mathbb{Z}_p$ having the properties (a), (b), (c), then any regular map $\mathcal{E} \to \mathcal{E}'$ is an isomorphism.

Moreover, Néron classified the possible special fibres, and obtained essentially the same list as Kodaira.

The complete Néron model has some defects: unlike the Weierstrass minimal model, not every point in $E(\mathbb{Q}_p)$ need extend to a point in $\mathcal{E}(\mathbb{Z}_p)$; it doesn't have a group structure; it's special fibre $\bar{E}$ may be singular. All three defects are eliminated by simply removing all singular points and multiple curves in the special fibre. One then obtains the *smooth Néron minimal model*, which however has the defect that it not complete.

Given an elliptic curve E over $\mathbb{Q}_p$ with now have three models over $\mathbb{Z}_p$:

(a) $\mathcal{E}^w$, the Weierstrass minimal model of E;
(b) $\mathcal{E}$, the complete Néron minimal model of E;
(c) $\mathcal{E}'$, the smooth Néron minimal model of E.

They are related as follows: to get $\mathcal{E}'$ from $\mathcal{E}$ remove all multiple curves and singular points; when we remove from $\mathcal{E}'$ all connected components of the special fibre except that containing O, we obtain the Weierstrass model with the singular point in the closed fibre removed.

Example 9.3. We describe three of the possible eleven different types of models. Some of the statements below are only valid when $p \neq 2, 3$. We describe the special fibre over $\mathbb{F}_p^{\mathrm{al}}$ rather than $\mathbb{F}_p$. For example, in (b), over $\mathbb{F}_p$ the zero component of G may be a twisted $\mathbb{G}_m$, and not all n points in the quotient G/G^0 need have coordinates in $\mathbb{F}_p$.

(a) For an elliptic curve E with good reduction, all three models are the same.

(b) For an elliptic curve E which has nodal reduction, and $\mathrm{ord}_p(\Delta) = n$, the special fibres for the three models are: (a) a cubic curve with a node; (b) n curves, each of genus 0, each intersecting exactly two other of the curves; (c) an algebraic group G such that the connected component G^0 of G containing zero is $\mathbb{G}_m$, and such that G/G^0 is a cyclic group of order n.

[[Diagram omitted]]

(c) For an elliptic curve E which has cuspidal reduction and $\mathrm{ord}_p(\Delta) = 5$, the special fibres for the three models are: (a) a cubic curve with a cusp; (b) five curves of genus 0, one with multiplicity 2, intersecting as below; (b) an algebraic group G whose zero component is $\mathbb{G}_a$ and such that G/G^0 is a group of order 4 killed by 2.

[[Diagram omitted.]]

Finally, the mysterious quotient $E(\mathbb{Q}_p)/E^0(\mathbb{Q}_p)$ is equal to $G(\mathbb{F}_p)/G^0(\mathbb{F}_p)$ where G is the special fibre of the smooth Néron model and G^0 is its zero component. In the above three examples, it is (a) the trivial group; (b) a subgroup of a cyclic group of order n (and equal to a cyclic group of order n if E has split nodal reduction); (c) a subgroup of $(\mathbb{Z}/2\mathbb{Z})^2$.

Summary. [[The top three E's are $\bar{E}$'s]]

Minimal Model	Weierstrass	complete Néron	smooth Néron
Plane curve	Yes	Not always	Not always
Regular?	Not always	Yes	Yes
$\overline{E}$ complete?	Yes	Yes	Not always
$\overline{E}$ nonsingular?	Not always	Not always	Yes
$\overline{E}$ a group?	Not always	Not always	Yes
$E(\mathbb{Z}_p) = E(\mathbb{Q}_p)$?	Yes	Not always	Yes

Tate has given an algorithm for determining the Néron model of an elliptic curve.

10. Elliptic Curves over the Complex Numbers

In this section, we review some of the theory of elliptic curves over $\mathbb{C}$.

Lattices and bases. A *lattice* in $\mathbb{C}$ is the subgroup generated by two complex numbers linearly independent over $\mathbb{R}$: thus

$$\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2.$$

Since neither ω_1 nor ω_2 is a real multiple of the other, we can order them so that $\Im(\omega_1/\omega_2) > 0$. If $\{\omega_1', \omega_2'\}$ is a second pair of elements of Λ, then

$$\omega_1' = a\omega_1 + b\omega_2, \quad \omega_2' = c\omega_1 + d\omega_2, \quad a, b, c, d \in \mathbb{Z},$$

that is,

$$\begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} = A \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix},$$

with A a 2×2 matrix with integer coefficients. The pair (ω_1', ω_2') will be a basis for Λ if and only if A has determinant ± 1, and $\Im(\omega_1'/\omega_2') > 0$ if and only if $\det A > 0$. Therefore, if we let $\mathrm{SL}_2(\mathbb{Z})$ be the group of matrices with integer coefficients and determinant 1, then $\mathrm{SL}_2(\mathbb{Z})$ acts transitively on the set of bases (ω_1, ω_2) for Λ for which $\Im(\omega_1/\omega_2) > 0$. We have proved the following statement:

Proposition 10.1. *Let M be the set of pairs of complex numbers (ω_1, ω_2) such that $\Im(\omega_1/\omega_2) > 0$, and let $\mathcal{L}$ be the set of lattices in $\mathbb{C}$. Then the map $(\omega_1, \omega_2) \mapsto \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ induces a bijection*

$$\mathrm{SL}_2(\mathbb{Z})\backslash M \to \mathcal{L}.$$

Here $\mathrm{SL}_2(\mathbb{Z})\backslash M$ means the set of orbits in M for the action

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} a\omega_1 + b\omega_2 \\ c\omega_1 + d\omega_2 \end{pmatrix}$$

Let $\mathbb{H}$ be the complex upper half-plane:

$$\mathbb{H} = \{z \in \mathbb{C} \mid \Im(z) > 0\}.$$

Let $z \in \mathbb{C}^\times$ act on M by the rule $z(\omega_1, \omega_2) = (z\omega_1, z\omega_2)$ and on $\mathcal{L}$ by the rule $z\Lambda = \{z\lambda \mid \lambda \in \Lambda\}$. The map $(\omega_1, \omega_2) \mapsto \omega_1/\omega_2$ induces a bijection $M/\mathbb{C}^\times \to \mathbb{H}$. The action of $\mathrm{SL}_2(\mathbb{Z})$ on M corresponds to the action

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tau = \frac{a\tau + b}{c\tau + d}$$

on $\mathbb{H}$. We have bijections

$$\begin{array}{ccccc} \mathcal{L}/\mathbb{C}^\times & \longleftrightarrow & \mathrm{SL}_2(\mathbb{Z})\backslash M/\mathbb{C}^\times & \longleftrightarrow & \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}. \\ \mathbb{Z}\tau + \mathbb{Z} & & (\tau, 1) & & \tau \end{array}$$

For a lattice Λ, the interior of any parallelogram with vertices z_0, $z_0+\omega_1$, $z_0+\omega_2$, $z_0+\omega_1+\omega_2$, where $\{\omega_1, \omega_2\}$ is a basis for Λ, is called a *fundamental domain* or *period parallelogram D* for Λ. We usually choose D to contain 0.

Quotients of $\mathbb{C}$ by lattices. . Let Λ be a lattice in $\mathbb{C}$. Topologically the quotient $\mathbb{C}/\Lambda \approx \mathbb{R}^2/\mathbb{Z}^2$, which is a one-holed torus (the surface of a donut).

Write $\pi : \mathbb{C} \to \mathbb{C}/\Lambda$ for the quotient map. Then $\mathbb{C}/\Lambda$ can be given the structure of a Riemann surface (i.e., complex manifold of dimension 1) such that a function $\varphi : U \to \mathbb{C}$ on an open subset U of $\mathbb{C}/\Lambda$ is holomorphic (resp. meromorphic) if and only if the composite $\varphi \circ \pi : \pi^{-1}(U) \to \mathbb{C}$ is holomorphic (resp. meromorphic) in the usual sense. It is the unique structure for which π is a local isomorphism of Riemann surfaces.

We shall see that, although any two quotients $\mathbb{C}/\Lambda$, $\mathbb{C}/\Lambda'$ are homeomorphic, they will be isomorphic as Riemann surfaces only if $\Lambda' = z\Lambda$ for some $z \in \mathbb{C}$.

Doubly periodic functions. Let Λ be a lattice in $\mathbb{C}$. According to the above discussion, a meromorphic function on $\mathbb{C}/\Lambda$ is simply a meromorphic function $f(z)$ on $\mathbb{C}$ such that

$$f(z + \omega) = f(z) \text{ for all } \omega \in \Lambda.$$

This condition is equivalent to

$$f(z + \omega_1) = f(z), \quad f(z + \omega_2) = f(z)$$

for $\{\omega_1, \omega_2\}$ a basis for Λ. Such a meromorphic function on $\mathbb{C}$ is said to be *doubly periodic* for Λ.

Proposition 10.2. *Let $f(z)$ be a doubly periodic function for Λ, not identically zero, and let D be a fundamental domain for Λ such that f has no zeros or poles on the boundary of D. Then*

(a) $\sum_{P\in D}\mathrm{Res}_P(f)=0$;
(b) $\sum_{P\in D}\mathrm{ord}_P(f)=0$;
(c) $\sum_{P\in D}\mathrm{ord}_P(f)\cdot P\equiv 0 \mod \Lambda$.

The first sum is over the points in D where f has a pole, and the other sums are over the points where it has a zero or pole (and $ord_P(f)$ is the order of the zero or the negative of the order of the pole). Each sum is finite.

Proof. According to the residue theorem,

$$\int_\Gamma f(z)dz=2\pi i(\sum_{P\in D}\mathrm{Res}_P(f)),$$

where Γ is the boundary of D. Because f is periodic, the integrals of it over opposite sides of D cancel, and so the integral is zero. This gives (a). For (b) one applies the residue theorem to f'/f, noting that this is again doubly periodic and that $\mathrm{Res}_P(f'/f)=\mathrm{ord}_P(f)$. For (c) one applies the residue theorem to $z\cdot f'(z)/f(z)$. This is no longer doubly periodic, but the integral of it around Γ lies in Λ. □

Corollary 10.3. *A nonconstant doubly periodic function has at least two poles.*

Proof. A holomorphic doubly periodic function is bounded on the closure of any fundamental domain (by compactness), and hence on the entire plane (by periodicity). It is constant by Liouville's theorem. It is impossible for a doubly periodic function to have a single simple pole in a period parallelogram, because by (a) of proposition the residue at the pole would have to be zero there, which contradicts the fact that it has a simple pole there. □

The holomorphic maps $\mathbb{C}/\Lambda\to\mathbb{C}/\Lambda'$. Let Λ and Λ' be lattices in $\mathbb{C}$. The map $\pi:\mathbb{C}\to\mathbb{C}/\Lambda$ realizes $\mathbb{C}$ as the universal covering space of $\mathbb{C}/\Lambda$. Since the same is true of $\pi':\mathbb{C}\to\mathbb{C}/\Lambda'$, a continous map $\varphi:\mathbb{C}/\Lambda\to\mathbb{C}/\Lambda'$ such that $\varphi(0)=0$ will lift uniquely to a continuous map $\widetilde{\varphi}:\mathbb{C}\to\mathbb{C}$ such that $\widetilde{\varphi}(0)=0$:

$$\begin{tikzcd}
\mathbb{C} \arrow[r, "\widetilde{\varphi}"] \arrow[d] & \mathbb{C} \arrow[d] \\
\mathbb{C}/\Lambda \arrow[r, "\varphi"] & \mathbb{C}/\Lambda'
\end{tikzcd}$$

(see, for example, Greenberg, Lectures on Algebraic Topology, 5.1, 6.4). The map φ will be holomorphic (i.e., a morphism of Riemann surfaces) if and only if $\widetilde{\varphi}$ is holomorphic. [Nonexperts can take this as a definition of a holomorphic map $\varphi:\mathbb{C}/\Lambda\to\mathbb{C}/\Lambda'$.]

Proposition 10.4. *Let Λ and Λ' be lattices in $\mathbb{C}$. A complex number α such that $\alpha\Lambda\subset\Lambda'$ defines a holomorphic map*

$$[z]\mapsto[\alpha z]:\mathbb{C}/\Lambda\to\mathbb{C}/\Lambda'$$

sending 0 to 0, and every holomorphic map $\mathbb{C}/\Lambda\to\mathbb{C}/\Lambda'$ is of this form (for a unique α).

Proof. It is obvious that α defines a holomorphic map $\mathbb{C}/\Lambda\to\mathbb{C}/\Lambda'$. Conversely, let $\varphi:\mathbb{C}/\Lambda\to\mathbb{C}/\Lambda'$ be a holomorphic map such that $\varphi(0)=0$, and let $\widetilde{\varphi}$ be its unique lifting to a holomorphic map $\mathbb{C}\to\mathbb{C}$ sending 0 to 0. For any $\omega\in\Lambda$, $z\mapsto\widetilde{\varphi}(z+\omega)-\widetilde{\varphi}(z)$ takes values in $\Lambda'\subset\mathbb{C}$. But a continuous map from a connected set to a set with the discrete topology is constant, and so the derivative of this function is zero:

$$\widetilde{\varphi}'(z+\omega)=\widetilde{\varphi}'(z).$$

Therefore $\widetilde{\varphi}'(z)$ is doubly periodic. As it is holomorphic, it must be constant, say $\widetilde{\varphi}'(z) = \alpha$ for all z. On integrating, we find that $\widetilde{\varphi}(z) = \alpha z + \beta$, and $\beta = \widetilde{\varphi}(0) = 0$. □

Corollary 10.5. *The Riemann surfaces $\mathbb{C}/\Lambda$ and $\mathbb{C}/\Lambda'$ are isomorphic if and only if $\Lambda' = \alpha\Lambda$ for some $\alpha \in \mathbb{C}^{\times}$.*

Proof. This is obvious from the proposition. □

The proposition shows that[11]

$$\mathrm{Hom}(\mathbb{C}/\Lambda, \mathbb{C}/\Lambda') \cong \{\alpha \in \mathbb{C} \mid \alpha\Lambda \subset \Lambda'\},$$

and the corollary shows that there is a one-to-one correspondence

$$\{\mathbb{C}/\Lambda\}/\approx \xleftrightarrow{1:1} \mathcal{L}/\mathbb{C}^{\times}.$$

The Weierstrass $\wp$ function. Let Λ be a lattice in $\mathbb{C}$. We don't yet know any nonconstant doubly periodic functions[12] for Λ. When G is a *finite* group acting on a set S, then it is easy to construct functions invariant under the action of G : take f to be any function $f : S \to \mathbb{C}$, and define

$$F(s) = \sum_{g \in G} f(gs);$$

then $F(g's) = \sum_{g\in G} f(g'gs) = F(s)$, and so F is invariant (and all invariant functions are of this form, obviously). When G is not finite, one has to verify that the series converges—in fact, in order to be able to change the order of summation, one needs absolute convergence. Moreover, when S is a Riemann surface and f is holomorphic, to ensure that F is holomorphic, one needs that the series converges absolutely uniformly on compact sets.

Now let $\varphi(z)$ be a holomorphic function $\mathbb{C}$ and write

$$\Phi(z) = \sum_{\omega \in \Lambda} \varphi(z + \omega).$$

Assume that as $|z| \to \infty$, $\varphi(z) \to 0$ so fast that the series for $\Phi(z)$ is absolutely convergent for all z for which none of the terms in the series has a pole. Then $\Phi(z)$ is doubly periodic with respect to Λ; for replacing z by $z + \omega_0$ for some $\omega_0 \in \Lambda$ merely rearranges the terms in the sum. This is the most obvious way to construct doubly periodic functions; similar methods can be used to construct functions on other quotients of domains.

To prove the absolute uniform convergence on compact subsets of such series, the following test is useful.

Lemma 10.6. *Let D be a bounded open subset of the complex plane and let $c > 1$ be constant. Suppose that $\psi(z, \omega)$, $\omega \in \Lambda$, is a function that is meromorphic in z for each ω and which satisfies the condition: there are constants A and B such that*

$$|\psi(z, m\omega_1 + n\omega_2)| < B(m^2 + n^2)^{-c}$$

whenever $m^2 + n^2 > A$. Then the series $\sum_{\omega\in\Lambda} \psi(z, \omega)$, with finitely many terms which have poles in D deleted, is uniformly absolutely convergent in D.

[11] I use $X \approx Y$ to mean that X and Y are isomorphic, and $X \cong Y$ to mean that they are isomorphic by a canonical (or given) isomorphism.

[12] For a lattice Λ in $\mathbb{C}^n$, $n > 1$, there frequently won't be any nonconstant holomorphic functions on $\mathbb{C}^n/\Lambda$.

Proof. That only finitely many terms can have poles in D follows from the condition. To prove the lemma it suffices to show that, given any $\varepsilon > 0$, there is an integer N such that $S < \varepsilon$ for every finite sum $S = \sum |\psi(z, m\omega_1 + n\omega_2)|$ in which all the terms are distinct and each one of them has $m^2 + n^2 \geq 2N^2$. Now S consists of eight subsums, a typical member of which consists of the terms for which $m \geq n \geq 0$. (There is some overlap between these sums, but that is harmless.) In this subsum we have $m \geq N$ and $\psi < Bm^{-2c}$, assuming as we may that $2N^2 > A$; and there are at most $m+1$ possible values of n for a given m. Thus

$$S \leq \sum_{m=N}^{\infty} B\ m^{-2c}(m+1) < B_1 N^{2-2c}$$

for a suitable constant B_1, and this proves the lemma. □

We know from Corollary 10.3 that the simplest possible nonconstant doubly periodic function is one with a double pole at each point of Λ and no other poles. Suppose $f(z)$ is such a function. Then $f(z) - f(-z)$ is a doubly periodic function with no poles except perhaps simple ones at the points of Λ. Hence it must be constant, and since it is an odd function it must vanish. Thus $f(z)$ is even, and we can make it unique by imposing the normalization condition $f(z) = z^{-2} + O(z^2)$ near $z = 0$—it turns out to be convenient to force the constant term in this expansion to vanish rather than to assign the zeros of $f(z)$. There is such an $f(z)$—indeed it is the Weierstrass function $\wp(z)$—but we can't define it by the method at the start of this subsection because if $\varphi(z) = z^{-2}$, the series $\Phi(z)$ is not absolutely convergent. However, if $\varphi(z) = -2z^{-3}$, we can apply this method, and it gives $\wp'$, the derivative of the Weierstrass $\wp$-function. Define

$$\wp'(z;\Lambda) = \wp'(z;\omega_1,\omega_2) = \sum_{\omega\in\Lambda} \frac{-2}{(z-\omega)^{-3}}$$

and

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega\in\Lambda,\omega\neq 0} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}\right).$$

They are both meromorphic doubly periodic functions on $\mathbb{C}$, and $\wp' = \frac{d\wp}{dz}$.

Eisenstein series. Let Λ be a lattice in $\mathbb{C}$, and consider the sum

$$\sum_{\omega\in\Lambda,\ \omega\neq 0} \frac{1}{\omega^n}.$$

The map $\omega \mapsto -\omega : \Lambda \to \Lambda$ has order 2, and its only fixed point is 0. Therefore $\Lambda \setminus \{0\}$ is a disjoint union of its orbits, and it follows that the sum is zero if n is odd. We write

$$G_k(\Lambda) = \sum_{\omega\in\Lambda,\ \omega\neq 0} \frac{1}{\omega^{2k}},$$

and we let $G_k(\tau) = G_k(\mathbb{Z}\tau + \mathbb{Z})$, $\tau \in \mathbb{H}$.

Proposition 10.7. *For all integers $k \geq 2$, $G_k(\tau)$ converges to a holomorphic function on $\mathbb{H}$.*

Proof. Apply Lemma 10.6. □

The functions $G_k(\Lambda)$ and $G_k(\tau)$ and are called *Eisenstein series.* Note that $G_k(c\Lambda) = c^{-2k}G_k(\Lambda)$ for $c \in \mathbb{C}^\times$.

The field of doubly periodic functions. Let Λ be a lattice in $\mathbb{C}$. The doubly periodic functions for Λ form a field, which the next two propositions determine.

Proposition 10.8. *There is the following relation between $\wp$ and $\wp'$:*

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3$$

where $g_2 = 60G_2(\Lambda)$ and $g_3 = 140G_3(\Lambda)$.

Proof. We compute the Laurent expansion of $\wp(z)$ near 0. Recall (from Math 115) that for $|t| < 1$,

$$\frac{1}{1-t} = 1 + t + t^2 + \cdots .$$

On differentiating this, we find that

$$\frac{1}{(1-t)^2} = \sum_{n\geq 1} nt^{n-1} = \sum_{n\geq 0}(n+1)t^n.$$

Hence, for $|z| < |\omega|$,

$$\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} = \frac{1}{\omega^2}\left(\frac{1}{\left(1-\frac{z}{\omega}\right)^2} - 1\right) = \sum_{n\geq 1}(n+1)\frac{z^n}{\omega^{n+2}}.$$

On putting this into the definition of $\wp(z)$ and changing the order of summation, we find that for $|z| < |\omega|$

$$\begin{aligned} \wp(z) &= \frac{1}{z^2} + \sum_{n\geq 1}\sum_{\omega\neq 0}(n+1)\frac{z^n}{\omega^{n+2}} \\ &= \frac{1}{z^2} + \sum_{k\geq 1}(2k+1)G_{k+1}(\Lambda)z^{2k} \\ &= \frac{1}{z^2} + 3G_2z^2 + 5G_3z^4 + \cdots . \end{aligned}$$

This last expression contains enough terms to show that the Laurent expansion of

$$\wp'(z)^2 - 4\wp(z)^3 + 60Gs_2(\Lambda)\wp(z) + 140G_3(\Lambda)$$

has no nonzero term in z^n with $n \leq 0$. Therefore this function is holomorphic at 0 and takes the value 0 there. Since it is doubly periodic and has no other poles in a suitable fundamental domain containing 0, we see that it is constant, and in fact zero. □

Proposition 10.9. *The doubly periodic functions for Λ are precisely the rational functions of $\wp(z)$ and $\wp'(z)$, i.e., if f is doubly periodic, then there exist $F(X,Y), G(X,Y) \in \mathbb{C}[X,Y]$, $G \neq 0$, such that $f(z) = F(\wp(z), \wp'(z))/G(\wp(z), \wp'(z))$.*

Proof. Omitted. □

Proposition 10.8 shows that $(X,Y) \mapsto (\wp(z), \wp'(z))$ defines a homomorphism

$$\mathbb{C}[x,y] =_{df} \mathbb{C}[X,Y]/(Y^2 - 4X^3 + g_2X + g_3) \to \mathbb{C}[\wp, \wp'],$$

where $\mathbb{C}[\wp, \wp']$ is the $\mathbb{C}$-algebra of meromorphic functions on $\mathbb{C}$ generated by $\wp$ and $\wp'$. I claim[13] that the map is an isomorphism. For this, we have to show that a polynomial

[13] Those who know some commutative algebra will be able to give a simpler proof.

$g(X,Y) \in \mathbb{C}[X,Y]$ for which $g(\wp,\wp') = 0$ is divisible by $f(X,Y) =_{df} Y^2 - X^3 + g_2X + g_3$. The theory of resultants (see the end of this section) shows that for any polynomial $g(X,Y)$, there exist polynomials $a(X,Y)$ and $b(X,Y)$ such that

$$a(X,Y)f(X,Y) + b(X,Y)g(X,Y) = R(X) \in \mathbb{C}[X]$$

with $\deg_Y(b(X,Y)) < \deg_Y(f(X,Y))$. Hence if $g(\wp,\wp') = 0$, then $R(\wp) = 0$, but it is easy to see that $\wp$ is transcendental over $\mathbb{C}$ (for example, it has infinitely many poles). Therefore $R = 0$, and so $f(X,Y)$ divides $b(X,Y)g(X,Y)$. Any polynomial with the form of $f(X,Y)$ is irreducible, and so $f(X,Y)$ divides either $b(X,Y)$ or $g(X,Y)$. Because of the degrees, it can't divide b, and so it must divide g.

The isomorphism $\mathbb{C}[x,y] \to \mathbb{C}[\wp,\wp']$ induces an isomorphism of the fields of fractions

$$\mathbb{C}(x,y) \to \mathbb{C}(\wp,\wp').$$

Proposition 10.9 shows that $\mathbb{C}(\wp,\wp')$ is the field of all double periodic functions for Λ.

The elliptic curve $E(\Lambda)$. Let Λ be a lattice in $\mathbb{C}$.

Lemma 10.10. *The polynomial $f(X) = 4X^3 - g_2(\Lambda)X - g_3(\Lambda)$ has distinct roots.*

Proof. The function $\wp'(z)$ is odd and doubly periodic, and so

$$\wp'(\frac{\omega_1}{2}) = -\wp'(-\frac{\omega_1}{2}), = \wp'(-\frac{\omega_1}{2}).$$

Hence $\wp'(z)$ has a zero at $\omega_1/2$, and so Propositiion 10.8 shows that $\wp(\omega_1/2)$ is a root of $f(X)$. The same argument shows that $\wp(\omega_2/2)$ and $\wp((\omega_1+\omega_2)/2)$ are also roots. It remains to prove that these three numbers are distinct.

The function $\wp(z) - \wp(\omega_1/2)$ has a zero at $\omega_1/2$, which must be a double zero because its derivative is also 0 there. Since $\wp(z) - \wp(\omega_1/2)$ has only one (double) pole in a fundamental domain D containing 0, Proposition 10.2 shows that $\omega_1/2$ is the only zero of $\wp(z) - \wp(\omega_1/2)$ in D, i.e., that $\wp(z)$ takes the value $\wp(\omega_1/2)$ only at $z = \omega_1/2$ within D. In particular, $\wp(\omega_1/2)$ is not equal to $\wp(\omega_2/2)$ or $\wp((\omega_1+\omega_2)/2)$. □

From the lemma, we see that

$$E(\Lambda) : Y^2Z = 4X^3 - g_2(\Lambda)XZ^2 - g_3(\Lambda)Z^3$$

is an elliptic curve. Recall that $c^4g_2(c\Lambda) = g_2(\Lambda)$ and $c^6g_3(c\Lambda) = g_3(\Lambda)$ for any $c \in \mathbb{C}^\times$, and so $c\Lambda$ defines essentially the same elliptic curve as Λ.

Proposition 10.11. *The map*

$$\begin{matrix} z \mapsto (\wp(z) : \wp'(z) : 1) \\ 0 \mapsto (0 : 1 : 0) \end{matrix} : \mathbb{C}/\Lambda \to E(\Lambda)$$

is an isomorphism of Riemann surfaces.

Proof. It is certainly a well-defined map. The function $\wp(z)$ is $2:1$ in a period parallelogram containing 0, except at the points $\frac{\omega_1}{2}$, $\frac{\omega_2}{2}$, $\frac{\omega_1+\omega_2}{2}$, where it is one-to-one. Since the function $(x : y : 1) \mapsto x : E(\Lambda) \setminus \{O\} \to \mathbb{C}$ has the same property, and both maps have image the whole of $\mathbb{C}$, this shows that the map in $z \mapsto (\wp(z) : \wp'(z) : 1)$ is one-to-one. Finally, one can verify that it induces isomorphisms on the tangent spaces. □

The addition formula. Consider $\wp(z+z')$. It is a doubly periodic function of z, and therefore it is a rational function of $\wp$ and $\wp'$. The next result exhibits the rational function.

Proposition 10.12. *The following formula holds:*

$$\wp(z+z') = \frac{1}{4}\left\{\frac{\wp'(z)-\wp'(z')}{\wp(z)-\wp(z')}\right\}^2 - \wp(z) - \wp(z').$$

Proof. Let $f(z)$ denote the difference between the left and the right sides. Its only possible poles (in D) are at 0, or $\pm z'$, and by examining the Laurent expansion of $f(z)$ near these points one sees that it has no pole at 0 or z, and at worst a simple pole at z'. Since it is doubly periodic, it must be constant, and since $f(0) = 0$, it must be identically zero. □

Corollary 10.13. *The map* $z \mapsto (\wp(z) : \wp'(z) : 1) : \mathbb{C}/\Lambda \to E(\Lambda)$ *is a homomorphism of groups.*

Proof. The above formula agrees with the formula for the x-coordinate of the sum of two points on $E(\Lambda)$. [Let $Y = mX + c$ be the line through the points $P = (x, y)$ and $P' = (x', y')$ on the curve $Y^2 = 4X^3 - g_2X - g_3$. Then the x, x', and $x(P + P')$ are the roots of the polynomial

$$(mX + c)^2 - 4X^3 + g_2X + g_3,$$

and so

$$x(P+P') + x + x' = m^2 = \left(\frac{y-y'}{x-x'}\right)^2.]$$

□

Classification of elliptic curves over $\mathbb{C}$.

Theorem 10.14. *Every elliptic curve E over $\mathbb{C}$ is of the form $E(\Lambda)$ for some lattice Λ.*

Proof. This follows from the next two lemmas. □

Lemma 10.15. *Two elliptic curve*

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in k$$

and

$$E : Y^2Z = X^3 + a'XZ^2 + b'Z^3, \quad a', b' \in k$$

over an algebraically closed field k of characteristic $\neq 2, 3$ are isomorphic if and only if $j(E) = j(E')$.

Proof. According to Theorem 5.3, E and E' are isomorphic if and only if there exists a $c \in k^\times$ such that $a' = c^4a$ and $b' = c^6b$. Since $j(E) = \frac{1728(4a^3)}{4a^3+27b^2}$, it is clear that $E \approx E' \implies j(E) = j(E')$. Conversely, suppose $j(E) = j(E')$. Note first that

$$a = 0 \iff j(E) = 0 \iff j(E') = 0 \implies a' = 0.$$

Hence we may suppose that a and a' are both nonzero. After replacing (a, b) with (c^4a, c^6b) where $c = \sqrt[4]{\frac{a'}{a}}$ we will have that $a = a'$. Now $j(E) = j(E') \implies b = \pm b'$. A minus sign can be removed by a change of variables with $c = \sqrt{-1}$. □

For any lattice Λ in $\mathbb{C}$, the curve

$$E(\Lambda) : Y^2Z = 4X^3 - g_2(\Lambda)XZ^2 - g_3Z^3$$

has discriminant $\Delta(\Lambda) = g_2(\Lambda)^3 - 27g_3(\Lambda)^2$ and j-invariant

$$j(\Lambda) = \frac{1728g_2(\Lambda)^3}{g_2(\Lambda)^3 - 27g_3(\Lambda)^2}.$$

For $c \in \mathbb{C}^\times$, $g_2(c\Lambda) = c^{-4}g_2(\Lambda)$ and $g_3(c\Lambda) = c^{-6}g_3(\Lambda)$, and so the isomorphism class of $E(\Lambda)$ depends only on Λ up to scaling. Define

$$j(\tau) = j(\mathbb{Z}\tau + \mathbb{Z}).$$

Then, for any $\left(\begin{array}{cc} a & b \\ c & d \end{array}\right) \in \mathrm{SL}_2(\mathbb{Z})$,

$$j(\frac{a\tau + b}{c\tau + d}) = j(\tau).$$

Hence j defines a function on the quotient space $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$.

Lemma 10.16. *The function j defines an isomorphism* $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H} \to \mathbb{C}$.

Proof. We omit the proof (and hope to return to it later). □

Summary. For any subfield k of $\mathbb{C}$, we have the diagram:

$$\begin{tikzcd}
\{\text{Elliptic curves}/\mathbb{C}\}/\approx \arrow[r, leftrightarrow, "1:1"] & \mathcal{L}/\mathbb{C}^\times \arrow[r, leftrightarrow, "1:1"] & \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H} \arrow[r, "j", "\approx"'] & \mathbb{C} \\
\{\text{Elliptic curves}/k\}/\approx \arrow[u] \arrow[rrr, "j"] & & & k \arrow[u]
\end{tikzcd}$$

The bottom map is surjective, because for any $j \neq 0, 1728$, the curve

$$Y^2Z = X^3 - \frac{27}{4}\frac{j}{j-1728}XZ^2 - \frac{27}{4}\frac{j}{j-1728}$$

has j-invariant j. The left hand vertical map and the bottom map are injective if k is algebraically closed.

Aside 10.17. The above picture can be made a little more precise. Consider the isomorphism

$$z \mapsto (\wp(z) : \wp'(z) : 1) : \mathbb{C}/\Lambda \to E(\mathbb{C}).$$

Since $x = \wp(z)$ and $y = \wp'(z)$,

$$\frac{dx}{y} = \frac{\wp'(z)dz}{\wp'(z)} = dz.$$

Thus the differential dz on $\mathbb{C}$ maps to the differential $\frac{dx}{y}$ on $E(\mathbb{C})$. Conversely, from a holomorphic differential ω on $E(\mathbb{C})$ we can obtain an realization of E as a quotient $\mathbb{C}/\Lambda$ as follows. For $P \in E(\mathbb{C})$, consider $\varphi(P) = \int_O^P \omega \in \mathbb{C}$. This is a not well defined because it depends on the choice of a path from O to P. However, if we choose a $\mathbb{Z}$-basis (γ_1, γ_2) for $H_1(E(\mathbb{C}), \mathbb{Z})$, and set $\omega_1 = \int_{\gamma_1} \omega$, $\omega_2 = \int_{\gamma_2} \omega$, then $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ is a lattice in $\mathbb{C}$, and $P \mapsto \varphi(P)$ is an isomorphism $E(\mathbb{C}) \to \mathbb{C}/\Lambda$. In this way, we obtain a natural one-to-one correspondence between $\mathcal{L}$ and the set of isomorphism classes of pairs (E, ω) consisting of an elliptic curve E over $\mathbb{C}$ and a holmorphic differential ω on E.

Torsion points. Frequently, I write $X_n = \{x \in X \mid nx = 0\}$. For an elliptic curve E over $\mathbb{C}$, from $E(\mathbb{C}) = \mathbb{C}/\Lambda$ we see that

$$E(\mathbb{C})_n = \frac{1}{n}\Lambda/\Lambda = \{\frac{a}{n}\omega_1 + \frac{b}{n}\omega_2 \mid a, b \in \mathbb{Z}\}/\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2.$$

This is a free $\mathbb{Z}/n\mathbb{Z}$-module of rank 2. Because of this description over $\mathbb{C}$, torsion points on elliptic curves are often called *division points.*

Theorem 10.18. *For any elliptic curve E over an algebraically closed field k of characteristic zero, $E(k)_n$ is a free $\mathbb{Z}/n\mathbb{Z}$-module of rank 2.*

Proof. There will exist an algebraically closed subfield k_0 of finite transcendance degree over $\mathbb{Q}$ such that E arises from a curve E_0 over k_0. Now k_0 can be embedded into $\mathbb{C}$, and so we can apply the next lemma (twice). □

Lemma 10.19. *Let E be an elliptic curve over an algebraically closed field k, and let Ω be an algebraically closed field containing k. Then the map $E(k) \to E(\Omega)$ induces an isomorphism on the torsion subgroups.*

Proof. Let E be the curve $Y^2Z = X^3 + aXZ^2 + bZ^3$. There are inductively defined universal polynomials $\psi_m(X,Y) \in \mathbb{Z}[X,Y]$ (depending on a, b), such that for any point $P = (x : y : 1)$ of E, $mP = (X\psi_m^4 - \psi_{m-1}\psi_m^2\psi_{m+1} : \frac{1}{2}\psi_{2m} : \psi_m^4)$. See for example $[C_2]$ p133. Therefore $P \in E(k)_m$ if and only if $\psi_m(x,y) = 0$. Thus this lemma follows from the next. □

Lemma 10.20. *Let $k \subset \Omega$ be algebraically closed fields. If $F(X,Y), G(X,Y) \in k[X,Y]$ have no common factor, then any common solution to the equations*

$$\begin{cases} F(X,Y) &= 0 \\ G(X,Y) &= 0 \end{cases}$$

with coordinates in Ω in fact has coordinates in k.

Proof. From the theory of resultants, we know that there exist polynomials $a(X,Y)$, $b(X,Y)$, and $R(X)$ with coefficients in k such that

$$a(X,Y)F(X,Y) + b(X,Y)G(X,Y) = R(X)$$

and $R(x_0) = 0$ if and only if $F(x_0,Y)$ and $G(x_0,Y)$ have a common zero. In other words, the roots of R are the x-coordinates of the common zeros of $F(X,Y)$ and $G(X,Y)$. Since $R(X)$ is a polynomial in one variable, its roots all lie in k. Moreover, for a given $x_0 \in k$, all the common roots of $F(x_0,Y)$ and $G(x_0,Y)$ lie in k. □

Remark 10.21. (a) Theorem 10.18 holds for elliptic curves over algebraically closed fields of characteristic $p \neq 0$ if (and only if) n is not divisible by p.

(b) In contrast to $E(\mathbb{Q}^{\mathrm{al}})$, the torsion subgroup of $E(\mathbb{Q})$ is quite small. It was conjectured by Beppo Levi at the International Congress in 1906 and proved by Mazur in 1975 that the $E(\mathbb{Q})_{\mathrm{tors}}$ is isomorphic to one of the following groups:

$$\begin{array}{lll} \mathbb{Z}/m\mathbb{Z} & \text{for} & m = 1, 2, \ldots, 10, 12; \\ \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z} & \text{for} & m = 2, 4, 6, 8. \end{array}$$

The 15 curves in Exercise 8.11 exhibit all possible torsion subgroups (in order). The fact that $E(\mathbb{Q})_{\mathrm{tors}}$ is so much smaller than $E(\mathbb{Q}^{\mathrm{al}})_{\mathrm{tors}}$ shows that the image of the Galois group in the automorphism group of $E(\mathbb{Q}^{\mathrm{al}})_{\mathrm{tors}}$ is large.

Endomorphisms. A field K of finite degree over $\mathbb{Q}$ is called an *algebraic number field.* Each $\alpha \in K$ satisfies an equation,

$$\alpha^m + a_1\alpha^{m-1} + \cdots + a_m = 0, \quad a_i \in \mathbb{Q}.$$

If it satisfies such an equation with the $a_i \in \mathbb{Z}$, then α is said to be an *(algebraic) integer* of K. The algebraic integers form a subring $\mathcal{O}_K$ of K, which is a free $\mathbb{Z}$-module of rank $[K : \mathbb{Q}]$. (Experts in commutative algebra will recognize $\mathcal{O}_K$ as being the integral closure of $\mathbb{Z}$ in K.) For example, if $K = \mathbb{Q}[\sqrt{d}]$ with $d \in \mathbb{Z}$ and square-free, then

$$\mathcal{O}_K = \begin{cases} \mathbb{Z}1 + \mathbb{Z}\sqrt{d} & d \not\equiv 1 \mod 4 \\ \mathbb{Z}1 + \mathbb{Z}\frac{1+\sqrt{d}}{2} & d \equiv 1 \mod 4. \end{cases}$$

Proposition 10.22. *Let $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ be a lattice in $\mathbb{C}$ with $\tau = \omega_1/\omega_2 \in \mathbb{H}$. The ring* $\mathrm{End}(\mathbb{C}/\Lambda) = \mathbb{Z}$ *unless* $[\mathbb{Q}[\tau] : \mathbb{Q}] = 2$*, in which case $R = \mathrm{End}(\mathbb{C}/\Lambda)$ is a subring of $\mathbb{Q}[\tau]$ of rank 2 as a $\mathbb{Z}$-module.*

Proof. Let $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ with $\tau =_{df} \omega_1/\omega_2 \in \mathbb{H}$, and suppose that there exists an $\alpha \in \mathbb{C}$, $\alpha \notin \mathbb{Z}$, such that $\alpha\Lambda \subset \Lambda$. Then

$$\begin{aligned} \alpha\omega_1 &= a\omega_1 + b\omega_2 \\ \alpha\omega_2 &= c\omega_1 + d\omega_2, \end{aligned}$$

with $a, b, c, d \in \mathbb{Z}$. On dividing through by ω_2 we obtain the equations

$$\begin{aligned} \alpha\tau &= a\tau + b \\ \alpha &= c\tau + d. \end{aligned}$$

As $\alpha \notin \mathbb{Z}$, $c \neq 0$.

On eliminating α from between the two equations, we find that

$$c\tau^2 + (d-a)\tau + b = 0.$$

Therefore $\mathbb{Q}[\tau]$ is of degree 2 over $\mathbb{Q}$.

On eliminating τ from between the two equations, we find that

$$\alpha^2 - (a+d)\alpha + bc = 0.$$

Therefore α is integral over $\mathbb{Z}$, and hence is contained in the ring of integers of $\mathbb{Q}[\tau]$. □

Example 10.23. (a) Consider $E : Y^2Z = X^3 + aXZ^2$. Then $(x : y : z) \mapsto (-x : iy : z)$ is an endomorphism of E of order 4, and so $\mathrm{End}(E) = \mathbb{Z}[i]$. Note that E has j-invariant 1728.

(b) Consider $E : Y^2Z = X^3 + bZ^3$, and let $\rho = e^{2\pi i/3}$. Then $(x : y : z) \mapsto (\rho x : y : z)$ is an endomorphism of E of order 3 of E. In this case, E has j-invariant is 0.

Aside 10.24. (For the experts.) Recall that a complex number α is said to be *algebraic* if it is algebraic over $\mathbb{Q}$, and is otherwise said to be *transcendental.* There is a general philosophy that a transcendental meromorphic function f should take transcendental values at the algebraic points in $\mathbb{C}$, except at some "special" points, where it has interesting "special values". We illustrate this for two functions.

(a) Define $e(z) = e^{2\pi iz}$. If z is algebraic but not rational, then $e(z)$ is transcendental. [More generally, if α and β are algebraic, $\alpha \neq 0, 1$, and β is irrational, then α^{β} is transcendental—Hilbert stated this as the seventh of his famous problems, and Gelfond and Schneider proved[14] it in 1934. It implies our statement because $e(z) = (e^{\pi i})^{2z}$.]

On the other hand, if $z \in \mathbb{Q}$, then $e(z)$ is algebraic—in fact, it is a root of 1, and so $\mathbb{Q}[e(z)]$ is a finite extension of $\mathbb{Q}$ with abelian Galois group (see Math 594). It is a famous theorem (the Kronecker-Weber theorem) that every such extension of $\mathbb{Q}$ is contained in $\mathbb{Q}[e(\frac{1}{m})]$ for some m (see Math 776).

Let $\tau \in \mathbb{H}$ be algebraic. If τ generates a quadratic extension of $\mathbb{Q}$, then $j(\tau)$ is algebraic, and otherwise $j(\tau)$ is transcendental (the second statement was proved by Siegel in 1949).

In fact, when $[\mathbb{Q}[\tau] : \mathbb{Q}] = 2$, one can say much more. Assume that $\mathbb{Z}[\tau]$ is the ring of integers in $K =_{df} \mathbb{Q}[\tau]$. Then $j(\tau)$ is an algebraic integer, and

$$[\mathbb{Q}[j(\tau)] : \mathbb{Q}] = [K[j(\tau)] : K] = h_K$$

where h_K is the class number of K. Moreover, $K[j(\tau)]$ is the Hilbert class field of K (the largest unramified abelian extension of K).

Appendix: Resultants. Let $f(X) = s_0X^m + s_1X^{m-1} + \cdots + s_m$ and $g(X) = t_0X^n + t_1X^{n-1} + \cdots + t_n$ be polynomials with coefficients in a field k. The *resultant* $\mathrm{Res}(f, g)$ of f and g is defined to be the determinant

$$\begin{vmatrix} s_0 & s_1 & \dots & s_m & & \\ & s_0 & \dots & & s_m & \\ & & \dots & & & \dots \\ t_0 & t_1 & \dots & t_n & & \\ & t_0 & \dots & & t_n & \\ & & \dots & & & \dots \end{vmatrix}$$

There are n rows of s's and m rows of t's, so that the matrix is $(m+n) \times (m+n)$; all blank spaces are to be filled with zeros. The resultant is a polynomial in the coefficients of f and g.

Proposition 10.25. *The resultant* $\mathrm{Res}(f, g) = 0$ *if and only if*

(i) *both s_0 and t_0 are zero; or*
(ii) *the two polyomials have a common root in k^{al}.*

Proof. If (i) holds, then the first column of the determinant is zero, and so certainly $\mathrm{Res}(f, g) = 0$. Suppose that α is a common root of f and g, so that there exist polynomials f_1 and g_1 in $k^{\mathrm{al}}[X]$ of degrees $m-1$ and $n-1$ respectively such that

$$f(X) = (X - \alpha)f_1(X), \qquad g(X) = (X - \alpha)g_1(X).$$

From these equations we find that

$$f(X)g_1(X) - g(X)f_1(X) = 0. \qquad (*)$$

[14] At about the time he stated his problems (1900), Hilbert gave a lecture in which he said that he expected the Riemann hypothesis to be proved within his lifetime, that Fermat's last theorem would be proved within the lifetimes of the youngest members of his audience, but that no one in the audience would see his seventh problem proved. He was close with Fermat's last problem.

On equating the coefficients of $X^{m+n-1}, \ldots, X, 1$ in (*) to zero, we find that the coefficients of f_1 and g_1 are the solutions of a system of $m+n$ linear equations in $m+n$ unknowns. The matrix of coefficients of the system is the transpose of the matrix

$$\begin{pmatrix} s_0 & s_1 & \ldots & s_m & & \\ & s_0 & \ldots & & s_m & \\ & & \ldots & & & \ldots \\ t_0 & t_1 & \ldots & t_n & & \\ & t_0 & \ldots & & t_n & \\ & & \ldots & & & \ldots \end{pmatrix}$$

The existence of the solution shows that this matrix has determinant zero, which implies that $\operatorname{Res}(f,g) = 0$.

Conversely, suppose that $\operatorname{Res}(f,g) = 0$ but neither s_0 nor t_0 is zero. Because the above matrix has determinant zero, we can solve the linear equations to find polynomials f_1 and g_1 satisfying (*). If α is a root of f, then it must also be a root of f_1 or g. If the former, cancel $X - \alpha$ from the left hand side of (*) and continue. As $\deg f_1 < \deg f$, we eventually find a root of f that is not a root of f_1, and so must be a root of g. □

Let $c_1, \ldots, c_{m+n}$ be the columns of the above matrix. Then

$$\begin{pmatrix} X^{m-1}f(X) \\ X^{m-2}f \\ \vdots \\ f(X) \\ X^{n-1}g(X) \\ \vdots \\ g(X) \end{pmatrix} = X^{m+n-1}c_0 + \cdots + 1c_{m+n},$$

and so

$$\operatorname{Res}(f,g) =_{df} \det(c_0, \ldots, c_{m+n}) = \det(c_0, \cdots, c_{m+n-1}, c)$$

where c is the vector on the left of the above equation. On expanding out this last determinant, we find that

$$\operatorname{Res}(f,g) = a(X)f(X) + b(X)g(X)$$

where $a(X)$ and $b(X)$ are polynomials of degrees $\leq n-1$ and $\leq m-1$ respectively.

Remark 10.26. If the $f(X)$ and $g(X)$ have coefficients in an integral domain R, for example, $\mathbb{Z}$ or $k[Y]$, then $\operatorname{Res}(f,g) \in R$, and the polynomials $a(X)$ and $b(X)$ have coefficients in R.

For a monic polynomial $f(X) = X^m + \cdots + s_m$, the resultant of $f(X)$ and $f'(X)$ is called the *discriminant* of f (apart possibly for a minus sign).

The resultant of homogeneous polynomials $F(X,Y) = s_0X^m + s_1X^{m-1}Y + \cdots + s_mY^m$ and $G(X,Y) = t_0X^n + t_1X^{n-1}Y + \cdots + t_nY^n$ is defined as for inhomogeneous polynomials.

Proposition 10.27. *The resultant* $\operatorname{Res}(F,G) = 0$ *if and only if* F *and* G *have a nontrivial zero in* $\mathbb{P}^1(k^{al})$.

Proof. The nontrivial zeros of $F(X,Y)$ in $\mathbb{P}^1(k^{\mathrm{al}})$ are of the form:

(i) $(a:1)$ with a a root of $F(X,1)$, or

(ii) $(1:0)$ in the case that $s_0 = 0$.

Since a similar statement is true for $G(X,Y)$, this proposition is a restatement of the previous proposition. □

Clearly, the statement is more pleasant in the homogeneous case.

Maple can find the resultant of two polynomials in one variable: for example, entering "resultant$((x+a)^5,(x+b)^5,x)$" gives the answer $(-a+b)^{25}$. Explanation: the polynomials have a common root if and only if $a=b$, and this can happen in 25 ways.

Aside 10.28. There is a geometric interpretation of the last proposition. Take k to be algebraically closed, and regard the coefficients of F and G as indeterminants. Let V be the subset of $\mathbb{A}^{m+n+2}\times\mathbb{P}^1$ where both $F(s_0,\ldots,s_m;X,Y)$ and $G(t_0,\ldots,t_n;X,Y)$ vanish. The proposition says that the projection of V on $\mathbb{A}^{m+n+2}$ is the set where $\mathrm{Res}(F,G)$, regarded as a polynomial in the s_i and t_i, vanishes. In other words, the proposition tells us that the projection of the particular Zariski-closed set V is the Zariski-closed set defined by the resultant of F and G.

Elimination theory does this in general. Given polynomials $P_i(T_1,\ldots,T_m;X_0,\ldots,X_n)$, homogeneous in the X_i, it provides an algorithm for finding polynomials $R_j(T_1,\ldots,T_n)$ such that the $P_i(a_1,\ldots,a_m;X_0,\ldots,X_n)$ have a common zero if and only if $R_j(a_1,\ldots,a_n)=0$ for all j. See, for example, Cox et al, Ideals, Varieties, and Algorithms, p388.

Exercise 10.29. (a) Prove that, for all z_1, z_2,

$$\begin{vmatrix} \wp(z_1) & \wp'(z_1) & 1 \\ \wp(z_2) & \wp'(z_2) & 1 \\ \wp(z_1+z_2) & -\wp'(z_1+z_2) & 1 \end{vmatrix} = 0.$$

(b) Compute sufficiently many initial terms for the Laurent expansions of $\wp'(z)$, $\wp'(z)^2$, etc., to verify the equation in Proposition 10.8.

11. The Mordell-Weil Theorem: Statement and Strategy

We state the Mordell-Weil (or finite basis) theorem, and outline the strategy for proving it.

Theorem 11.1 (Mordell-Weil). *For any elliptic curve E over a number field K, $E(K)$ is finitely generated.*

The theorem was proved by Mordell (1922) when $K=\mathbb{Q}$, and for all number fields by Weil in his thesis (1928). Weil in fact proved a much more general result, namely, he showed that for any nonsingular projective curve C over a number field K, the group $\mathrm{Pic}^0(C)$ is finitely generated. As we noted in (4.7), for an elliptic curve $C(K)=\mathrm{Pic}^0(C)$. The theorem was proved for all abelian varieties over number fields by Taniyama in 1954.

The first step in proving the theorem is to prove a weaker result:

Theorem 11.2 (Weak Mordell-Weil Theorem). *For any elliptic curve E over a number field K and any integer n, $E(K)/nE(K)$ is finite.*

Clearly, for an abelian group M,

$$M \text{ finitely generated } \implies M/nM \text{ finite for all } n > 1,$$

but the converse statement is false. For example, $\mathbb{Q}$ regarded as a group under addition has the property that $\mathbb{Q} = n\mathbb{Q}$ and so $\mathbb{Q}/n\mathbb{Q} = 0$, but the elements of any finitely generated subgroup of $\mathbb{Q}$ will have bounded denominators, and $\mathbb{Q}$ is not finitely generated.

We assume now that $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite, and sketch how one deduces that $E(\mathbb{Q})$ is finitely generated. Recall that the height of a point $P \in \mathbb{P}^2(\mathbb{Q})$ is $H(P) = \max(|a|, |b|, |c|)$ where $P = (a : b : c)$ and a, b, c have been chosen to be integers with no common factor. For points P and Q on an elliptic curve E, not necessarily distinct, we shall relate $H(P + Q)$ to $H(P)$ and $H(Q)$.

Let $P_1, \dots, P_s \in E(\mathbb{Q})$ be a set of representatives for the elements of $E(\mathbb{Q})/2E(\mathbb{Q})$. Then any $Q \in E(\mathbb{Q})$ can be written

$$Q = P_i + 2Q'$$

for some i and for some $Q' \in E(\mathbb{Q})$. We shall show that, $H(Q') < H(Q)$, at least provided $H(Q)$ is greater than some fixed constant H_0. If $H(Q') > H_0$, we can repeat the argument for Q', etc., to obtain

$$Q = P_i + 2Q' = P_i + 2(P_{i'} + 2Q'') = \cdots .$$

Let $Q_1, \dots, Q_t$ be the set of points in $E(\mathbb{Q})$ with height $< H_0$. Then the above equation exhibits Q as a linear combination of P_i's plus a Q_j, and so the P_i's and Q_j's generate $E(\mathbb{Q})$.

Remark 11.3. The argument in the last paragraph is called "proof by descent". Fermat is generally credited with originating this method in his proof of Fermat's last theorem for the exponent 4 (which *was* short enough to fit in the margin). However, in some sense it goes back to the Greeks. Consider the proof that $Y^2 = 2X^2$ has no solution in integers. Define the height of a pair (m, n) of integers to be $\max(|m|, |n|)$. One proves that if (m, n) is one solution to the equation, then there exists another of smaller height, which leads to a contradiction.

12. Group cohomology

In proving the weak Mordell-Weil theorem, and also later in the study of the Tate-Shafarevich group, we shall use a little of the theory of the cohomology of groups.

Cohomology of finite groups. Let G be a finite group. A *G-module* is an abelian group M together with an action of G, i.e., a map $G \times M \to M$ such that

(a) $\sigma(m + m') = \sigma m + \sigma m'$ for all $\sigma \in G$, $m, m' \in M$;
(b) $(\sigma\tau)(m) = \sigma(\tau m)$ for all $\sigma, \tau \in G$, $m \in M$;
(c) $1m = m$ for all $m \in M$.

Thus, to give an action of G on M is the same as to give a homomorphism $G \to \mathrm{Aut}(M)$ (automorphisms of M as an abelian group).

Example 12.1. Let L be a finite Galois extension of a field K with Galois group G, and let E be an elliptic curve over K. Then L, $L^\times$, and $E(L)$ are all G-modules.

Let M be a G-module. We define

$$H^0(G,M) \quad = \quad M^G = \{m \in M \mid \sigma m = m, \text{ all } \sigma \in G\}.$$

For the G-modules in (12.1),

$$H^0(G,L) = K, \quad H^0(G,L^\times) = K^\times, \text{ and } H^0(G,E(L)) = E(K).$$

A *crossed homomorphism* is a map $f : G \to M$ such that

$$f(\sigma\tau) = f(\sigma) + \sigma f(\tau).$$

Note that the condition implies that $f(1) = f(1 \cdot 1) = f(1) + f(1)$, and so $f(1) = 0$. For any $m \in M$, we obtain a crossed homomorphism by putting

$$f(\sigma) = \sigma m - m, \qquad \text{all } \sigma \in G.$$

Such a crossed homomorphism is said to be *principal.* The sum of two crossed homomorphisms is again a crossed homomorphism, and the sum of two principal crossed homomorphisms is again principal. Thus we can define

$$H^1(G,M) = \frac{\{\text{crossed homomorphisms}\}}{\{\text{principal crossed homomorphisms}\}}.$$

There are also cohomology groups $H^n(G,M)$ for $n > 1$, but we won't need them.

Example 12.2. If G acts trivially on M, i.e., $\sigma m = m$ for all $\sigma \in G$ and $m \in M$, then a crossed homomorphism is simply a homomorphism, and every principal crossed homomorphism is zero. Hence $H^1(G,M) = \mathrm{Hom}(G,M)$.

Proposition 12.3. *Let L be a finite Galois extension of K with group G; then $H^1(G,L^\times) = 0$, i.e., every crossed homomorphism $G \to L^\times$ is principal.*

Proof. Let f be a crossed homomorphism $G \to L^\times$. In multiplicative notation, this means,

$$f(\sigma\tau) = f(\sigma) \cdot \sigma(f(\tau)), \quad \sigma, \tau \in G,$$

and we have to find a $\gamma \in L^\times$ such that $f(\sigma) = \sigma\gamma/\gamma$ for all $\sigma \in G$. Because the $f(\tau)$ are nonzero, Dedekind's theorem on the independence of characters (see Math 594) implies that

$$\sum f(\tau)\tau : L \to L$$

is not the zero map, i.e., that there exists an $\alpha \in L$ such that

$$\beta = \sum_{\tau \in G} f(\tau)\tau\alpha \neq 0.$$

But then, for $\sigma \in G$,

$$\sigma\beta = \sum_{\tau \in G} \sigma(f(\tau)) \cdot \sigma\tau(\alpha) = \sum_{\tau \in G} f(\sigma)^{-1} \cdot f(\sigma\tau) \cdot \sigma\tau(\alpha) = f(\sigma)^{-1} \sum_{\tau \in G} f(\sigma\tau)\sigma\tau(\alpha) = f(\sigma)^{-1}\beta,$$

which shows that $f(\sigma) = \beta/\sigma\beta = \sigma(\beta^{-1})/\beta^{-1}$. □

Proposition 12.4. *For any exact sequence of G-modules*

$$0 \to M \to N \to P \to 0,$$

there is a canonical exact sequence

$$0 \to H^0(G,M) \to H^0(G,N) \to H^0(G,P) \xrightarrow{\delta} H^1(G,M) \to H^1(G,N) \to H^1(G,P)$$

Proof. The map δ is defined as follows. Let $p \in P^G$. There exists an $n \in N$ mapping to p, and $\sigma n - n \in M$ for all $\sigma \in G$. The map $\sigma \mapsto \sigma n - n : G \to M$ is a crossed homomorphism, whose class we define to be $\delta(p)$. Another n' mapping to p gives rise to a crossed homomorphism differing from the first by a principal crossed homomorphism, and so $\delta(p)$ is well-defined. The rest of the proof is routine. □

Let H be a subgroup of G. The restriction map $f \mapsto f|H$ defines a homomorphism $\mathrm{Res} : H^1(G, M) \to H^1(H, M)$.

Proposition 12.5. *If G has order m, then m kills $H^1(G, M)$.*

Proof. In general, if H is a subgroup of G of index m, then there exists a homomorphism $\mathrm{Cor} : H^i(H, M) \to H^i(G, M)$ such that the composite $\mathrm{Res} \circ \mathrm{Cor}$ is multiplication by m. The proposition is proved by taking $H = 1$. □

Remark 12.6. Let H be a normal subgroup of a group G, and let M be a G-module. Then M^H is a G/H-module, and a crossed homomorphism $f : G/H \to M^H$ defines a crossed homomorphism $G \to M$ by composition:

$$\begin{tikzcd}
G \arrow[r, dashed] \arrow[d] & M \\
G/H \arrow[r, "f"] & M^H. \arrow[u, phantom, "\cup"]
\end{tikzcd}$$

In this way we obtain an "inflation" homomorphism

$$\mathrm{Inf} : H^1(G/H, M^H) \to H^1(G, M),$$

and one verifies easily that the sequence

$$0 \to H^1(G/H, M^H) \xrightarrow{\mathrm{Inf}} H^1(G, M) \xrightarrow{\mathrm{Res}} H^1(H, M)$$

is exact.

Cohomology of infinite Galois groups. Let k be a perfect field, and let k^{al} be an algebraic closure of k. The automorphisms of k^{al} fixing the elements of k form a group G, which when endowed with the topology for which the open subgroups are those fixing some finite extension of k, is called the *Galois group* of k^{al} over k. The group G is compact, and so any open subgroup of G is of finite index. Infinite Galois theory says that the intermediate fields K, $k \subset K \subset k^{\mathrm{al}}$, are in natural one-to-one correspondence with the closed subgroups of G. Under the correspondence intermediate fields of finite degree over k correspond to open subgroups of G.

A G-module M is said to be *discrete* if the map $G \times M \to M$ is continuous when M is given the discrete topology and G is given its natural topology. This is equivalent to requiring that

$$M = \cup_H M^H, \quad H \text{ open in } G,$$

i.e., to requiring that every element of M is fixed by the subgroup of G fixing some finite extension of k. For example, $M = k^{\mathrm{al}}$, $M = k^{\mathrm{al}\times}$, and $M = E(k^{\mathrm{al}})$ are all discrete G-modules because

$$k^{\mathrm{al}} = \cup K, \quad k^{\mathrm{al}} = \cup K^\times, \text{ and } E(k^{\mathrm{al}}) = \cup E(K)$$

where, in each case, the union runs over the finite extensions K of k contained in k^{al}.

For an infinite Galois group G, we define $H^1(G, M)$ to be the group of continuous crossed homomorphisms $f : G \to M$ modulo the subgroup of principal crossed homomorphisms. With this definition

$$H^1(G, M) = \varinjlim_H H^1(G/H, M^H)$$

where H runs through the open normal subgroups of G. Explicitly, this means that:

(a) $H^1(G, M)$ is the union of the images of the inflation maps $\mathrm{Inf} : H^1(G/H, M^H) \to H^1(G, M)$, H an open normal subgroup of G;

(b) an element $\gamma \in H^1(G/H, M^H)$ maps to zero in $H^1(G, M)$ if and only if it maps to zero $H^1(G/H', M^{H'})$ for some open normal subgroup H' of G contained in H.

In particular, the group $H^1(G, M)$ is torsion.

Example 12.7. (a) Proposition 12.3 shows that

$$H^1(G, k^{\mathrm{al}\times}) = \varinjlim_H H^1(\mathrm{Gal}(K/k), K^\times) = 0.$$

(b) For a field L, let $\mu_n(L) = \{\zeta \in L^\times \mid \zeta^n = 1\}$. From the exact sequence

$$1 \to \mu_n(k^{\mathrm{al}}) \to k^{\mathrm{al}\times} \xrightarrow{n} k^{\mathrm{al}\times} \to 1$$

we obtain an exact sequence of cohomology groups

$$1 \to \mu_n(k) \to k^\times \xrightarrow{n} k^\times \to H^1(G, \mu_n(k^{\mathrm{al}}) \to 1,$$

and hence a canonical isomorphism $H^1(G, \mu_n(k^{\mathrm{al}})) \xrightarrow{\approx} k^\times/k^{\times n}$. Note that for $k = \mathbb{Q}$, $\mathbb{Q}^\times/\mathbb{Q}^{\times n}$ is infinite if $n > 1$. For example, the numbers

$$(-1)^{\varepsilon(\infty)} \prod_{p \text{ prime}} p^{\varepsilon(p)},$$

where $\varepsilon(p) = 0$ or 1 and all but finitely many are zero, form a set of representatives for the elements of $\mathbb{Q}^\times/\mathbb{Q}^{\times 2}$, which is therefore an infinite-dimensional vector space over $\mathbb{F}_2$.

(c) If G acts trivially on M, then $H^1(G, M)$ is the set of continuous homomorphisms $G \to M$. This set can be identified with the set of pairs (K, α) consisting of a finite Galois extension K of k contained in k^{al} and an injective homomorphism $\alpha : \mathrm{Gal}(K/k) \to M$.

For an elliptic curve E over k, we abbreviate $H^i(\mathrm{Gal}(k^{\mathrm{al}}/k), E(k^{\mathrm{al}}))$ to $H^i(k, E)$.

Now consider an elliptic curve E over $\mathbb{Q}$. Let $\mathbb{Q}^{\mathrm{al}}$ be the algebraic closure of $\mathbb{Q}$ in $\mathbb{C}$, and choose an algebraic closure $\mathbb{Q}_p^{\mathrm{al}}$ for $\mathbb{Q}_p$. The embedding $\mathbb{Q} \hookrightarrow \mathbb{Q}_p$ extends to an embedding $\mathbb{Q}^{\mathrm{al}} \hookrightarrow \mathbb{Q}_p^{\mathrm{al}}$,

$$\begin{tikzcd}
\mathbb{Q}^{\mathrm{al}} \arrow[r, hook] & \mathbb{Q}_p^{\mathrm{al}} \\
\mathbb{Q} \arrow[u] \arrow[r, hook] & \mathbb{Q}_p. \arrow[u]
\end{tikzcd}$$

The action of $\mathrm{Gal}(\mathbb{Q}_p^{\mathrm{al}}/\mathbb{Q}_p)$ on $\mathbb{Q}^{\mathrm{al}} \subset \mathbb{Q}_p^{\mathrm{al}}$ defines an inclusion

$$\mathrm{Gal}(\mathbb{Q}_p^{\mathrm{al}}/\mathbb{Q}_p) \hookrightarrow \mathrm{Gal}(\mathbb{Q}^{\mathrm{al}}/\mathbb{Q}).$$

Hence any crossed homomorphism

$$\mathrm{Gal}(\mathbb{Q}^{\mathrm{al}}/\mathbb{Q}) \to E(\mathbb{Q}^{\mathrm{al}})$$

induces (by composition) a crossed homomorphism

$$\mathrm{Gal}(\mathbb{Q}_p^{\mathrm{al}}/\mathbb{Q}) \to E(\mathbb{Q}_p^{\mathrm{al}}).$$

In this way, we obtain a homomorphism

$$H^1(\mathbb{Q},E)\to H^1(\mathbb{Q}_p,E)$$

that (slightly surprisingly) is independent of the choice of the embedding $\mathbb{Q}^{\mathrm{al}}\hookrightarrow\mathbb{Q}_p^{\mathrm{al}}$. A similar remark applies to the cohomology groups of μ_n and E_n. Later we'll give a more natural interpretation of these "localization" homomorphisms.

13. The Selmer and Tate-Shafarevich groups

Lemma 13.1. *For any elliptic curve E over an algebraically closed field k and any integer n, the map $n: E(k)\to E(k)$ is surjective.*

Proof. The simplest proof uses algebraic geometry. The map of varieties $n: E\to E$ has finite fibres (because $E(k)_n$ is finite and it is a homomorphism) and has Zariski-closed image (because E is complete) of dimension one (because its fibres have dimension 0). Hence it is surjective as a morphism of algebraic varieties.

Alternatively, let $P=(x:y:1)\in E(k)$. To find a point $Q=(x':y':1)$ such that $nQ=P$ one has to solve a pair of polynomial equations in the variables X,Y. In characteristic zero, these equations can't be inconsistent, because $n: E(\mathbb{C})\to E(\mathbb{C})$ is surjective, and so, by the Hilbert Nullstellensatz, they have a solution in k. In characteristic p one has to work a little harder. □

From the lemma we obtain an exact sequence

$$0\to E_n(\mathbb{Q}^{\mathrm{al}})\to E(\mathbb{Q}^{\mathrm{al}})\xrightarrow{n}E(\mathbb{Q}^{\mathrm{al}})\to 0$$

and a cohomology sequence

$$0\to E_n(\mathbb{Q})\to E(\mathbb{Q})\xrightarrow{n}E(\mathbb{Q})\to H^1(\mathbb{Q},E_n)\to H^1(\mathbb{Q},E)\xrightarrow{n}H^1(\mathbb{Q},E),$$

from which we extract the sequence

$$\boxed{0\to E(\mathbb{Q})/nE(\mathbb{Q})\to H^1(\mathbb{Q},E_n)\to H^1(\mathbb{Q},E)_n\to 0.}$$

Here, as usual, $H^1(\mathbb{Q},E)_n$ is the group of elements in $H^1(\mathbb{Q},E)_n$ killed by n. If $H^1(\mathbb{Q},E_n)$ were finite, then we could deduce that $E(\mathbb{Q})/nE(\mathbb{Q})$ is finite, but unfortunately, it isn't. Instead, we proceed as follows. When we consider E as an elliptic curve over $\mathbb{Q}_p$ we obtain a similar exact sequence, and there is a commutative diagram:

$$\begin{tikzcd}
0 \arrow[r] & E(\mathbb{Q})/nE(\mathbb{Q}) \arrow[r] \arrow[d] & H^1(\mathbb{Q},E_n) \arrow[r] \arrow[d] & H^1(\mathbb{Q},E)_n \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & E(\mathbb{Q}_p)/nE(\mathbb{Q}_p) \arrow[r] & H^1(\mathbb{Q}_p,E_n) \arrow[r] & H^1(\mathbb{Q}_p,E)_n \arrow[r] & 0.
\end{tikzcd}$$

We want replace $H^1(\mathbb{Q},E_n)$ by a subset that contains the image of $E(\mathbb{Q})/nE(\mathbb{Q})$ but which we'll be able to prove finite. We do this as follows: if $\gamma\in H^1(\mathbb{Q},E_n)$ comes from an element of $E(\mathbb{Q})$, then certainly its image γ_p in $H^1(\mathbb{Q}_p,E_n)$ comes from an element of $E(\mathbb{Q}_p)$. This suggests defining[15]

$$\begin{aligned} S^{(n)}(E/\mathbb{Q}) &= \{\gamma\in H^1(\mathbb{Q},E_n)\mid \forall p,\ \gamma_p \text{ comes from } E(\mathbb{Q}_p)\}\\ &= \operatorname{Ker}(H^1(\mathbb{Q},E_n)\to\prod_p H^1(\mathbb{Q}_p,E)).\end{aligned}$$

[15] In the definitions of both the Selmer and Tate-Shafarevich groups, we should require that the elements become zero also in $H^1(\mathbb{R},E)$. We ignore this for the present.

The group $S^{(n)}(E/\mathbb{Q})$ is called the *Selmer group.* In the same spirit, we define[16] the *Tate-Shafarevich group* to be

$$\mathrm{TS}(E/\mathbb{Q}) = \mathrm{Ker}(H^1(\mathbb{Q}, E) \to \prod_p H^1(\mathbb{Q}_p, E)).$$

It is a torsion group. Later we shall give a geometric interpretation of $\mathrm{TS}(E/\mathbb{Q})$ which shows that it provides a measure of the failure of the Hasse principle for curves of genus 1. One can similarly define Selmer and Tate-Shafarevich groups for elliptic curves over number fields.

The next lemma is as trivial to prove as it is useful.

Lemma 13.2. *From any pair of maps of abelian groups (or modules etc.)*

$$A \xrightarrow{\alpha} B \xrightarrow{\beta} C$$

there is an exact sequence

$$0 \to \mathrm{Ker}(\alpha) \to \mathrm{Ker}(\beta \circ \alpha) \xrightarrow{\alpha} \mathrm{Ker}(\beta) \to \mathrm{Coker}(\alpha) \to \mathrm{Coker}(\beta \circ \alpha) \to \mathrm{Coker}(\beta) \to 0.$$

When we apply the lemma to the maps

$$H^1(\mathbb{Q}, E_n) \to H^1(\mathbb{Q}, E)_n \to \prod_p H^1(\mathbb{Q}_p, E)_n,$$

we obtain the fundamental exact sequence

$$\boxed{0 \to E(\mathbb{Q})/nE(\mathbb{Q}) \to S^{(n)}(E/\mathbb{Q}) \to \mathrm{TS}(E/\mathbb{Q})_n \to 0.}$$

We shall prove $E(\mathbb{Q})/nE(\mathbb{Q})$ to be finite by showing that $S^{(n)}(E/\mathbb{Q})$ is finite.

14. The Finiteness of the Selmer Group

Theorem 14.1. *For any elliptic curve E over a number field L and any integer n, the Selmer group $S^{(n)}(E/L)$ is finite (and, in fact, computable).*

Lemma 14.2. *Let E be an elliptic curve over $\mathbb{Q}_p$ with good reduction, and let n be an integer not divisible by p. A point $P \in E(\mathbb{Q}_p)$ is of the form nQ for some $Q \in E(\mathbb{Q}_p)$ if and only if $\bar{P} \in E(\mathbb{F}_p)$ is of the form $n\bar{Q}$ for some $\bar{Q} \in E(\mathbb{F}_p)$.*

Proof. Clearly $P = nQ \implies \bar{P} = n\bar{Q}$. For the converse, we make use of the filtration defined in Section 7:

$$E(\mathbb{Q}_p) \supset E^1(\mathbb{Q}_p) \supset \cdots \supset E^n(\mathbb{Q}_p) \supset E^{n+1}(\mathbb{Q}_p) \supset \cdots,$$

$$E(\mathbb{Q}_p)/E^1(\mathbb{Q}_p) \cong \bar{E}(\mathbb{F}_p), \quad E^n(\mathbb{Q}_p)/E^{n+1}(\mathbb{Q}_p) \cong \mathbb{F}_p.$$

By hypothesis there exists a $Q_0 \in E(\mathbb{Q}_p)$ such that

$$nQ_0 \equiv P \mod E^1(\mathbb{Q}_p).$$

Consider $P - nQ_0 \in E^1(\mathbb{Q}_p)$. Because $E^1(\mathbb{Q}_p)/E^2(\mathbb{Q}_p) \approx \mathbb{F}_p$ and p doesn't divide n, multiplication by n is an isomorphism on $E^1(\mathbb{Q}_p)/E^2(\mathbb{Q}_p)$. Therefore there exists a $Q_1 \in E^1(\mathbb{Q}_p)$ such that

$$P - nQ_0 = nQ_1 \mod E^2(\mathbb{Q}_p).$$

[16] The name, I believe, is due to Cassels, who certainly knows the alphabet. Recently, it has become fashionable to reverse the order of the names. In the absence of an argument for doing this, I prefer to follow Cassels. In the original, the initial Russian letter of Shafarevich was used for TS.

Continuing in this fashion, we find a sequence $Q_0, Q_1, \dots$ of points in $E(\mathbb{Q}_p)$ such that

$$Q_i \in E^i(\mathbb{Q}_p), \quad P - n\sum_{i=0}^{m} Q_i \in E^{m+1}(\mathbb{Q}_p).$$

The first condition implies that $\sum Q_i$ converges to a point in $E(\mathbb{Q}_p)$ (recall that $E(\mathbb{Q}_p)$ is compact), and the second condition implies that its limit Q has the property that $P = nQ$. □

We now need a result from algebraic number theory.

Lemma 14.3. *For any finite extension k of $\mathbb{F}_p$, there exists an extension K of $\mathbb{Q}_p$ with the following properties:*

(a) $[K : \mathbb{Q}_p] = [k : \mathbb{F}_p]$*;*
(b) *the integral closure R of $\mathbb{Z}_p$ in K is a principal ideal domain with p as its only prime element (up to associates), and $R/pR = k$.*

Proof. Omitted. □

The field K in the lemma is unique (up to a unique isomorphism inducing the identity map on the residue fields). It is called the *unramified extension of $\mathbb{Q}_p$ with residue field k.*

Because R is a principal ideal domain with p as its only prime element and K is the field of fractions of R, every element α in K can be written uniquely in the form up^m with $u \in R^\times$. Define $\mathrm{ord}_p(\alpha) = m$. Then ord_p is a homomorphism $K^\times \to \mathbb{Z}$ extending $\mathrm{ord}_p : \mathbb{Q}^\times \to \mathbb{Z}$.

Remark 14.4. Let $K \supset R \to k$ be as in the lemma. Then pR is the unique maximal ideal of R, and Hensel's lemma (Theorem 2.8) holds for R, and so all the roots of $X^q - X$ in k lift to R. Therefore, K contains the splitting field of $X^q - X$, and, in fact, is equal to it.

The theory in Section 7 holds, word for word, with $\mathbb{Q}_p$ replaced by an unramified extension[17] K, except that now

$$E(K)/E^1(K) \cong \bar{E}(k), \quad E^n(K)/E^{n+1}(K) \cong k.$$

Therefore, Lemma 14.2 is also valid with $\mathbb{Q}_p$ replaced by K.

Consider an elliptic curve E over $\mathbb{Q}_p$ and an n satisfying the hypotheses of Lemma 14.2. Let $P \in E(\mathbb{Q}_p)$. According to Lemma 14.2, $P = nQ$ for some Q with coordinates in a field $K \supset \mathbb{Q}_p$, which we may choose to be of finite degree over $\mathbb{Q}_p$, and which (the generalization of) (14.2) allows us to take to be unramified over $\mathbb{Q}_p$. We have proved:

Lemma 14.5. *Let E and n satisfy the hypotheses of Lemma 14.2, and let $P \in E(\mathbb{Q}_p)$. Then there exists a finite unramified extension K of $\mathbb{Q}_p$ such that $P \in nE(K)$.*

Proposition 14.6. *Let E be an elliptic curve over $\mathbb{Q}$, and let T be the set of primes dividing $2n\Delta$. For any $\gamma \in S^{(n)}(\mathbb{Q})$ and any $p \notin T$, there exists a finite unramified extension K of $\mathbb{Q}_p$ such that γ maps to zero in $H^1(K, E_n)$.*

[17] In fact, it holds even for a ramified extension.

Proof. From the definition of the Selmer group, we know that there exists a $P \in E(\mathbb{Q}_p)$ mapping to $\gamma_p \in H^1(\mathbb{Q}_p, E_n)$. Since p does not divide 2Δ, E has good reduction at p, and so there is an unramified extension K of $\mathbb{Q}_p$ such that $P \in nE(K)$. Now the following diagram shows that γ maps to zero in $H^1(K, E_n)$:

$$\begin{tikzcd}
E(\mathbb{Q}) \arrow[r, "n"] \arrow[d] & E(\mathbb{Q}) \arrow[r] \arrow[d] & H^1(\mathbb{Q}, E_n) \arrow[d] \\
E(\mathbb{Q}_p) \arrow[r, "n"] \arrow[d] & E(\mathbb{Q}_p) \arrow[r] \arrow[d] & H^1(\mathbb{Q}_p, E_n) \arrow[d] \\
E(K) \arrow[r, "n"] & E(K) \arrow[r] & H^1(K, E_n).
\end{tikzcd}$$

$\square$

Proof of the finiteness of the Selmer group in a special case. We prove that $S^{(2)}(E/\mathbb{Q})$ is finite in the case that the points of order 2 on E have coordinates in $\mathbb{Q}$. This condition means that the equation for E has the form:

$$Y^2Z = (X - \alpha Z)(X - \beta Z)(X - \gamma Z), \quad \alpha, \beta, \gamma \in \mathbb{Q}.$$

It implies that

$$E_2(\mathbb{Q}^{\mathrm{al}}) = E_2(\mathbb{Q}) \approx (\mathbb{Z}/2\mathbb{Z})^2 = (\mu_2)^2,$$

all with the trivial action of $\mathrm{Gal}(\mathbb{Q}^{\mathrm{al}}/\mathbb{Q})$, and so

$$H^1(\mathbb{Q}, E_2) \approx H^1(\mathbb{Q}, \mu_2)^2 = (\mathbb{Q}^\times/\mathbb{Q}^{\times 2})^2.$$

Let $\gamma \in S^{(2)}(E/\mathbb{Q}) \subset H^1(\mathbb{Q}, E_2)$. For any prime p_0 not dividing 2Δ, there exists a finite unramified extension K of $\mathbb{Q}_{p_0}$ such that γ maps to zero under the vertical arrows:

$$\begin{tikzcd}
H^1(\mathbb{Q}, E_2) \arrow[r, phantom, "\approx"] \arrow[d] & (\mathbb{Q}^\times/\mathbb{Q}^{\times 2})^2 \arrow[d] \\
H^1(K, E_2) \arrow[r, phantom, "\approx"] & (K^\times/K^{\times 2})^2.
\end{tikzcd}$$

Suppose

$$\gamma \leftrightarrow ((-1)^{\varepsilon(\infty)} \prod p^{\varepsilon(p)}, (-1)^{\varepsilon'(\infty)} \prod p^{\varepsilon'(p)}), \quad 0 \leq \varepsilon(p), \varepsilon'(p) \leq 1.$$

Now

$$\mathrm{ord}_{p_0}\left((-1)^{\varepsilon(\infty)} \prod p^{\varepsilon(p)}\right) = \varepsilon(p_0),$$

and so if $(-1)^{\varepsilon(\infty)} \prod p^{\varepsilon(p)}$ is a square in K, then $\varepsilon(p_0) = 0$. Therefore the only p that can occur in the factorizations are those dividing 2Δ. This allows only finitely many possibilities for γ.

Remark 14.7. It is possible to prove that $E(\mathbb{Q})/2E(\mathbb{Q})$ is finite in this case without mentioning cohomology groups. Consider an elliptic curve

$$Y^2Z = (X - \alpha Z)(X - \beta Z)(X - \gamma Z), \quad \alpha, \beta, \gamma \in \mathbb{Z}.$$

Define $\varphi_\alpha : E(\mathbb{Q})/2E(\mathbb{Q}) \to \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ by

$$\varphi_\alpha((x:y:z)) = \begin{cases} (x/z - \alpha)\mathbb{Q}^{\times 2} & z \neq 0, \quad x \neq \alpha z; \\ (\alpha - \beta)(\alpha - \gamma)\mathbb{Q}^\times & z \neq 0, \quad x = \alpha z \\ \mathbb{Q}^\times & (x:y:z) = (0:1:0). \end{cases}$$

One can prove directly that φ_α is a homomorphism, that the kernel of $(\varphi_\alpha, \varphi_\beta) : E(\mathbb{Q}) \to (\mathbb{Q}^\times/\mathbb{Q}^{\times 2})^2$ is $2E(\mathbb{Q})$, and that $\varphi_\alpha(P)$ and $\varphi_\beta(P)$ are represented by ± a product of primes dividing 2Δ (see [Kn] pp85–91).

Proof of the finiteness of the Selmer group in the general case. In the above proof we made use of the following facts:

(a) $\mathbb{Q}$ contains a primitive square root of 1;
(b) $E(\mathbb{Q})_2 = E(\mathbb{Q}^{\mathrm{al}})_2$;
(c) for any finite set T of prime numbers, the kernel of

$$r \mapsto (\operatorname{ord}_p(r) \mod 2) : \mathbb{Q}^\times/\mathbb{Q}^{\times 2} \to \bigoplus_{p \notin T} \mathbb{Z}/2\mathbb{Z}$$

is finite.

For some finite extension L of $\mathbb{Q}$, L will contain a primitive nth root of 1 and $E(L)$ will contain all the points of order n on $E(\mathbb{Q}^{\mathrm{al}})$. The next lemma shows that, in order to prove that $S^{(n)}(E/\mathbb{Q})$ is finite, it suffices to prove that $S^{(n)}(E/L)$ is finite.

Lemma 14.8. *For any finite Galois extension L of $\mathbb{Q}$ and any n, the kernel of*

$$S^{(n)}(E/\mathbb{Q}) \to S^{(n)}(E/L)$$

is finite.

Proof. Since $S^{(n)}(E/\mathbb{Q})$ and $S^{(n)}(E/L)$ are subgroups of $H^1(\mathbb{Q}, E_n)$ and $H^1(L, E_n)$ respectively, it suffices to prove that the kernel of

$$H^1(\mathbb{Q}, E_n) \to H^1(L, E_n)$$

is finite. But (cf. 12.6), this kernel is $H^1(\mathrm{Gal}(L/\mathbb{Q}), E_n(L))$, which is finite because both $\mathrm{Gal}(L/\mathbb{Q})$ and $E_n(L)$ are finite. □

It remains to consider (c). The proof of its analogue for L requires the three fundamental theorems in any course on algebraic number theory. We review their statements.

Review of algebraic number theory. In the following, L is a finite extension of $\mathbb{Q}$ and R is the ring of all algebraic integers in L (see p53).

Every element of R is a product of irreducible (i.e., "unfactorable") elements, but this factorization may not be unique. For example, in $\mathbb{Z}[\sqrt{-5}]$ we have

$$6 = 2 \cdot 3 = (1+\sqrt{-5})(1-\sqrt{-5})$$

and 2, 3, $1+\sqrt{-5}$, $1-\sqrt{-5}$ are irreducible with no two associates. The idea of Kummer and Dedekind to remedy this problem was to enlarge the set of numbers with "ideal numbers", now called ideals, to recover unique factorization. For ideals $\mathfrak{a}$ and $\mathfrak{b}$, define

$$\mathfrak{a}\mathfrak{b} = \{\sum a_i b_i \mid a_i \in \mathfrak{a}, \quad b_i \in \mathfrak{b}\}.$$

It is again is an ideal.

Theorem 14.9 (Dedekind). *Every ideal in R can be written uniquely as a product of prime ideals.*

For example, in $\mathbb{Z}[\sqrt{-5}]$,

$$(6) = (2, 1+\sqrt{-5})(2, 1-\sqrt{-5})(3, 1+\sqrt{-5})(3, 1-\sqrt{-5}).$$

For an element $a \in R$ and a prime ideal $\mathfrak{p}$ in R, let $\mathrm{ord}_{\mathfrak{p}}(a)$ be the exponent of $\mathfrak{p}$ in the unique factorization of the ideal (a), so that

$$(a) = \prod_{\mathfrak{p}} \mathfrak{p}^{\mathrm{ord}_{\mathfrak{p}}(a)}.$$

For $x = \frac{a}{b} \in L$, define $\mathrm{ord}_{\mathfrak{p}}(x) = \mathrm{ord}_{\mathfrak{p}}(a) - \mathrm{ord}_{\mathfrak{p}}(b)$. The *ideal class group* C of R is defined to be the cokernel of the homomorphism

$$\begin{array}{ccccccc} L^{\times} & \to & \bigoplus_{\mathfrak{p}\subset R, \mathfrak{p} \text{ prime}} \mathbb{Z} & \to & C & \to & 0 \\ x & \mapsto & (\mathrm{ord}_{\mathfrak{p}}(x)). & & & & \end{array}$$

It is 0 if and only if R is a principal ideal domain, and so C can be regarded as giving a measure of the failure of unique factorization of elements in R.

Theorem 14.10 (Finiteness of the class number). *The ideal class group C is finite.*

We next need to understand the group U of units in R. For $R = \mathbb{Z}$, $U = \{\pm 1\}$, but already for $R = \mathbb{Z}[\sqrt{2}]$, U is infinite because $\sqrt{2}+1$ is a unit in $\mathbb{Z}[\sqrt{2}]$. One can show that

$$\mathbb{Z}[\sqrt{2}]^{\times} = \{\pm(1+\sqrt{2})^n \mid n \in \mathbb{Z}\} \approx \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}.$$

Theorem 14.11 (Dedekind unit theorem). *The group U of units of R is finitely generated.*

In fact, the full theorem gives a formula for the rank of U.

As in any commutative ring, a is a unit in R if and only if $(a) = R$. In our case, this is equivalent to saying that $\mathrm{ord}_{\mathfrak{p}}(a) = 0$ for all prime ideals $\mathfrak{p}$, and so we have an exact sequence

$$0 \to U \to L^{\times} \to \oplus_{\mathfrak{p}}\mathbb{Z} \to C \to 0$$

with U finitely generated and C finite.

The fundamental theorems of algebraic number theory show, more generally, that, when T is a finite set of prime ideals in L, the groups U_T and C_T defined by the exactness of

$$0 \to U_T \to L^{\times} \xrightarrow{a \mapsto (\mathrm{ord}_{\mathfrak{p}}(a))} \oplus_{\mathfrak{p}\notin T}\mathbb{Z} \to C_T \to 0$$

are, respectively, finitely generated and finite.

Completion of the proof of the finiteness of the Selmer group.

Lemma 14.12. *Let N be the kernel of*

$$a \mapsto (ord_{\mathfrak{p}}(a) \mod n) : \mathrm{Ker}(L^{\times}/L^{\times n}) \to \oplus_{\mathfrak{p}\notin T}\mathbb{Z}/n\mathbb{Z}).$$

Then there is an exact sequence

$$0 \to U_T/U_T^n \to N \to (C_T)_n$$

Proof. Let $\alpha \in N$. Then $n|\mathrm{ord}_{\mathfrak{p}}(\alpha)$ for all $\mathfrak{p} \notin T$, and so we can map α to the class c of $(\frac{\mathrm{ord}_{\mathfrak{p}}(\alpha)}{n})$ in C_T. Clearly $nc = 0$, and any element of C_T killed by n arises in this way. If $c = 0$, then there exists a $\beta \in L^{\times}$ such that $\mathrm{ord}_{\mathfrak{p}}(\beta) = \mathrm{ord}_{\mathfrak{p}}(\alpha)/n$ for all $\mathfrak{p}$. Now α/β^n lies in U_T, and is well-defined up to an element of U_T^n. □

Now the argument used in the special case shows that $S^{(n)}(E/L)$ is finite.

Remark 14.13. The above proof of the finiteness of the Selmer group is taken from my book, Etale Cohomology, p133. It is simpler than the standard proof (see [S1] p190–196) which unnecessarily "translate[s] the putative finiteness of $E(K)/mE(K)$ into a statement about certain field extensions of K."

15. Heights

Let $P = (a_0 : \ldots : a_n) \in \mathbb{P}^n(\mathbb{Q})$. We shall say that $(a_0 : \ldots : a_n)$ is a *primitive representative* for P if

$$a_i \in \mathbb{Z}, \quad \gcd(a_0, \ldots, a_n) = 1.$$

The *height* $H(P)$ of P is then defined to be

$$H(P) = \max_j |a_i|.$$

Here $| * |$ is the usual absolute value. The *logarithmic height* $h(P)$ of P is defined to be $\log H(P)$.

Heights on $\mathbb{P}^1$. Let $F(X,Y)$ and $G(X,Y)$ be homogeneous polynomials of degree m in $\mathbb{Q}[X,Y]$, and let $V(\mathbb{Q})$ be the set of their common zeros. Then F and G define a map

$$\varphi : \mathbb{P}^1(\mathbb{Q}) \setminus V(\mathbb{Q}) \to \mathbb{P}^1(\mathbb{Q}), \quad (x : y) \mapsto (F(x,y) : G(x,y)).$$

Proposition 15.1. *If $F(X,Y)$ and $G(X,Y)$ have no common zero in $\mathbb{P}^1(\mathbb{Q}^{al})$, then there exists a constant B such that*

$$|h(\varphi(P)) - mh(P)| \leq B, \quad \text{all } P \in \mathbb{P}^1(\mathbb{Q}).$$

Proof. We may suppose that F and G have integer coefficients. Let $(a : b)$ be a primitive representative for P. Then, for a monomial $H(X,Y) = cX^iY^{m-i}$, $|H(a,b)| \leq |c| \max(|a|^m, |b|^m)$, and so

$$|F(a,b)|, |G(a,b)| \leq C \left(\max(|a|, |b|\right)^m$$

with

$$C = (m+1)\max(|\text{coeff. of } F \text{ or } G|).$$

Now

$$H(\varphi(P)) \leq \max(|F(a,b)|, |G(a,b)|) \leq C(\max(|a|,|b|)^m = C \cdot H(P)^m.$$

On taking logs, we obtain the inequality

$$h(\varphi(P)) \leq mh(P) + \log C.$$

The problem with proving a reverse inequality is that $F(a,b)$ and $G(a,b)$ may have a large common factor, and so the first inequality in the second last equation may be strict. We use the hypothesis that F and G have no common zero in $\mathbb{Q}^{\mathrm{al}}$ to limit this problem.

Let R be the resultant of F and G—the hypothesis says that $R \neq 0$. Consider $Y^{-m}F(X,Y) = F(\frac{X}{Y}, 1)$ and $Y^{-m}G(X,Y) = G(\frac{X}{Y}, 1)$. When regarded as polynomials in the single variable $\frac{X}{Y}$, $F(\frac{X}{Y}, 1)$ and $G(\frac{X}{Y}, 1)$ have the same resultant as $F(X,Y)$ and $G(X,Y)$, and so (see p55), there are polynomials $U(\frac{X}{Y})$, $V(\frac{X}{Y}) \in \mathbb{Z}[\frac{X}{Y}]$ of degree $m-1$ such that

$$U(\frac{X}{Y})F(\frac{X}{Y}, 1) + V(\frac{X}{Y})G(\frac{X}{Y}, 1) = R.$$

On multiplying through by Y^{2m-1} and renaming $Y^{m-1}U(\frac{X}{Y})$ as $U(X,Y)$ and $Y^{m-1}V(\frac{X}{Y})$ as $V(X,Y)$, we obtain the equation

$$U(X,Y)F(X,Y)+V(X,Y)G(X,Y)=RY^{2m-1}.$$

Similarly, there are homogenous polynomials $U'(X,Y)$ and $V'(X,Y)$ of degree $m-1$ such that

$$U'(X,Y)F(X,Y)+V'(X,Y)G(X,Y)=RX^{2m-1}.$$

Substitute (a,b) for (X,Y) to obtain the equations

$$U(a,b)F(a,b)+V(a,b)G(a,b)=Rb^{2m-1},$$

$$U'(a,b)F(a,b)+V'(a,b)G(a,b)=Ra^{2m-1}.$$

From these equations we see that

$$\gcd(F(a,b),G(a,b)) \text{ divides } \gcd(Ra^{2m-1},Rb^{2m-1})=R.$$

Moreover, as in the first part of the proof, there is a $C>0$ such that

$$U(a,b),U'(a,b),V(a,b),V'(a,b)\leq C\left(\max|a|,|b|\right)^{m-1}.$$

Therefore

$$2C\left(\max|a|,|b|\right)^{m-1}\left(\max|F(a,b)|,|G(a,b)|\right)\geq|R||a|^{2m-1},\ |R||b|^{2m-1}.$$

Together with $\gcd(F(a,b),G(a,b))|R$, these inequalities imply that

$$H(\varphi(P))\geq\frac{1}{|R|}\max(|F(a,b)|,|G(a,b)|)\geq\frac{1}{2C}H(P)^m.$$

On taking logs, we obtain the inequality

$$h(\varphi(P))\geq mh(P)-\log 2C.$$

□

There is a well-defined map (special case of the Veronese map)

$$(a:b),(c:d)\mapsto(ac:ad+bc:bd):\mathbb{P}^1\times\mathbb{P}^1\to\mathbb{P}^2.$$

Let R be the image of (P,Q).

Lemma 15.2.

$$\frac{1}{2}\leq\frac{H(R)}{H(P)H(Q)}\leq 2.$$

Proof. Choose $(a:b)$ and $(c:d)$ to be primitive representatives of P and Q. Then

$$H(R)\leq\max(|ac|,|ad+bc|,|bd|)\leq 2\max(|a|,|b|)\max(|c|,|d|)=2H(P)H(Q).$$

If a prime p divides both ac and bd, then either it divides a and d but not b or c, or the other way round. In either case, it doesn't divide $ad+bc$, and so $(ac:ad+bc:bd)$ is a primitive representative for R. It remains to show that

$$\max(|ac|,|ad+bc|,|bd|)\geq\frac{1}{2}\left(\max(|a|,|b|)\left(\max|c||d|\right),\right.$$

but this is an elementary exercise. □

Heights on E. Let E be the elliptic curve

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in \mathbb{Q}, \quad \Delta = 4a^3 + 27b^2 \neq 0.$$

For $P \in E(\mathbb{Q})$, define

$$H(P) = \begin{cases} H((x(P) : z(P))) & \text{if } z(P) \neq 0 \\ 0 & \text{if } P = (0 : 1 : 0). \end{cases}$$

and

$$h(P) = \log H(P).$$

Other definitions of h are possible, but they differ by bounded amounts, and therefore lead to the same canonical height (see below).

Lemma 15.3. *For any constant B, the set of $P \in E(\mathbb{Q})$ such that $h(P) < B$ is finite.*

Proof. Certainly, for any constant B, $\{P \in \mathbb{P}^1(\mathbb{Q}) \mid H(P) \leq B\}$ is finite. But for every point $(x_0 : z_0) \in \mathbb{P}^1(\mathbb{Q})$, there are at most two points $(x_0 : y : z_0) \in E(\mathbb{Q})$, and so $\{P \in E(\mathbb{Q}) \mid H(P) \leq B\}$ is finite. □

Proposition 15.4. *There exists a constant A such that*

$$|h(2P) - 4h(P)| \leq A.$$

Proof. Let $P = (x : y : z)$ and $2P = (x_2 : y_2 : z_2)$. According to the duplication formula (p37),

$$(x_2 : z_2) = (F(x, z) : G(x, z))$$

where $F(X, Z)$ and $G(X, Z)$ are polynomials of degree 4 such that

$$\begin{aligned} F(X, 1) &= (3X^2 + a)^2 - 8X(X^3 + ax + b) \\ G(X, 1) &= 4(X^3 + aX + b). \end{aligned}$$

Since $X^3 + aX + b$ and its derivative $3X^2 + b$ have no common root, neither do $F(X, 1)$ and $G(X, 1)$, and so Proposition 15.1 shows that

$$|h(2P) - 4h(P)| \leq A$$

for some constant A. □

Theorem 15.5. *There exists a unique function $\widehat{h} : E(\mathbb{Q}) \to \mathbb{R}$ satisfying the conditions (a) and (b):*

(a) *$\widehat{h}(P) - h(P)$ is bounded;*
(b) *$\widehat{h}(2P) = 4\widehat{h}(P)$.*

In fact,

$$\widehat{h}(P) = \lim_{n\to\infty} \frac{h(2^nP)}{4^n}$$

and it has the following additional properties:

(c) *for any $C \geq 0$, the set $\{P \in E(\mathbb{Q}) \mid \widehat{h}(P) \leq C\}$ is finite;*
(d) *$\widehat{h}(P) \geq 0$, with equality if and only if P has finite order.*

Proof. We first prove uniqueness. If h' satisfies (a) with bound B, then

$$|h'(2^nP) - h(2^nP)| \leq B.$$

If in addition it satisfies (b), then

$$\left| h'(P) - \frac{h(2^nP)}{4^n} \right| \leq \frac{B}{4^n},$$

and so $h(2^nP)/4^n$ converges to $h'(P)$.

To prove the existence, we first verify that $h(2^nP)/4^n$ is a Cauchy sequence. From Proposition 15.4, we know that there exists a constant A such that

$$|h(2P) - 4h(P)| \leq A$$

for all P. For $N \geq M \geq 0$ and $P_0 \in E(\mathbb{Q})$,

$$\begin{aligned}
\left| \frac{h(2^NP_0)}{4^N} - \frac{h(2^MP_0)}{4^M} \right| &= \left| \sum_{n=M}^{N-1} \left(\frac{h(2^{n+1}P_0)}{4^{n+1}} - \frac{h(2^nP_0)}{4^n} \right) \right| \\
&\leq \sum_{n=M}^{N-1} \frac{1}{4^{n+1}} |h(2^{n+1}P_0) - 4h(2^nP_0)| \\
&\leq \sum_{n=M}^{N-1} \frac{1}{4^{n+1}} A \\
&\leq \frac{A}{4^{M+1}} \left(1 + \frac{1}{4} + \frac{1}{4^2} + \cdots \right) \\
&= \frac{A}{3 \cdot 4^M}.
\end{aligned}$$

This shows that the sequence $h(2^nP)/4^n$ is Cauchy, and we define $\widehat{h}(P)$ to be its limit. Because $H(P)$ is an integer ≥ 1, $h(P) \geq 0$ and $\widehat{h}(P) \geq 0$.

When $M = 0$ the displayed equation becomes

$$\left| \frac{h(2^NP)}{4^N} - h(P) \right| \leq \frac{A}{3},$$

and on letting $N \to \infty$ we obtain (a).

For (b), note that

$$\widehat{h}(2P) = \lim_{n\to\infty} \frac{h(2^{n+1}P)}{4^n} = 4 \cdot \lim_{n\to\infty} \frac{h(2^{n+1}P)}{4^{n+1}} = 4 \cdot \widehat{h}(P).$$

The set of P for which $\widehat{h}(P) \leq C$ is finite, because h has this property and the difference $\widehat{h}(P) - h(P)$ is bounded.

If P is torsion, then $\{2^nP \mid n \geq 0\}$ is finite, so $\widehat{h}$ is bounded on it, by D say, and $\widehat{h}(P) = \widehat{h}(2^nP)/4^n \leq D/4^n$ for all n. On the other hand, if P has infinite order, then $\{2^nP \mid n \geq 0\}$ is infinite and $\widehat{h}$ is unbounded on it. Hence $\widehat{h}(2^nP) > 1$ for some n, and so $\widehat{h}(P) > 4^{-n} > 0$. □

The function $\widehat{h}$ is called[18] the *canonical*, or *Néron-Tate, height.* If was defined independently by Tate using the above method, and by Néron using a much more elaborate method which, however, gives more information about $\widehat{h}$.

Let $f : M \to K$ be a function from an abelian group M into a field K of characteristic $\neq 2$. Such an f is called a *quadratic form* if $f(2x) = 4f(x)$ and

$$B(x,y) =_{df} f(x+y) - f(x) - f(y)$$

is bi-additive. Then B is symmetric, and it is the only symmetric bi-additive form $B : M \times M \to K$ such that $f(x) = \frac{1}{2}B(x,x)$. We shall need the following criterion:

Lemma 15.6. *A function $f : M \to K$ from an abelian group into a field K of characteristic $\neq 2$ is a quadratic form if and only if it satisfies the parallelogram[19] law:*

$$f(x+y) + f(x-y) = 2f(x) + 2f(y) \quad \text{all } x, y \in M.$$

Proof. On taking $x = y = 0$ in the parallelogram law, we find that $f(0) = 0$, on taking $x = y$ we find that $f(2x) = 4f(x)$, and on taking $x = 0$ we find that $f(-y) = f(y)$. By symmetry, it remains to show that $B(x+y,z) = B(x,z) + B(y,z)$, i.e., that

$$f(x+y+z) - f(x+y) - f(x+z) - f(y+z) + f(x) + f(y) + f(z) = 0.$$

Now four applications of the parallelogram law show that:

$$\begin{gathered} f(x+y+z) + f(x+y-z) - 2f(x+y) - 2f(z) = 0 \\ f(x-y+z) + f(x+y-z) - 2f(x) - 2f(y-z) = 0 \\ f(x-y+z) + f(x+y+z) - 2f(x+z) - 2f(y) = 0 \\ 2f(y+z) + 2f(y-z) - 4f(y) - 4f(z) = 0. \end{gathered}$$

The alternating sum of these equations is the required equation. □

Proposition 15.7. *The height function $\widehat{h} : E(\mathbb{Q}) \to \mathbb{R}$ is a quadratic form.*

We have to prove the parallelogram law.

Lemma 15.8. *There exists a constant C such that*

$$H(P_1 + P_2)H(P_1 - P_2) \leq C \cdot H(P_1)^2 H(P_2)^2$$

for all $P_1, P_2 \in E(\mathbb{Q})$.

Proof. Let $P_1 + P_2 = P_3$ and $P_1 - P_2 = P_4$, and let $P_i = (x_i : y_i : z_i)$. Then

$$(x_3x_4 : x_3z_4 + x_4z_3 : z_3z_4) = (W_0 : W_1 : W_2)$$

where (see p37)

$$\begin{aligned} W_0 &= (X_2Z_1 - X_1Z_2)^2 \\ W_1 &= 2(X_1X_2 + aZ_1Z_2)(X_1Z_2 + X_2Z_1) + 4bZ_1^4Z_2^4 \\ W_2 &= X_1^2X_2^2 - 2aX_1X_2Z_1Z_2 - 4b(X_1Z_1Z_2^2 + X_2Z_1^2Z_2) + a^2Z_1^2Z_2^2. \end{aligned}$$

It follows that

$$H(W_0 : W_1 : W_2) \leq CH(P_1)^2H(P_2)^2.$$

[18] Unfortunately, there are different definitions of the "canonical" height, which differ by a constant factor.

[19] In elementary linear algebra, the parallelogram law says that, for vectors u and v in $\mathbb{R}^n$, $\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$.

According to Lemma 15.2,

$$H(W_0 : W_1 : W_2) \geq \frac{1}{2}H(P_3)H(P_4).$$

□

Lemma 15.9. *The canonical height function* $\widehat{h} : E(\mathbb{Q}) \to \mathbb{R}$ *satisfies the parallelogram law:*

$$\widehat{h}(P+Q) + \widehat{h}(P-Q) = 2\widehat{h}(P) + 2\widehat{h}(Q).$$

Proof. On taking logs in the previous lemma, we find that

$$h(P+Q) + h(P-Q) \leq 2h(P) + 2h(Q) + B.$$

On replacing P and Q with 2^nP and 2^nQ, dividing through by 4^n, and letting $n \to \infty$, we obtain the inequality

$$\widehat{h}(P+Q) + \widehat{h}(P-Q) \leq 2\widehat{h}(P) + 2\widehat{h}(Q).$$

Putting $P' = P+Q$ and $Q' = P-Q$ in this gives the reverse inequality:

$$\widehat{h}(P') + \widehat{h}(Q') \leq 2\widehat{h}\left(\frac{P'+Q'}{2}\right) + 2\widehat{h}\left(\frac{P'-Q'}{2}\right) = \frac{1}{2}\widehat{h}(P'+Q') + \frac{1}{2}\widehat{h}(P'-Q').$$

□

Aside 15.10 (For the experts). Let K be a number field. For each prime v of K, let $|\cdot|_v$ be the normalized valuation, for which the product formula holds:

$$\prod_v |a|_v = 1, \quad a \in K^\times.$$

Define the height of a point $P = (a_0 : a_1 : \ldots : a_n) \in \mathbb{P}^n(K)$ to be

$$H(P) = \prod_v \max_i(|a_i|_v).$$

Because of the product formula, $H(P)$ doesn't depend on the choice of $(a_0 : \ldots : a_n)$ representing P. When $K = \mathbb{Q}$, we can choose the a_i to be integers with no common factor, which makes $\max_i |a_i|_p = 1$ for all p, and leaves $H(P) = \max_i |a_i|_\infty$.

With this definition, all the above results extend to elliptic curves over number fields.

16. Completion of the Proof of the Mordell-Weil Theorem, and Further Remarks

Let $P_1, \ldots, P_s$ be a set of representatives for $E(\mathbb{Q})/2E(\mathbb{Q})$. For any $Q \in E(\mathbb{Q})$ there exists an i such that $Q \pm P_i \in 2E(\mathbb{Q})$ for both choices of signs. According to the parallelogram law,

$$h(Q \pm P_i) \leq h(Q) + h(P_i)$$

for (at least) one choice of signs. For that choice, let $Q \pm P_i = 2Q'$. Then

$$4h(Q') = h(Q \pm P_i) \leq h(Q) + h(P_i) \leq h(Q) + C$$

where $C = \max h(P_i)$. Hence $h(Q') < \frac{1}{2}h(Q)$ provided $h(Q) > C$. Now the argument sketched in Section 11 shows that $E(\mathbb{Q})$ is generated by $P_1, \ldots, P_s$ and the Q with $h(Q) \leq C$.

The Problem of Computing the Rank of $E(\mathbb{Q})$. According to André Weil, one of the two oldest outstanding problems in mathematics is that of determining the group $E(\mathbb{Q})$. We know that $E(\mathbb{Q})$ is finitely generated, say $E(\mathbb{Q}) \approx E(\mathbb{Q})_{\text{tors}} \oplus \mathbb{Z}^r$, and the problem is to find an algorithm for determining r, or better, for finding a set of generators for $E(\mathbb{Q})/E(\mathbb{Q})_{\text{tors}}$. Since we know how to compute $E(\mathbb{Q})_{\text{tors}}$, this amounts to being able to find a basis for $E(\mathbb{Q})/2E(\mathbb{Q})$. We can regard $S^{(2)}(E/\mathbb{Q})$ as giving a computable upper bound for r, with $\text{TS}(E/\mathbb{Q})_2$ as the error. The problem is to determine the image of $E(\mathbb{Q})$ in $S^{(2)}(\mathbb{Q})$.

Consider the following commutative diagram:

$$\begin{tikzcd}
0 \arrow[r] & E(\mathbb{Q})/2E(\mathbb{Q}) \arrow[r] & S^{(2)}(E/\mathbb{Q}) \arrow[r] & \text{TS}(E/\mathbb{Q})_2 \arrow[r] & 0 \\
0 \arrow[r] & E(\mathbb{Q})/4E(\mathbb{Q}) \arrow[r] \arrow[u] & S^{(4)}(E/\mathbb{Q}) \arrow[r] \arrow[u] & \text{TS}(E/\mathbb{Q})_4 \arrow[r] \arrow[u, "2"'] & 0 \\
 & \vdots \arrow[u] & \vdots \arrow[u] & \vdots \arrow[u, "2"'] & \\
0 \arrow[r] & E(\mathbb{Q})/2^nE(\mathbb{Q}) \arrow[r] \arrow[u] & S^{(2^n)}(E/\mathbb{Q}) \arrow[r] \arrow[u] & \text{TS}(E/\mathbb{Q})_{2^n} \arrow[r] \arrow[u, "2"'] & 0.
\end{tikzcd}$$

Define $S^{(2,n)}(E/\mathbb{Q})$ to be the image of $S^{(2^n)}(E/\mathbb{Q})$ in $S^{(2)}(E/\mathbb{Q})$.

Proposition 16.1. *The group*

$$E(\mathbb{Q})/2E(\mathbb{Q}) \subset \cap_n S^{(2,n)}(E/\mathbb{Q}),$$

and is equal to it if and only if TS$(E/\mathbb{Q})$ contains no nonzero element divisible by all powers of 2, in which case $S^{(2,n)}(E/\mathbb{Q})$ is constant for sufficiently large n.

Proof. Clearly the image of $E(\mathbb{Q})/2^nE(\mathbb{Q})$ in $S^{(2)}(E/\mathbb{Q})$ is independent of n, and is contained in $S^{(2,n)}(E/\mathbb{Q})$ for all n. Conversely, let $\gamma \in \cap S^{(2,n)}(E/\mathbb{Q})$. By definition, there is, for each n, an element $\gamma_n \in S^{(2^n)}$ mapping γ. Let δ_n be the image of γ_n in $\text{TS}(E/\mathbb{Q})_{2^n}$. Then $2^{n-1}\delta_n = \delta_1$ for all n, and so δ_1 is divisible by all powers of 2. If $\text{TS}(E/\mathbb{Q})$ contains no such element other than zero, then γ is in the image of $E(\mathbb{Q})/2E(\mathbb{Q})$. It is not difficult to show that, in this case, the 2-primary component of $\text{TS}(E/\mathbb{Q})$ is finite (using that $\text{TS}(E/\mathbb{Q})_2$ is finite), and the map $S^{(2^{n+1})}(\mathbb{Q}) \to S^{(2^n)}(\mathbb{Q})$ is onto if $\text{TS}(E/\mathbb{Q})_{2^n} = 0$. □

This gives a strategy for computing r. Calculate $S^{(2)}$, and then leave your computer running overnight searching for points in $E(\mathbb{Q})$. If the subgroup $T(1)$ of $E(\mathbb{Q})$ generated by the points the computer has found maps onto $S^{(2)}$ we have found r, and even a set of generators for $E(\mathbb{Q})$. If not, calculate $S^{(2^2)}$, and have the computer run overnight again finding a bigger group $T(2) \subset E(\mathbb{Q})$. If the image of $T(2)$ in $S^{(2)}$ is $S^{(2,2)}$, then we have found r. If not, we compute $S^{(2^3)} \ldots$.

Nightmare possibility: The Tate-Shafarevich group contains a nonzero element divisible by all powers of 2, in which case the calculation goes on for all eternity. This would happen, for example, if $\text{TS}(E/\mathbb{Q})$ contains a copy of $\mathbb{Q}/\mathbb{Z}$.

Conjecture 16.2. *The Tate-Shafarevich group is always finite.*

When the conjecture is true, then the above argument shows that we have an algorithm for computing $E(\mathbb{Q})$.

Until the work of Rubin and Kolyvagin about 1987, the Tate-Shafarevich group was not known to be finite for a single elliptic curve over a number field, and the conjecture is still far from being proved in that case. For an elliptic curve E over a function field K in one variable over a finite field k, I proved that $\mathrm{TS}(E/K)$ is finite when $j(E) \in k$ in my thesis (1967). Later (1975) I showed that the curve

$$E(j) : Y^2Z = X^3 - \frac{27}{4}\frac{j}{j-1728}XZ^2 - \frac{27}{4}\frac{j}{j-1728}Z^3$$

over the field $K = \mathbb{F}_p(j)$ has finite Tate-Shafarevich group. Not much more is known now.

The Néron-Tate Pairing. We saw in Section 15, that there is a canonical $\mathbb{Z}$-bilinear pairing

$$B : E(\mathbb{Q}) \times E(\mathbb{Q}) \to \mathbb{R}, \quad B(x, y) = \widehat{h}(x + y) - \widehat{h}(x) - \widehat{h}(y).$$

This pairing extends uniquely to an $\mathbb{R}$-bilinear pairing

$$B : E(\mathbb{Q}) \otimes \mathbb{R} \times E(\mathbb{Q}) \otimes \mathbb{R} \to \mathbb{R}.$$

If we choose a $\mathbb{Z}$-basis $e_1, \ldots , e_r$ for $E(\mathbb{Q})/E(\mathbb{Q})_{\mathrm{tors}}$, then $E(\mathbb{Q}) \otimes \mathbb{R}$ has $\mathbb{R}$-basis $(e_1 \otimes 1, \ldots , e_r \otimes 1)$ with respect to which B has matrix

$$(B(e_i, e_j))$$

Theorem 16.3. *The bilinear pairing*

$$E(\mathbb{Q}) \otimes \mathbb{R} \times E(\mathbb{Q}) \otimes \mathbb{R} \to \mathbb{R}$$

defined by $\widehat{h}$ is positive definite (and, in particular, nondegenerate).

This follows from what we have proved already, plus the following result from linear algebra. By a lattice in a real vector space, I mean a $\mathbb{Z}$-submodule generated by a basis for V (sometimes this called a full, or complete, lattice).

Lemma 16.4. *Let $q : V \to \mathbb{R}$ be a quadratic form on a finite-dimensional real vector space V. If there exists a lattice Λ in V such that*

(a) *$q(P) = 0$, $P \in \Lambda$, $\implies P = 0$,*
(b) *for every constant C, the set $\{P \in \Lambda \mid q(P) \leq C\}$ is finite,*

then q is positive definite on V.

Proof. According to Sylvester's theorem (see Math 593), there exists a basis for V relative to which q takes the form

$$q(x) = x_1^2 + \cdots + x_s^2 - x_{s+1}^2 - \cdots - x_t^2, \quad t \leq \dim V.$$

Use the basis to identify V with $\mathbb{R}^n$. Let λ be the length of the shortest vector in Λ, i.e.,

$$\lambda = \inf\{q(P) \mid P \in \Lambda, P \neq 0\}.$$

From (b) we know that $\lambda > 0$. Consider the set

$$B(\delta) = \{(x_i) \mid x_1^2 + \cdots + x_s^2 \leq \frac{\lambda}{2}, \quad x_{s+1}^2 + \cdots + x_t^2 \leq \delta\}.$$

The length (using q) of any vector in $B(\delta)$ is $\leq \lambda/2$, and so $B(\delta) \cap \Lambda = \{0\}$, but the volume of $B(\delta)$ can be made arbitrarily large by taking δ large, and so this violates the following theorem of Minkowski. □

Theorem 16.5 (Minkowski). *Let Λ be a lattice in $\mathbb{R}^n$ with fundamental parallelopiped D_0, and let B be a subset of $\mathbb{R}^n$ that is compact, convex, and symmetric in the origin. If*

$$\mathrm{Vol}(B) \geq 2^n \,\mathrm{Vol}(D)$$

then B contains a point of Λ other than the origin.

Proof. We first show that a measurable set S in $\mathbb{R}^n$ with $\mathrm{Vol}(S) > \mathrm{Vol}(D_0)$ contains distinct points α, β such that $\alpha - \beta \in \Lambda$. Clearly

$$\mathrm{Vol}(S) = \sum \mathrm{Vol}(S \cap D)$$

where the sum is over all the translates of D by elements of Λ. The fundamental parallelopiped D_0 will contain a unique translate (by an element of Λ) of each set $S \cap D$. Since $\mathrm{Vol}(S) > \mathrm{Vol}(D_0)$, at least two of these sets will overlap, and so there exist elements $\alpha, \beta \in S$ such that

$$\alpha - \lambda = \beta - \lambda', \quad \text{some } \lambda \neq \lambda' \in \Lambda.$$

Then $\alpha - \beta = \lambda - \lambda' \in \Lambda \setminus \{0\}$.

We apply this to $\frac{1}{2}B =_{df} \{\frac{x}{2} \mid x \in B\}$. It has volume $\frac{1}{2^n}\mathrm{Vol}(B) > \mathrm{Vol}(D_0)$, and so there exist $\alpha, \beta \in B$, $\alpha \neq \beta$, such that $\alpha/2 - \beta/2 \in \Lambda$. Because B is symmetric about the origin, $-\beta \in B$, and because it is convex, $(\alpha + (-\beta))/2 \in B$. □

Remark 16.6. Systems consisting of a real vector space V, a lattice Λ in V, and a positive-definite quadratic form q on V are of great interest in mathematics. By Sylvester's theorem, we can choose a basis for V that identifies (V, q) with $(\mathbb{R}^n, X_1^2 + \cdots + X_n^2)$. Finding a dense sphere (lattice) packing in $\mathbb{R}^n$ amounts to finding a lattice Λ such that

$$\frac{\|\text{shortest vector}\|^n}{\mathrm{Vol}(\text{fundamental parallelopiped})}$$

is large. Many lattices, for example, the Leech lattice, have very interesting automorphism groups. See Conway and Sloane, Sphere Packings, Lattices and Groups.

From an elliptic curve E over $\mathbb{Q}$, one obtains such a system, namely, $V = E(\mathbb{Q}) \otimes \mathbb{R}$, $\Lambda = E(\mathbb{Q})/E(\mathbb{Q})_{\mathrm{tors}}$, $q = \widehat{h}$. As far as I know, they aren't interesting—at present no elliptic curve is known with $\mathrm{rank}(E(\mathbb{Q})) > 19$. However, for elliptic curves over function fields in one variable over a finite field, Elkies, Shioda, Dummigan, and others have shown that one gets (infinite families of) very interesting lattices.

Computing the rank. Computing the rank r of $E(\mathbb{Q})$ can be difficult (perhaps impossible), but occasionally it is straightforward. In order to avoid the problem of having to work with a number field L other than $\mathbb{Q}$, we assume that the elliptic curve has all its points of order 2 rational over $\mathbb{Q}$:

$$E : Y^2Z = (X - \alpha Z)(X - \beta Z)(X - \gamma Z), \quad \alpha, \beta, \gamma \text{ distinct integers.}$$

The discriminant of $(X - \alpha)(X - \beta)(X - \gamma)$ is

$$\Delta = (\alpha - \beta)^2(\beta - \gamma)^2(\gamma - \alpha)^2.$$

Proposition 16.7. *The rank r of $E(\mathbb{Q})$ satisfies the inequality*

$$r \leq 2 \times \#\{p \mid p \text{ divides } 2\Delta\}.$$

Proof. Since $E(\mathbb{Q}) \approx T \oplus \mathbb{Z}^r$, $T = E(\mathbb{Q})_{\text{tors}}$, we have $E(\mathbb{Q})/2E(\mathbb{Q}) \approx T/2T \oplus (\mathbb{Z}/2\mathbb{Z})^r$. Because T is finite, the kernel and cokernel of $T \xrightarrow{2} T$ have the same order, and so $T/2T \approx (\mathbb{Z}/2\mathbb{Z})^2$. Theorem provides us with an injection $E(\mathbb{Q})/2E(\mathbb{Q}) \hookrightarrow (\mathbb{Q}^\times/\mathbb{Q}^{\times 2})^2$, and the image is contained in the product of the subgroups of $\mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ generated by -1 and the primes where E has bad reduction, namely, those dividing 2Δ. □

It is possible to improve this estimate. Let T_1 be the set of prime numbers dividing Δ for which the reduction is nodal, and let T_2 be the set of prime numbers dividing Δ for which the reduction is cuspidal. Thus T_1 comprises the prime numbers modulo which two of the roots of

$$(X-\alpha)(X-\beta)(X-\gamma)$$

coincide, and T_2 comprises those modulo which all three coincide. Let t_1 and t_2 respectively be the numbers of elements of T_1 and T_2.

Proposition 16.8. *The rank r of $E(\mathbb{Q})$ satisfies $r \leq t_1 + 2t_2 - 1$.*

Proof. Define $\varphi_\alpha : E(\mathbb{Q})/2E(\mathbb{Q}) \to \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ as in (14.7):

$$\varphi_\alpha((x:y:z)) = \begin{cases} (\frac{x}{z}-\alpha)\mathbb{Q}^{\times 2} & z \neq 0, \quad x \neq \alpha z; \\ (\alpha-\beta)(\alpha-\gamma)\mathbb{Q}^{\times} & z \neq 0, \quad x = \alpha z \\ \mathbb{Q}^{\times} & (x:y:z) = (0:1:0). \end{cases}$$

Define φ_β similarly—the map

$$P \mapsto (\varphi_\alpha(P), \varphi_\beta(P)) : E(\mathbb{Q})/2E(\mathbb{Q}) \to (\mathbb{Q}^\times/\mathbb{Q}^{\times 2})^2$$

is injective. For each prime p, let $\varphi_p(P)$ be the element of $(\mathbb{Z}/2\mathbb{Z})^2$ whose components are

$$\operatorname{ord}_p(\varphi_\alpha(P)) \mod 2, \quad \text{and } \operatorname{ord}_p(\varphi_\beta(P)) \mod 2$$

and let $\varphi_\infty(P)$ be the element of $\{\pm\}^2$ whose components are

$$\operatorname{sign}(\varphi_\alpha(P)), \quad \text{and } \operatorname{sign}(\varphi_\beta(P)).$$

The proposition is proved by showing:

(a) if p does not divide Δ, then $\varphi_p(P) = 0$ for all P;
(b) if $p \in T_1$, then $\varphi_p(P)$ is contained in the diagonal of $\mathbb{F}_2^2$ for all P;
(c) when α, β, γ are ordered so that $\alpha < \beta < \gamma$, $\varphi_\infty(P)$ equals $(+,+)$ or $(+,-)$.

Except for $p = 2$, (a) was proved in the paragraph preceding (14.7).

We prove (b) in the case $\alpha \equiv \beta \mod p$ and $P = (x:y:1)$, $x \neq \alpha, \beta, \gamma$. Let

$$a = \operatorname{ord}_p(x-\alpha), \quad b = \operatorname{ord}_p(x-\beta), \quad c = \operatorname{ord}_p(x-\gamma).$$

Because

$$(x-\alpha)(x-\beta)(x-\gamma)$$

is a square, $a + b + c \equiv 0 \mod 2$.

If $a < 0$, then (because $\alpha \in \mathbb{Z}$) p^{-a} occurs as a factor of the denominator of x (in its lowest terms), and it follows that $b = a = c$. Since $a + b + c \equiv 0 \mod 2$, this implies that $a \equiv b \equiv c \equiv 0 \mod 2$, and so $\varphi_p(P) = 0$. The same argument applies if $b < 0$ or $c < 0$.

If $a > 0$, then p divides the numerator of $x - \alpha$. Because p doesn't divide $(\alpha - \gamma)$, it doesn't divide $(\alpha-\gamma)+(x-\alpha) = (x-\gamma)$, and so $c = 0$. Now $a + b \equiv 0 \mod 2$ implies that $\varphi_p(P)$ lies in the diagonal of $\mathbb{F}_2^2$. A similar argument applies if $b > 0$ or $c > 0$.

The remaining cases of (b) are proved similarly.

We prove (c). Let $P = (x : y : 1)$, $x \neq \alpha, \beta, \gamma$. We may suppose that $\alpha < \beta < \gamma$, so that $(x-\alpha) > (x-\beta) > (x-\gamma)$. Then $\varphi_\infty(P) = (+,+)$, $(+,-)$, or $(-,-)$. However, because

$$(x-\alpha)(x-\beta)(x-\gamma)$$

is a square in $\mathbb{Q}$, the pair $(-,-)$ is impossible. The cases $x = \alpha$ etc. are equally easy. □

Example 16.9. The curve

$$E : Y^2Z = X^3 - XZ^2$$

is of the above form with $(\alpha, \beta, \gamma) = (-1, 0, 1)$. The only bad prime is 2, and here the reduction is nodal. Therefore $r = 0$, and E has no point of infinite order:

$$E(\mathbb{Q}) \approx (\mathbb{Z}/2\mathbb{Z})^2.$$

Exercise 16.10. Hand in *one* of the following two problems (those who know the quadratic reciprocity law should do (2)).

(1) Show that $E(\mathbb{Q})$ is finite if E has equation

$$Y^2Z = X^3 - 4XZ^2.$$

Hint: Let P be a point of infinite order in $E(\mathbb{Q})$, and show that, after possibly replacing P with $P + Q$ where $2Q = 0$, $\varphi_2(P)$ is zero. Then show that $\varphi_\infty(P) = (+,+)$—contradiction.

(2) Let E be the elliptic curve

$$Y^2Z = X^3 - p^2XZ^2$$

where p is an odd prime. Show that the rank r of $E(\mathbb{Q})$ satisfies:

$$\begin{aligned} &r \leq 2 && \text{if } p \equiv 1 \mod 8 \\ &r = 0 && \text{if } p \equiv 3 \mod 8 \\ &r \leq 1 && \text{otherwise.} \end{aligned}$$

Hint: Let P be a point of infinite order in $E(\mathbb{Q})$, and show that, after possibly replacing P with $P + Q$ where $2Q = 0$, $\varphi_p(P)$ is zero.

Note: These are fairly standard examples. You should do them without looking them up in a book.

17. Geometric Interpretation of the Cohomology Groups; Jacobians

For simplicity throughout this section we take k to be a perfect field, for example, a field of characteristic zero or a finite field. Everything still holds when k is not perfect except that then the algebraic closure k^{al} of k must be replaced with its separable algebraic closure (the union of all subfields of k^{al} finite and separable over k).

For any elliptic curve E over a field k, we have an exact sequence of cohomology groups:

$$0 \to E(k)/nE(k) \to H^1(k, E_n) \to H^1(k, E)_n \to 0.$$

Here $H^1(k, E_n)$ and $H^1(k, E)$ are defined to be the groups of crossed homomorphisms from $\mathrm{Gal}(k^{\mathrm{al}}/k)$ to $E(k^{\mathrm{al}})_n$ and $E(k^{\mathrm{al}})$ respectively, modulo the principal crossed homomorphisms. In this section, we shall give a geometric interpretation of these groups, and hence also of the Selmer and Tate-Shafarevich groups. We shall attach to any curve W of genus 1 over

k, possibly without a point with coordinates in k, an elliptic curve E, called its Jacobian. The Tate-Shafarevich group of an elliptic curve E classifies the curves of genus 1 over k for which the Hasse principle fails, i.e., such that the curve has a point in $\mathbb{Q}_p$ for all p and in $\mathbb{R}$, but which doesn't have a point in $\mathbb{Q}$.

In general, $H^1(k, ?)$ classifies objects over k that become isomorphic over k^{al} to a fixed object with automorphism group ?. We shall see several examples of this.

Principal homogeneous spaces (of sets). Let A be an abelian group. A right A-set

$$(w, a) \mapsto w + a : W \times A \to W$$

is called a *principal homogeneous space* for A if $W \neq \emptyset$ and the map

$$(w, a) \mapsto (w, w + a) : W \times A \to W \times W$$

is bijective, i.e., if for every pair $w_1, w_2 \in W$, there is a unique $a \in A$ such that $w_1 + a = w_2$.

Example 17.1. (a) Addition $A \times A \to A$ makes A into a principal homogeneous space for A, called the *trivial* principal homogeneous space.

(b) An affine space (for example, the universe according to Newton) is (by definition) a principal homogeneous space for a vector space—essentially, it is a vector space without a preferred origin.

A *morphism* $\varphi : W \to W'$ of principal homogeneous spaces is simply a map A-sets.

Proposition 17.2. *Let W and W' be principal homogeneous spaces for A.*

(a) For any points $w_0 \in W$, $w_0' \in W'$, there exists a unique morphism $\varphi : W \to W'$ sending w_0 to w_0'.

(b) Every morphism $W \to W'$ is an isomorphism (i.e., has an inverse that is also a morphism).

Proof. (a) Uniqueness: Any $w \in W$ can be written uniquely in the form $w = w_0 + a$, $a \in A$, and then $\varphi(w) = w_0' + a$. Existence: This formula defines a morphism.

(b) If φ maps w_0 to w_0', then the unique morphism $W' \to W$ sending w_0' to w_0 is an inverse to φ. □

Corollary 17.3. *(a) Let W be a principal homogeneous space over A. For any point $w_0 \in W$, there is a unique morphism $A \to W$ (of principal homogeneous spaces) sending* 0 *to w_0.*

(b) An $a \in A$ defines an automorphism $w \mapsto w + a$ of W, and every automorphism of W is of this form for a unique $a \in A$.

Hence

$$\text{Aut}(W) = A;$$

—for any abelian group A, we have defined a class of objects having A as their groups of automorphisms.

Principal homogeneous spaces (of curves). Let E be an elliptic curve over a field k. A *principal homogeneous space* for E is a curve W over k together with a right action of E given by a regular [20] map

$$(w,P)\mapsto w+P : W\times E\to W$$

such that

$$(w,P)\mapsto (w,w+P) : W\times E\to W\times W$$

is an isomorphism of algebraic varieties. The conditions imply that, for any field $K\supset k$, $W(K)$ is either empty or is a principal homogeneous space for the group $E(K)$ (in the sense of the previous subsection). A *morphism* of principal homogeneous spaces for E is a regular map $\varphi : W\to W'$ such that

$$\begin{tikzcd}
W\times E \arrow[r] \arrow[d] & W \arrow[d] \\
W'\times E \arrow[r] & W'
\end{tikzcd}$$

commutes. Much of the theory in the previous subsection extends to principal homogeneous spaces for elliptic curves:

Addition $E\times E\to E$ makes E into a principal homogeneous space for E—any principal homogeneous space isomorphic to this principal homogeneous space is said to be *trivial.*

Let W and W' be principal homogeneous spaces for E. For any field $K\supset k$ and any points $w_0\in W(K)$, $w_0'\in W'(K)$, there exists a unique morphism $\varphi : W\to W'$ over K sending w_0 to w_0', and φ is automatically an isomorphism of principal homogeneous spaces over K.

Let W be a principal homogeneous space for E. For any point $w_0\in W(k)$, there is a unique homomorphism $E\to W$ (of principal homogeneous spaces) sending 0 to w_0. Thus W is trivial if and only if $W(k)\neq\emptyset$. Since W will have a point with coordinates in some finite extension K of k (this follows from the Hilbert Nullstellensatz), it becomes trivial over such a K.

A point $P\in E(K)$ defines an automorphism $w\mapsto w+P$ of W, and every automorphism of W over K is of this form for a unique $P\in E(K)$.

The classification of principal homogeneous spaces. Let W be a principal homogeneous space for E, and choose a point $w_0\in W(k^{\mathrm{al}})$. For any $\sigma\in\mathrm{Gal}(k^{\mathrm{al}}/k)$, $\sigma w_0\in W(k^{\mathrm{al}})$, and so can be expressed $\sigma w_0 = w_0+f(\sigma)$ for a unique $f(\sigma)\in E(k^{\mathrm{al}})$. Note that

$$(\sigma\tau)w_0=\sigma(\tau w_0)=\sigma(w_0+f(\tau))=\sigma w_0+\sigma f(\tau)=w_0+f(\sigma)+\sigma f(\tau),$$

and so

$$f(\sigma\tau)=f(\sigma)+\sigma f(\tau).$$

Thus f is a crossed homomorphism $\mathrm{Gal}(k^{\mathrm{al}}/k)\to E(k^{\mathrm{al}})$. Because w_0 has coordinates in a finite extension of k, f is continuous. A second point $w_1\in W(k^{\mathrm{al}})$ will define a second crossed homomorphism f_1, but $w_1=w_0+P$ for some $P\in E(\mathbb{Q}^{\mathrm{al}})$, and so

$$\sigma w_1=\sigma(w_0+P)=\sigma w_0+\sigma P=w_0+f(\sigma)+\sigma P=w_1+f(\sigma)+\sigma P-P.$$

Hence

$$f_1(\sigma)=f(\sigma)+\sigma P-P,$$

[20] That is, one defined by polynomials.

i.e., f and f' differ by a principal crossed homomorphism, and so we have attached a well-defined cohomology class to W.

If the cohomology class is zero, then $f(\sigma) = \sigma P - P$ for some $P \in E(k^{\mathrm{al}})$, and

$$\sigma(w_0 - P) = \sigma w_0 - \sigma P = w_0 + \sigma P - P - \sigma P = w_0 - P.$$

This implies that $w_0 - P \in W(k)$, and so W is a trivial principal homogeneous space.

Theorem 17.4. *The map $W \mapsto [f]$ defines a one-to-one correspondence*

$$\{\text{Principal homogeneous spaces for } E\}/\approx \ \overset{1:1}{\longleftrightarrow} H^1(k, E).$$

Proof. Let $\varphi : W \to W'$ be an isomorphism of principal homogeneous spaces for E (over k), and let $w_0 \in W(k^{\mathrm{al}})$. One checks immediately that (W, w_0) and $(W', \varphi(w_0))$ define the same crossed homomorphism, and hence the map

$$\{\text{Principal homogeneous spaces for } E\}/\approx \ \longrightarrow H^1(k, E)$$

is well-defined. If W and W' define the same cohomology class, we can choose w_0 and w_0' so that (W, w_0) and (W', w_0') define the same crossed homomorphism. There is a unique regular map $\varphi : W \to W'$ over k^{al} sending w_0 to w_0'. Let $w \in W(k^{\mathrm{al}})$, and write $w = w_0 + P$. Then

$$\varphi(\sigma w) = \varphi(\sigma(w_0+P)) = \varphi(\sigma w_0+\sigma P) = \varphi(w_0+f(\sigma)+\sigma P) = w_0'+f(\sigma)+\sigma P = \sigma w_0'+\sigma P = \sigma\varphi(w),$$

which implies that the map φ is defined over k (i.e., it is defined by polynomials with coordinates in k rather than k^{al}). Hence the map is one-to-one. We discuss the surjectivity in the next subsubsection. □

Defining algebraic curves over subfields of algebraically closed fields. Two plane affine curves C_1 and C_2 over k may become isomorphic over k^{al} without being isomorphic over k. The simplest example is the pair of curves

$$X^2 + Y^2 = 1, \quad X^2 + Y^2 = -1,$$

which are not isomorphic over $\mathbb{R}$ (one has no real points) but which become isomorphic over $\mathbb{C}$.

From an affine curve C over k, we obtain an affine curve C' over k^{al} together with an action of $\mathrm{Gal}(k^{\mathrm{al}}/k)$ on $C(k^{\mathrm{al}})$.

Proposition 17.5. *The functor sending a plane affine curve C over k to C' endowed with the action of* $\mathrm{Gal}(k^{al}/k)$ *on $C'(k^{al})$ is fully faithful, i.e., to give a regular map $C_1 \to C_2$ of curves over k is the same as to give a regular map $C_1' \to C_2'$ commuting with the Galois actions.*

We explain the statement. Suppose C_1 and C_2 are defined by the polynomials $F_1(X, Y)$, $F_2(X, Y) \in k[X, Y]$. The curves C_1' and C_2' are defined by the same polynomials now regarded as elements of $k^{\mathrm{al}}[X, Y]$.

By definition, a regular map $\varphi : C_1 \to C_2$ is of the form

$$(x, y) \mapsto (G(x, y), H(x, y)), \quad G(x, y), H(x, y) \in k[C_1] =_{df} k[X, Y]/(F_1(X, Y)).$$

A regular map $\varphi : C_1' \to C_2'$ is of the form

$$(x, y) \mapsto (G(x, y), H(x, y)), \quad G(x, y), H(x, y) \in k^{\mathrm{al}}[C_1'] =_{df} k^{\mathrm{al}}[X, Y]/(F_1(X, Y)).$$

To say that φ commutes with the Galois actions means that, for all $P \in C_1'(k^{\mathrm{al}})$ and all $\sigma \in \mathrm{Gal}(k^{\mathrm{al}}/k)$, $\varphi(\sigma P) = \sigma\varphi(P)$, i.e., $\sigma \circ \varphi \circ \sigma^{-1} = \varphi$. But $\sigma \circ \varphi \circ \sigma^{-1}$ is defined by ${}^{\sigma}G, {}^{\sigma}H$ where ${}^{\sigma}G$ and ${}^{\sigma}H$ are obtained from G and H by applying σ to their coefficients. Therefore, if $\sigma \circ \varphi \circ \sigma^{-1} = \varphi$ for all σ, then $G, H \in k^{\mathrm{al}}[C_1']^{\mathrm{Gal}(k^{\mathrm{al}}/k)} = k[C_1]$.

It follows from the proposition that if a curve C' endowed with an action of the Galois group on $C'(k^{\mathrm{al}})$ arises from a curve C over k, then C is unique (up to a unique isomorphism). We can ask: when does such a pair (C', action) arise from a curve C over k? A necessary and sufficient condition is the following:

(a) the orbits of $\mathrm{Gal}(k^{\mathrm{al}}/k)$ acting on $C'(k^{\mathrm{al}})$ are finite; and
(b) denote the given action of $\sigma \in \mathrm{Gal}(k^{\mathrm{al}}/k)$ on $P \in C'(k^{\mathrm{al}})$ by $\sigma * P$; let ${}^{\sigma}C'$ be the curve obtained from C' by applying σ to the coefficients of the polynomial defining C', and let $P \mapsto \sigma P : C'(k^{\mathrm{al}}) \to {}^{\sigma}C'(k^{\mathrm{al}})$ be the map $(x, y) \to (\sigma x, \sigma y)$; then the map $\sigma * P \mapsto \sigma P : C'(k^{\mathrm{al}}) \to^{\sigma} C'(k^{\mathrm{al}})$ should be regular.

Similar remarks apply to plane projective curves.

Geometric Interpretation of $H^1(\mathbb{Q}, E_n)$. We now give a geometric interpretation of $H^1(k, E_n)$. An *n-covering* is a pair (W, α) consisting of a principal homogeneous space W for E and a regular map $\alpha : W \to E$ (defined over k) with the property: for some $w_1 \in W(k^{\mathrm{al}})$, $\alpha(w_1 + P) = nP$ for all $P \in E(k^{\mathrm{al}})$. A *morphism* $(W, \alpha) \to (W', \alpha')$ (automatically an isomorphism) of n-coverings is a morphism $\varphi : W \to W'$ of principal homogeneous spaces such that $\alpha = \alpha' \circ \varphi$.

For $\sigma \in \mathrm{Gal}(k^{\mathrm{al}}/k)$, write $\sigma w_1 = w_1 + f(\sigma)$, $f(\sigma) \in E(k^{\mathrm{al}})$. As before, $f(\sigma\tau) = f(\sigma) + \sigma f(\tau)$. The equation $\sigma\alpha(w_1) = \alpha(\sigma w_1)$ implies that $nf(\sigma) = 0$, and so f is a crossed homomorphism with values in $E_n(k^{\mathrm{al}})$. The element $w_1 \in W(k^{\mathrm{al}})$ is uniquely determined by the property "$\alpha(w_1 + P) = nP$ for all P" up to replacement by $w_1 + Q$, $Q \in E_n(k^{\mathrm{al}})$. It follows easily that the class of f in $H^1(k, E_n)$ is independent of the choice of w_1.

Theorem 17.6. *The map $(W, \alpha) \mapsto [f]$ defines a bijection*

$$\{n\text{-coverings}\}/\approx \xleftrightarrow{1:1} H^1(k, E_n).$$

Proof. The proof is similar to that of Theorem 17.4. □

Geometric Interpretation of the Exact Sequence. We now give a geometric description of the exact sequence:

$$0 \to E(k)/nE(k) \to H^1(k, E_n) \to H^1(k, E)_n \to 0.$$

If $\gamma \in H^1(k, E_n)$ corresponds to the n-covering (W, α), then the image of γ in $H^1(k, E)$ corresponds to W. If W is trivial, so that there exists a $w_0 \in W(k)$, then γ is the image of the point $\alpha(w_0) \in E(k)$. If w_0' also $\in W(k)$, then $w_0' = w_0 + P$ for some $P \in E(k)$, and $\alpha(w_0') = \alpha(w_0) + nP$, and so $\alpha(w_0)$ is well-defined as an element of $E(k)/nE(k)$.

Twists of Elliptic Curves. In this subsection we study the following problem: given an elliptic curve E_0 over k, find all elliptic curves E over k that become isomorphic to E_0 over k^{al}. Such a curve E is often called a "twist" of E_0. Remember than an elliptic curve E over k has a distinguished point $O \in E(k)$. Throughout, I assume that the characteristic of k is $\neq 2, 3$.

Example 17.7. Consider an elliptic curve

$$E_1 : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in k, \quad \Delta = 4a^3 + 27b^2 \neq 0$$

over k. For any $d \in k^\times$,

$$E_d : dY^2Z = X^3 + aXZ^2 + bZ^3,$$

is an elliptic curve over k that becomes isomorphic to E_1 over k^{al}. Indeed, after making the change of variables $dZ \leftrightarrow Z$, the equation becomes

$$Y^2Z = X^3 + \frac{a}{d^2}XZ^2 + \frac{b}{d^3}Z^3,$$

and so E_d becomes isomorphic to E_1 over any field in which d is a square.

We first compute $\mathrm{Aut}(E, 0)$, the group of automorphisms of E fixing the zero element. According to Theorem 5.3, two elliptic curves

$$E(a, b) : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in k, \quad \Delta(a, b) \neq 0$$

$$E(a', b') : Y^2Z = X^3 + a'XZ^2 + b'Z^3, \quad a', b' \in k, \quad \Delta(a', b') \neq 0$$

are isomorphic if and only if there exists a $c \in k^\times$ such that $a' = c^4a$, $b' = c^6b$, in which case the isomorphisms are of the form

$$(x : y : z) \mapsto (c^2x : c^3y : z).$$

Since these maps not only send O to O', but also map straight lines in $\mathbb{P}^2$ to straight lines, they are homomorphisms. We apply this to the case: $(a', b') = (a, b)$.

Case $ab \neq 0$: Here we seek $c \in k^\times$ such that $c^4 = 1 = c^6$. These equations imply that $c = \pm 1$, and so the only automorphism of (E, O) other than the identity map is

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

Case $a = 0$: Here c can be any 6th root ζ of 1 in k, and the automorphisms of (E, O) are the maps

$$(x : y : z) \mapsto (\zeta^{2i}x : \zeta^{3i}y : z).$$

Case $b = 0$: Here c can be any 4th root ζ of 1 in k, and the automorphisms of (E, O) are the maps

$$(x : y : z) \mapsto (\zeta^{2i}x : \zeta^{3i}y : z).$$

Proposition 17.8. *The automorphism group of (E, O) is $\approx \{\pm 1\}$ unless $j(E) = 0$ or 1728, in which cases it is $\approx \mu_6(k)$ or $\approx \mu_4(k)$ respectively.*

Remark 17.9. (a) Notice that the proposition is consistent with Proposition 10.22, which says that (over $\mathbb{C}$), $\mathrm{End}(E)$ is isomorphic to $\mathbb{Z}$ or to a subring of the ring of integers in a field $\mathbb{Q}[\sqrt{-d}]$. The only units in such rings are roots of 1, and only $\mathbb{Q}[\sqrt{-1}]$ and $\mathbb{Q}[\sqrt{-3}]$ contain roots of 1 other than ±1.

(b) When we allow k to have characteristic 2 or 3, then it is still true that $\mathrm{Aut}(E, O) = \{\pm 1\}$ when $j(E) \neq 0, 1728$, but when $j = 0$ or 1728 the group of automorphisms of (E, O) can have as many as 24 elements.

Fix an elliptic curve E_0 over k, and let E be an elliptic curve over k that becomes isomorphic to E_0 over k^{al}. Choose an isomorphism $\varphi : E_0 \to E$ over k^{al}. For any $\sigma \in \mathrm{Gal}(k^{\mathrm{al}}/k)$, we obtain a second isomorphism $\sigma\varphi =_{df} \sigma \circ \varphi \circ \sigma^{-1} : E_0 \to E$ over k^{al}. For example, if φ is $(x : y : z) \mapsto (c^2x : c^3y : z)$, then $\sigma\varphi$ is $(x : y : z) \mapsto ((\sigma c)^2x : (\sigma c)^3y : z)$. The two isomorphisms $\varphi, \sigma\varphi : E_0 \to E$ (over k^{al}) differ by an automorphism of E_0 over k^{al}:

$$\sigma\varphi = \varphi \circ \alpha(\sigma), \quad \alpha(\sigma) \in \mathrm{Aut}_{k^{\mathrm{al}}}(E_0, O).$$

Note that

$$(\sigma\tau)\varphi = \sigma(\tau\varphi) = \sigma(\varphi \circ \alpha(\tau)) = \varphi \circ \alpha(\sigma) \circ \sigma\alpha(\tau),$$

and so α is a crossed homomorphism $\mathrm{Gal}(k^{\mathrm{al}}/k) \to \mathrm{Aut}_{k^{\mathrm{al}}}(E_0, O)$. Choosing a different isomorphism φ replaces $\alpha(\sigma)$ by its composite with a principal crossed homomorphism.

Theorem 17.10. *The map $E \mapsto [\alpha]$ defines a one-to-one correspondence*

$$\{\text{elliptic curves over } k, \text{ isomorphic to } E_0 \text{ over } k^{al}\}/{\approx} \xleftrightarrow{1:1} H^1(\mathrm{Gal}(k^{al}/k), \mathrm{Aut}_{k^{al}}(E_0)).$$

Proof. The proof is similar to that of Theorem 17.4. □

Corollary 17.11. *If $j(E_0) \neq 0, 1728$, then the list of twists of E_0 in Example* 17.7 *is complete.*

Proof. In this case, $\mathrm{Aut}_{k^{\mathrm{al}}}(E, O) = \mu_2$, and so, according to Example 12.7, $H^1(\mathrm{Gal}(k^{\mathrm{al}}/k), \mu_2) = k^\times/k^{\times 2}$. Under the correspondence in the theorem, $E_d \leftrightarrow d \bmod k^{\times 2}$. □

Remark 17.12. The same arguments can be used to obtain the description of the twisted multiplicative groups on p25. The only endomorphisms of $\mathbb{G}_m = \mathbb{A}^1 \setminus \{0\}$ are the maps $t \mapsto t^m$, some fixed $m \in \mathbb{Z}$. Hence $\mathrm{End}(\mathbb{G}_m) = \mathbb{Z}$ and $\mathrm{Aut}(\mathbb{G}_m) = (\mathrm{End}(\mathbb{G}_m))^\times = \{\pm 1\}$. The twisted forms of $\mathbb{G}_m$ are classified by $H^1(k, \{\pm 1\}) = H^1(k, \mu_2) = k^\times/k^{\times 2}$. The twisted multiplicative group corresponding to $a \in k^\times/k^{\times 2}$ is $\mathbb{G}_m[a]$.

Remark 17.13. Let $\mathrm{Aut}(E)$ be the group of all automorphisms of E, not necessarily preserving O. The map $Q \mapsto t_Q$, where t_Q is the translation $P \mapsto P + Q$, identifies $E(k)$ with a subgroup of $\mathrm{Aut}(E)$. I claim that $\mathrm{Aut}(E)$ is a semi-direct product,

$$\mathrm{Aut}(E) = E(k) \rtimes \mathrm{Aut}(E, O),$$

that is, that

(a) $E(k)$ is a normal subgroup of $\mathrm{Aut}(E)$;
(b) $E(k) \cap \mathrm{Aut}(E, O) = \{0\}$;
(c) $\mathrm{Aut}(E) = E(k) \cdot \mathrm{Aut}(E, O)$.

Let $Q \in E(k)$ and let $\alpha \in \mathrm{Aut}(E, O)$. As we noted above, α is a homomorphism, and so, for any $P \in E$,

$$(\alpha \circ t_Q \circ \alpha^{-1})(P) = \alpha(\alpha^{-1}(P) + Q) = P + \alpha(Q) = t_{\alpha(Q)}(P).$$

Therefore $\alpha \circ t_Q \circ \alpha^{-1} = t_{\alpha(Q)}$, which implies (a). Assertion (b) is obvious. For (c), let $\gamma \in \mathrm{Aut}(E)$, and let $\gamma(0) = Q$; then $\gamma = t_Q \circ (t_{-Q} \circ \gamma)$, and $t_{-Q} \circ \gamma \in \mathrm{Aut}(E, O)$.

Curves of genus 1. Let W be a principal homogeneous space for an elliptic curve E over k. Then W becomes isomorphic to E over k^{al}, and so W is projective, nonsingular, and of genus 1 (at least over k^{al}, which implies that it is also over k). The next theorem shows that, conversely, every projective nonsingular curve W of genus 1 over k occurs as a principal homogeneous space for some elliptic curve over k.

Theorem 17.14. *Let W be a nonsingular projective curve over k of genus* 1. *Then there exists an elliptic curve E_0 over k such that W is a principal homogeneous space for E_0. Moreover, E_0 is unique up to an isomorphism (over k).*

Proof. (Sketch). By assumption, there exists an isomorphism $\varphi : W \to E$ from W to an elliptic curve E over k^{al}, which we may suppose to be in our standard form

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in k, \quad \Delta = 4a^3 + 27b^2 \neq 0.$$

Let $\sigma \in \text{Gal}(k^{\text{al}}/k)$. Then $\sigma\varphi$ is an isomorphism $\sigma W \to \sigma E$. Here σW and σE are obtained from W and E by applying σ to the coefficients of the polynomials defining them (so $E = E(\sigma a, \sigma b)$). But W is defined by polynomials with coefficients in k, and so $\sigma W = W$. Therefore $E \approx W \approx \sigma E$, and $j(E) = j(\sigma E) = \sigma j(E)$. Since this is true for all $\sigma \in \text{Gal}(k^{\text{al}}/k)$, we have that $j(E) \in k$. Now (see bottom of p51) there is a curve E_0 over k with $j(E_0) = j(E)$. In fact, there will be many such curves over k, and so we have to make sure we have the correct one.

Choose an isomorphism $\varphi : E_0 \to W$ over k^{al}, and for $\sigma \in \text{Gal}(k^{\text{al}}/k)$, let $\sigma\varphi = \varphi \circ \alpha(\sigma)$ where $\alpha(\sigma) \in \text{Aut}_{k^{\text{al}}}(E_0)$. Then $\sigma \mapsto \alpha(\sigma)$ is a crossed homomorphism into $\text{Aut}_{k^{\text{al}}}(E_0)$, and hence defines a class $[\alpha]$ in $H^1(k, \text{Aut}_{k^{\text{al}}}(E_0))$. According to (17.13) there is an exact sequence

$$1 \to E_0(k^{\text{al}}) \to \text{Aut}_{k^{\text{al}}}(E_0) \to \text{Aut}_{k^{\text{al}}}(E_0, O) \to 1.$$

If $[\alpha]$ lies in the subgroup $H^1(k, E_0)$ of $H^1(k, \text{Aut}_{k^{\text{al}}}(E_0))$, then W is a principal homogeneous space for E_0. If not, then we use the image of $[\alpha]$ in $H^1(k, \text{Aut}_{k^{\text{al}}}(E_0))$ to twist E_0 to obtain a second curve E_1 over k with the same j-invariant. Now one can check that the class of the crossed homomorphism $[\alpha]$ lies in $H^1(k, E_1)$, and so W is a principal homogeneous space for E_1. □

The curve E_0 given by the theorem is called the *Jacobian* of W. It is characterized by having the following property: there is an isomorphism $\varphi : E_0 \to W$ over k^{al} such that, for all $\sigma \in \text{Gal}(k^{\text{al}}/k)$, there exists a point $Q_\sigma \in E_0(k^{\text{al}})$ such that

$$(\sigma\varphi)(P) = \varphi(P + Q_\sigma), \quad \text{all } P \in E(k^{\text{al}}).$$

Remark 17.15. In the above proof we spoke of a crossed homomorphism into $\text{Aut}_{k^{\text{al}}}(E_0)$, which need not be an abelian group. However, one can still define $H^1(G, M)$ when M is non-abelian as follows. Write M multiplicatively. As in the abelian case, a crossed homomorphism is a map $f : G \to M$ such that $f(\sigma\tau) = f(\sigma)\cdot\sigma f(\tau)$. Call two crossed homomorphisms f and g equivalent if there exists an $m \in M$ such that $g(\sigma) = m^{-1} \cdot f(\sigma) \cdot \sigma m$, and let $H^1(G, M)$ be the set of equivalence classes of crossed homomorphisms. It is a set with a distinguished element, namely, the map $\sigma \mapsto 1$.

Exercise 17.16. Find the Jacobian of the curve

$$W : aX^3 + bX^3 + cY^3 = 0, \quad a, b, c \in \mathbb{Q}^\times.$$

[First, by a change of variables over $\mathbb{Q}^{\mathrm{al}}$, obtain an isomorphism $W \approx E$ where E is an elliptic curve over $\mathbb{Q}^{\mathrm{al}}$ in standard form. Second, write down an elliptic curve E_0 over $\mathbb{Q}$ in standard form that becomes isomorphic to E over $\mathbb{Q}$. Third, modify E_0 if necessary so that it has the property characterizing the Jacobian.]

The classification of elliptic curves over $\mathbb{Q}$ (summary). Let (E, O) be an elliptic curve over $\mathbb{Q}$. We attach to it the invariant $j(E) \in \mathbb{Q}$. Every element of $\mathbb{Q}$ occurs as the j-invariant of an elliptic curve over $\mathbb{Q}$, and two elliptic curves over $\mathbb{Q}$ have the same j-invariant if and only if they become isomorphic over $\mathbb{Q}^{\mathrm{al}}$. See (10.15) et seq..

Fix a $j \in \mathbb{Q}$, and consider the elliptic curves (E, O) over $\mathbb{Q}$ with $j(E) = j$. The isomorphism classes of such curves are in one-to-one correspondence with the elements of $H^1(\mathbb{Q}, \mathrm{Aut}(E, O))$. For example, if $j \neq 0, 1728$, then $\mathrm{Aut}(E, O) = \mu_2$, $H^1(\mathbb{Q}, \mathrm{Aut}(E, O)) = \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$, and the curve corresponding to $d \in \mathbb{Q}^\times$ is the curve E_d of Example 17.7.

Fix an elliptic curve (E, O) over $\mathbb{Q}$, and consider the curves of genus 1 over $\mathbb{Q}$ having E as their Jacobian. Such a curve has the structure of a principal homogeneous space for E, and every principal homogeneous space for E has E as its Jacobian. The principal homogeneous spaces for E are classified by the group $H^1(\mathbb{Q}, E)$, which is a very large group.

Every curve of genus 1 over $\mathbb{Q}$ occurs as the Jacobian of an elliptic curve over $\mathbb{Q}$, and hence as a principal homogeneous space.

Consider the exact sequence of torsion groups

$$0 \to \mathrm{TS}(E/\mathbb{Q}) \to H^1(\mathbb{Q}, E) \to \oplus_{p,\infty} H^1(\mathbb{Q}_p, E) \to C \to 0.$$

Endow each group with the discrete topology. Cassels has shown that the Pontryagin dual of this sequence has the form

$$0 \leftarrow \mathrm{TS}(E/\mathbb{Q}) \leftarrow \Theta \leftarrow \prod_{p,\infty} H^1(\mathbb{Q}_p, E) \leftarrow \widehat{E}(\mathbb{Q}) \leftarrow 0,$$

where $\widehat{E}(\mathbb{Q})$ is the completion of $E(\mathbb{Q})$ for the topology for which the subgroups of finite index form a fundamental system of neighbourhoods of 0, provided that $\mathrm{TS}(E/\mathbb{Q})$ is finite.

Exercise 17.17. Find the Jacobian of the curve

$$W : aX^3 + bY^3 + cZ^3 = 0, \quad a, b, c \in \mathbb{Q}^\times.$$

[Hint: The curve $E : X^3 + Y^3 + dZ^3 = 0$, $d \in \mathbb{Q}^\times$, has the point $O : (1 : -1 : 0)$—the pair (E, O) is an elliptic curve over $\mathbb{Q}$. It can be put in standard form by the change of variables $X = X' + Y'$, $Y = X' - Y'$.]

18. The Tate-Shafarevich Group; Failure Of The Hasse Principle

We discuss a family of curves whose the Tate-Shafarevich groups are nonzero, and which therefore give examples of elliptic curves for which the Hasse principle fails. Full details on what follows can be found in [S1], pp309–318.

Proposition 18.1. *If $p \equiv 1 \mod 8$, then the 2-Selmer group $S^{(2)}(E/\mathbb{Q})$ of the elliptic curve*

$$E : Y^2 Z = X^3 + pXZ^2$$

is isomorphic to $(\mathbb{Z}/2\mathbb{Z})^3$.

The family of curves in the statement is similar to that in Exercise 16.10(2), but since only one of the points of order 2 on E have coordinates in $\mathbb{Q}$, we don't have a simple description of $H^1(\mathbb{Q}, E_2)$. Of course, one can pass to $\mathbb{Q}[\sqrt{p}]$, but it is easier to proceed as follows. One shows that there is a second curve E' and homomorphisms

$$E \xrightarrow{\phi} E' \xrightarrow{\psi} E$$

whose composite is multiplication by 2 and such that the kernel of ϕ is the subgroup of E generated by $P = (0 : 0 : 1)$. From the study of the cohomology sequences of

$$0 \to <P> \to E(\mathbb{Q}^{\mathrm{al}}) \xrightarrow{\phi} E'(\mathbb{Q}^{\mathrm{al}}) \to 0$$

and

$$0 \to \operatorname{Ker} \psi \to E'(\mathbb{Q}^{\mathrm{al}}) \xrightarrow{\psi} E(\mathbb{Q}^{\mathrm{al}}) \to 0$$

one can draw information about $E(\mathbb{Q})/2E(\mathbb{Q})$, $S^{(2)}(E/\mathbb{Q})$, $\mathrm{TS}(E/\mathbb{Q})_2$. For example, Lemma 13.2 applied to the maps

$$E(\mathbb{Q}) \xrightarrow{\phi} E'(\mathbb{Q}) \xrightarrow{\psi} E(\mathbb{Q})$$

shows that there is an exact sequence:

$$E(\mathbb{Q})/\phi(E(\mathbb{Q})) \to E(\mathbb{Q})/2E(\mathbb{Q}) \to E(\mathbb{Q})/\psi(E(\mathbb{Q})) \to 0.$$

Since $E(\mathbb{Q})_2 \approx \mathbb{Z}/2\mathbb{Z}$,

$$\operatorname{rank}(E(\mathbb{Q})) + \dim_{\mathbb{F}_2} \mathrm{TS}(E/\mathbb{Q})_2 = \dim_{\mathbb{F}_2} S^{(2)}(E/\mathbb{Q}) - 1.$$

Thus $r = 0, 1$, or 2, but $r = 1$ is conjecturally ruled out: Cassels has shown that $\mathrm{TS}(E/\mathbb{Q})$ carries a nondegenerate alternating form if it is finite, and the existence of such a form implies that $\dim_{\mathbb{F}_2}(\mathrm{TS}(E/\mathbb{Q}))$ is even. [21]

Proposition 18.2. *Let E be as in (18.1). If 2 is not a fourth power modulo p, then* $\operatorname{rank}(E(\mathbb{Q})) = 0$ *and $TS(E/\mathbb{Q})_2 \approx (\mathbb{Z}/2\mathbb{Z})^2$.*

Remark 18.3. It is, of course, easy (for a computer) to check for any particular prime whether 2 is a fourth power modulo p, but Gauss found a more efficient test.

From Math 593, we know that the ring of Gaussian integers, $\mathbb{Z}[i]$, is a principal ideal domain. An odd prime p either remains prime in $\mathbb{Z}[i]$ or it factors $p = (A + iB)(A - iB)$. In the first case, $\mathbb{Z}[i]/p\mathbb{Z}[i]$ is an field extension of $\mathbb{F}_p$ of degree 2. Therefore p remains prime if and only if $\mathbb{F}_p$ doesn't contain a primitive 4th root of 1. Because $\mathbb{F}_p^\times$ is cyclic, it contains an element of order 4 if and only if 4 divides its order. Therefore the second case occurs if and only if $4|p-1$. We conclude that a prime $p \equiv 1 \mod 4$ can be expressed $p = A^2 + B^2$, $A, B \in \mathbb{Z}$.

Gauss showed that for a prime $p \equiv 1 \mod 8$, 2 is a 4th power modulo p if and only if $8|AB$. Therefore, p satisfies the hypotheses of the proposition if p is

$$17 = 1^2 + 4^2, \quad 41 = 5^2 + 4^2, \quad 97 = 9^2 + 4^2, \quad 193 = 7^2 + 12^2 ...$$

The proof of this, which is quite elementary, can be found in [S1], p318. Number theorists will wish to prove that there are infinitely many such primes p (and find their density).

[21] Recall from Math 593 that a vector space carrying a nondegenerate skew-symmetric form has even dimension, provided the field is of characteristic $\neq 2$. When the form is assumed to be alternating, i.e., $\psi(x, x) = 0$ for all x, then the condition on the characteristic is unnecessary.

It is very difficult to show directly that the rank of an elliptic curve is smaller than the bound given by the Selmer group. Instead, in this case, one exhibits 3 nontrivial elements of $\text{TS}(E/\mathbb{Q})_2$. They are:

$$Y^2 = 4pX^4 - 1, \quad \pm Y^2 = 2pX^4 - 2.$$

One can (no doubt) check directly that these three curves are principal homogeneous spaces for $E : Y^2Z = X^3 + pZ^3$, but it can be more easily seen from the proof of Proposition 18.1 ([S1] 6.2b).

Remark 18.4. We should explain what we mean by these curves. Consider, more generally, the curve

$$C : Y^2 = aX^4 + bX^3 + cX^2 + dX + e$$

where the polynomial on the right has no repeated roots. Assume that the characteristic is $\neq 2, 3$. Then this is a nonsingular affine curve, but its projective closure

$$C' : Y^2Z^2 = aX^4 + bX^3Z + cX^2Z^2 + dXZ^3 + eZ^4$$

is singular: on setting $Y = 1$, we obtain the equation

$$Z^2 = \text{homogeneous polynomial of degree 4},$$

which is visibly singular at $(0,0)$. Recall (p9) that the genus of a plane projective curve of degree d is

$$g = \frac{(d-1)(d-2)}{2} - \sum_{P \text{ singular}} \delta_P.$$

For $P = (0,0)$, $\delta_P = 2$, and so the genus of C' is $3-2 = 1$. When one "blows up" the singular point, one obtains a nonsingular curve and a regular map $C'' \to C'$ that is an isomorphism except over the singular point. It is really C'' that one means when one writes C.

We shall prove that the curve

$$C : Y^2 = 2 - 2pX^4$$

has no points in $\mathbb{Q}$, but has points in $\mathbb{R}$ and $\mathbb{Q}_p$ for all p. For this we shall need to use the quadratic reciprocity law. For an integer a not divisible by the prime p, the *Legendre symbol* $\left(\frac{a}{p}\right) = +1$ or -1 according as a is, or is not, a square modulo p.

Theorem 18.5 (Quadratic reciprocity law). *For odd primes* p, q,

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}.$$

Moreover,

$$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}.$$

Proof. The theorem surely has more published proofs than any other in mathematics. The first proofs were found by Gauss. Most introductory books on number theory contain a proof. □

Proof. We now prove that

$$C : Y^2 = 2 - 2pX^4$$

has no points with coordinates in $\mathbb{Q}$. Suppose (x, y) is a point on the curve. Let $x = r/t$ with r and t integers having no common factor. Then

$$y^2 = \frac{2t^4 - 2pr^4}{t^4}.$$

The numerator and denominator on the right again have no common factor, and so $y = 2s/t^2$ for some integer s with

$$\boxed{2s^2 = t^4 - pr^4.}$$

Let q be an odd prime dividing s. Then $t^4 \equiv pr^4 \mod q$, and so $\left(\frac{p}{q}\right) = 1$. According to the quadratic reciprocity law, this implies that $\left(\frac{q}{p}\right) = 1$. From the quadratic reciprocity law, $\left(\frac{2}{p}\right) = 1$, and so all prime factors of s are squares modulo p. Hence s^2 is a 4th power modulo p. The equation

$$2s^2 \equiv t^4 \mod p$$

now shows that 2 is a 4th power modulo p, which contradicts our hypothesis.

We should also make sure that there is no point lurking at infinity. The projective closure of C is

$$C' : Y^2Z^2 = 2Z^4 - 2pX^4,$$

and we have just shown that C' has no rational point with $Z = 1$. For $Z = 0$, the equation becomes

$$2Z^4 - 2pX^4 = 0$$

which clearly has no rational solution. Since the nonsingular version C'' of C' maps to C', it can't have a rational point either. □

The curve C obviously has points in $\mathbb{R}$. In order to prove that C has a point in $\mathbb{Q}_q$ it suffices (by Hensel's lemma) to show that the reduction $\bar{C}$ of C modulo the prime q has a nonsingular point with coordinates in $\mathbb{F}_q$. For $q \neq 2, p$, the (affine) curve C has good reduction at q, and the results of the next section will show that it has a point with coordinates in $\mathbb{F}_p$ (at least for q not too small). Therefore, C automatically has a point with coordinates in $\mathbb{Q}_q$ except for $q = 2, p$, and perhaps a few additional small primes. The verification for these fields can safely be left to the reader (or the reader's computer).

19. Elliptic Curves Over Finite Fields

As usual, $\mathbb{F}_p$ is the field $\mathbb{Z}/p\mathbb{Z}$ with p elements, $\mathbb{F}$ is a fixed algebraic closure of $\mathbb{F}_p$, and $\mathbb{F}_{p^n}$ is the (unique) subfield of $\mathbb{F}$ with p^n elements. The elements of $\mathbb{F}_{p^n}$ are the roots of $X^{p^n} - X$, and $\mathbb{F}_{p^m} \subset \mathbb{F}_{p^n}$ if and only if $m|n$ (see Math 594).

The Frobenius map; curves of genus 1 over $\mathbb{F}_p$. Let C be a plane projective curve over $\mathbb{F}_p$, so that C is defined by an equation

$$F(X,Y,Z) = \sum_{i+j+k=d} a_{ijk} X^i Y^j Z^k, \quad a_{ijk} \in \mathbb{F}_p.$$

If $P = (x : y : z)$, $x, y, z \in \mathbb{F}$, lies on C, then

$$\sum_{i+j+k=d} a_{ijk} x^i y^j z^k = 0.$$

On raising this equation to the pth power, remembering that we are in characteristic p and that $a^p = a$ for all $a \in \mathbb{F}_p$, we obtain the equation

$$\sum_{i+j+k=d} a_{ijk} x^{ip} y^{jp} z^{kp} = 0,$$

which says that $(x^p : y^p : z^p)$ also lies on C. We therefore obtain a map $(x : y : z) \mapsto (x^p : y^p : z^p) : C \to C$, which, being defined by polynomials, is regular. It is called the *Frobenius map.*

Proposition 19.1. *For any elliptic curve E over $\mathbb{F}_p$, $H^1(\mathbb{F}_p, E) = 0$. Therefore, every principal homogeneous space for E is trivial.*

Proof. Let Γ be the Galois group of $\mathbb{F}$ over $\mathbb{F}_p$. We have to show that every continuous crossed homomorphism $f : \Gamma \to E(\mathbb{F})$ is principal.

We first determine the structure of Γ. The map $a \mapsto a^p$ is an automorphism of $\mathbb{F}$, which we call the *Frobenius automorphism*, and denote σ. As we noted above, for each $n \geq 1$, $\mathbb{F}_p$ has a unique extension of degree n contained in $\mathbb{F}$, namely, $\mathbb{F}_{p^n}$. Moreover, $\mathbb{F}_{p^n}$, being the splitting field of $X^{p^n} - X$, is Galois over $\mathbb{F}_p$, and $\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$ is generated by $\sigma|\mathbb{F}_{p^n}$. Therefore, by infinite Galois theory, for each $n \geq 1$, Γ has a unique open subgroup Γ_n of index n, and Γ/Γ_n is generated by $\sigma\Gamma_n$. It follows that σ has infinite order, and that Γ is the closure of the subgroup generated by σ—we say that σ generates Γ as a topological group. Note that for $P = (x : y : z) \in E(\mathbb{F})$, and $\varphi : E \to E$ the Frobenius map,

$$\varphi(P) = (x^p : y^p : z^p) = (\sigma x : \sigma y : \sigma z) = \sigma P.$$

Now consider a crossed homomorphism $f : \Gamma \to E(\mathbb{F})$. The map $P \mapsto \varphi(P) - P$ is a nonconstant regular map $E \to E$; it therefore induces a surjective[22] map $E(\mathbb{F}) \to E(\mathbb{F})$. In particular, there exists a $P \in E(\mathbb{F})$ such that $\varphi(P) - P = f(\sigma)$, i.e., such that $f(\sigma) = \sigma P - P$. Then

$$f(\sigma^2) = f(\sigma) + \sigma f(\sigma) = \sigma P - P + \sigma^2 P - \sigma P = \sigma^2 P - P,$$
$$\cdots$$
$$f(\sigma^n) = f(\sigma) + \sigma f(\sigma^{n-1}) = \sigma P - P + \sigma(\sigma^{n-1} P - P) = \sigma^n P - P.$$

Therefore f and the principal crossed homomorphism $\tau \mapsto \tau P - P$ agree on σ^n for all n. Because both crossed homomorphisms are continuous, this implies that they agree on the whole of Γ. □

[22] Any nonconstant regular map $\varphi : C \to C'$ from a connected projective curve to an irreducible curve is surjective as a map of algebraic curves (this implies that $C(k^{\mathrm{al}}) \to C'(k^{\mathrm{al}})$ is surjective, but not necessarily that $C(k) \to C'(k)$ is surjective): because C is projective the image of φ is Zariski-closed; because C' is connected, its only proper Zariski-closed subsets are finite; therefore, $\varphi(C) \neq C' \implies \varphi(C)$ is finite $\implies$ $\varphi(C) =$ a single point (because C is connected) $\implies$ φ is constant.

Corollary 19.2. *A nonsingular projective curve C of genus 1 over $\mathbb{F}_p$ has a point with coordinates in $\mathbb{F}_p$.*

Proof. According to (17.14), the curve C is a principal homogeneous space for its Jacobian E, and according to the Proposition, it is a trivial principal homogeneous space, i.e., $C(\mathbb{F}_p) \neq 0$. □

I next want to prove the Riemann hypothesis for an elliptic curve, namely, that if N is the number of points on the elliptic curve E with coordinates in $\mathbb{F}_p$, then $|N-p-1| \leq 2\sqrt{p}$. However, first I'll explain why this statement is called the Riemann hypothesis, which involves reviewing some of the formalism of zeta functions.

Zeta functions of number fields. First recall that the original (Riemann's) Riemann zeta function is

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1-p^{-s}} = \sum_{n\geq 1} n^{-s}, \quad s \text{ complex}, \quad \Re(s) > 1.$$

The second equality is an expression of unique factorization:

$$\zeta(s) = \prod_{p} \frac{1}{1-p^{-s}} = \prod_{p} \left(1 + p^{-s} + (p^{-s})^2 + (p^{-s})^3 + \cdots\right);$$

on multiplying out this product, we obtain a sum of terms

$$(p_1^{-s})^{r_1}(p_2^{-s})^{r_2}\cdots(p_t^{-s})^{r_t} = (p_1^{r_1}\cdots p_t^{r_t})^{-s}.$$

Both the sum and the product converge for $\Re(s) > 1$, and so $\zeta(s)$ is holomorphic and nonzero for $\Re(s) > 1$. In fact, $\zeta(s)$ extends to a meromorphic function on the whole complex plane with a simple pole at $s = 0$. Moreover, the function $\xi(s) = \pi^{-\frac{s}{2}}\Gamma(\frac{s}{2})\zeta(s)$ satisfies the functional equation $\xi(s) = \xi(1-s)$, has simple poles at $s = 0, 1$, and is otherwise holomorphic. Since $\Gamma(s)$ has poles at $s = 0, -1, -2, -3, \ldots$, this forces ζ to be zero at $s = -2n$, $n > 0$, $n \in \mathbb{Z}$. These are called the *trivial zeros* of the zeta function.

Conjecture 19.3 (Riemann hypothesis). *The nontrivial zeros of $\zeta(s)$ lie on the line $\Re(s) = \frac{1}{2}$.*

This is perhaps the most famous problem remaining in mathematics.

Dedekind extended Riemann's definition by attaching a zeta function $\zeta_K(s)$ to any number field K. He defined

$$\zeta_K(s) = \prod_{\mathfrak{p}} \frac{1}{1-\mathbb{N}(\mathfrak{p})^{-s}} = \sum_{\mathfrak{a}} \mathbb{N}(\mathfrak{a})^{-s}, \quad s \text{ complex}, \quad \Re(s) > 1.$$

The first product is over the nonzero prime ideals in K, and the second is over all ideals in $\mathcal{O}_K$. The *numerical norm* $\mathbb{N}\mathfrak{a}$ of an ideal $\mathfrak{a}$ is the order of the quotient ring $\mathcal{O}_K/\mathfrak{a}$. The proof of the second equality uses that every nonzero ideal $\mathfrak{a}$ has a unique factorization $\mathfrak{a} = \prod \mathfrak{p}_i^{r_i}$ into a product of powers of prime ideals, and that $\mathbb{N}(\mathfrak{a}) = \prod \mathbb{N}(\mathfrak{p}_i)^{r_i}$. Note that for $K = \mathbb{Q}$, this definition gives back $\zeta(s)$. The function $\zeta_K(s)$ extends to a meromorphic function on the whole complex plane with a simple pole at $s = 1$. Moreover, a certain multiple $\xi_K(s)$ of it satisfies a functional equation $\xi_K(s) = \xi_K(1-s)$. It is conjectured that $\zeta_K(s)$ has its nontrivial zeros on the line $\Re(s) = \frac{1}{2}$.

Zeta functions of affine curves over finite fields. Consider a plane affine curve

$$C : f(X,Y) = 0$$

over the field $\mathbb{F}_p$. In analogy with the above definition, we set

$$\zeta(C,s) = \prod_{\mathfrak{p}} \frac{1}{1-\mathbb{N}\mathfrak{p}^{-s}}, \quad \Re(s) > 1,$$

where $\mathfrak{p}$ runs over the nonzero prime ideals in $\mathbb{F}_p[C] =_{df} \mathbb{F}_p[X,Y]/(f(X,Y)) = \mathbb{F}_p[x,y]$. For any nonzero prime ideal $\mathfrak{p}$ of $\mathbb{F}_p[C]$, the quotient $\mathbb{F}_p[C]/\mathfrak{p}$ is finite, and so we can again define $\mathbb{N}\mathfrak{p}$ to be its order.

Because $\mathbb{F}_p[C]/\mathfrak{p}$ is finite and an integral domain, it is a field (and $\mathfrak{p}$ is maximal). We define $\deg \mathfrak{p}$ to be the degree of $\mathbb{F}_p[C]/\mathfrak{p}$ over $\mathbb{F}_p$, so that $\mathbb{N}\mathfrak{p} = p^{\deg \mathfrak{p}}$. This allows us to make a change of variables in the zeta function: when we define

$$Z(C,T) = \prod_{\mathfrak{p}} \frac{1}{1-T^{\deg \mathfrak{p}}},$$

then

$$\zeta(C,s) = Z(C,p^{-s}).$$

Here Z is a capital zeta. The product $\prod \frac{1}{1-T^{\deg \mathfrak{p}}}$ converges for all small T, but generally we simply regard it as a formal power series

$$Z(C,T) = \prod_{\mathfrak{p}} (1 + T^{\deg \mathfrak{p}} + T^{2\deg \mathfrak{p}} + T^{3\deg \mathfrak{p}} + \cdots) \in \mathbb{Z}[[T]].$$

In the following, I will often make use of the fact that many of the identities in Math 115 involving power series are valid for power series over any ring. For example, we can define $\log(1+t)$ to be the power series

$$\log(1+t) = t - \frac{t^2}{2} + \frac{t^3}{3} - \frac{t^4}{4} + \frac{t^5}{5} - \cdots,$$

and then

$$\log \frac{1}{1-t} = -\log(1-t) = t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \cdots.$$

Computation of the zeta function of $\mathbb{A}^1$*.* To fit it into the above scheme, we can regard $\mathbb{A}^1$ as the curve $Y = 0$, and so

$$\mathbb{F}_p[\mathbb{A}^1] = \mathbb{F}_p[X,Y]/(Y) = \mathbb{F}_p[X].$$

The nonzero (prime) ideals on $\mathbb{F}_p[X]$ are in one-to-one correspondence with the monic (irreducible) polynomials $f(X) \in \mathbb{F}_p[X]$, and so

$$Z(\mathbb{A}^1,T) = \prod_{f \text{ monic irreducible}} \frac{1}{1-T^{\deg f}}.$$

Take the log of both sides,

$$\log Z(\mathbb{A}^1,T) = -\sum_f \log(1-T^{\deg f}),$$

and then the derivatives

$$\begin{aligned}\frac{Z'(T)}{Z(T)} &= \sum_f \frac{\deg f\cdot T^{\deg f-1}}{1-T^{\deg f}}\\ &= \sum_{n\geq 0,f} \deg f\cdot T^{(n+1)\deg f-1}.\end{aligned}$$

In this power series, the coefficient of T^{m-1} is $\sum \deg f$, where f runs over all monic irreducible polynomials $f\in\mathbb{F}_p[T]$ of degree dividing m. Note that $\deg f|m$ if and only if all of the roots of f lie in $\mathbb{F}_{p^m}$, and that, conversely, every element of $\mathbb{F}_{p^m}$ is the root of a polynomial of degree dividing m. Therefore, the coefficient of T^{m-1} is p^m, and we have

$$\frac{Z'(T)}{Z(T)} = \sum p^m T^{m-1}.$$

On integrating, we find that

$$\log Z(\mathbb{A}^1,T) = \sum \frac{p^m T^m}{m} = \log\frac{1}{1-pT}.$$

We have shown:

Proposition 19.4. *The zeta function of* $\mathbb{A}^1$ *has the property that*

$$\log Z(\mathbb{A}^1,T) = \sum \frac{\#\mathbb{A}^1(\mathbb{F}_{p^m})T^m}{m}$$

and so

$$Z(\mathbb{A}^1,T) = \frac{1}{1-pT}.$$

Expression of $Z(C,T)$ in terms of the points of C. Let C be an affine curve over $\mathbb{F}_p$. As for $\mathbb{A}^1$,

$$\begin{aligned}\frac{Z'(C,T)}{Z(C,T)} &= \sum_{\mathfrak{p}} \frac{\deg\mathfrak{p}\cdot T^{\deg\mathfrak{p}-1}}{1-T^{\deg\mathfrak{p}}}\\ &= \sum_{n\geq 0,\mathfrak{p}} \deg\mathfrak{p}\cdot T^{(n+1)\deg\mathfrak{p}}/T.\end{aligned}$$

In this power series, the coefficient of T^{m-1} is $\sum\deg\mathfrak{p}$ where $\mathfrak{p}$ runs over the prime ideals of $k[C]$ such that $\deg\mathfrak{p}$ divides m. But $\deg\mathfrak{p} =_{df} [\mathbb{F}_p[C]/\mathfrak{p}:\mathbb{F}_p]$, and so the condition $\deg\mathfrak{p}|m$ means that there is a homomorphism $\mathbb{F}_p[C]/\mathfrak{p}\hookrightarrow\mathbb{F}_{p^m}$—there will in fact be exactly $\deg\mathfrak{p}$ such homomorphisms (because $\mathbb{F}_p[C]/\mathfrak{p}$ is separable over $\mathbb{F}_p$). Conversely, any homomorphism $\mathbb{F}_p[C]\to\mathbb{F}_{p^m}$ factors through $\mathbb{F}_p[C]/\mathfrak{p}$ for some prime ideal with $\deg\mathfrak{p}|m$. Therefore, the coefficient of T^{m-1} in the above power series is the number of homomorphisms (of $\mathbb{F}_p$-algebras)

$$\mathbb{F}_p[C]\to\mathbb{F}_{p^m}.$$

But $\mathbb{F}_p[C]=\mathbb{F}_p[X,Y]/(f(X,Y))$, and so a homomorphism $\mathbb{F}_p[C]\to\mathbb{F}_{p^m}$ is determined by the images a,b of X,Y, and conversely the homomorphism $P(X,Y)\mapsto P(a,b):\mathbb{F}_p[X,Y]\to\mathbb{F}_{p^m}$ determined by a pair $a,b\in\mathbb{F}_{p^m}$ factors through $\mathbb{F}_p[X,Y]/(f(X,Y))$ if and only if $f(a,b)=0$. Therefore there is a natural one-to-one correspondence

$$\{\text{homomorphisms } \mathbb{F}_p[C]\to\mathbb{F}_{p^m}\}\overset{1:1}{\leftrightarrow} C(\mathbb{F}_{p^m}).$$

We have proved:

Proposition 19.5. *The zeta function of an affine curve C has the property that*

$$\log Z(C,T) = \sum N_m \frac{T^m}{m}, \quad N_m = \#C(\mathbb{F}_{p^m}).$$

In other words,

$$Z(C,T) = \exp(\sum_{m\geq 1} N_m \frac{T^m}{m}),$$

where

$$\exp T = 1 + T + \frac{T^2}{2!} + \cdots + \frac{T^n}{n!} + \cdots .$$

Zeta functions of plane projective curves. For a plane projective curve C (in fact, an arbitrary curve) over $\mathbb{F}_p$ one usually defines $Z(C,T)$ to be the power series such that

$$\log Z(C,T) = \sum N_m(C) \frac{T^m}{m}, \quad N_m(C) = \#C(\mathbb{F}_{p^m}),$$

and then one defines

$$\zeta(C,s) = Z(C,p^{-s}).$$

If $C = C_0 \cup C_1$, then

$$N_m(C) = N_m(C_0) + N_m(C_1) - N_m(C_0 \cap C_1),$$

and so

$$\zeta(C,s) = \frac{\zeta(C_0,s)\zeta(C_1,s)}{\zeta(C_0 \cap C_1,s)}.$$

Thus, $\zeta(C,s)$ can also be expressed as a product of terms $\frac{1}{1-\mathbb{N}\mathfrak{p}^{-s}}$.

For example,

$$N_m(\mathbb{P}^1) = N_m(\mathbb{A}^1) + 1, \text{ for all } m,$$

and so

$$\log Z(\mathbb{P}^1,T) = \log Z(\mathbb{A}^1,T) + \log \frac{1}{1-T}.$$

Therefore

$$Z(\mathbb{P}^1,T) = \frac{1}{1-T} Z(\mathbb{A}^1,T) = \frac{1}{(1-T)(1-pT)}.$$

Similarly, for E an elliptic curve over $\mathbb{F}_p$ and E^{aff} the affine curve $E \cap \{Z \neq 0\}$,

$$Z(E,T) = \frac{1}{1-T} Z(E^{\mathrm{aff}},T).$$

The rationality of the zeta function of an elliptic curve.

Theorem 19.6. *Let E be an elliptic curve over $\mathbb{F}_p$. Then*

$$Z(E,T)=\frac{1+(N_1-p-1)T+pT^2}{(1-T)(1-pT)},\quad N_1=N_1(E).$$

Remark 19.7. (a) Factor

$$1+(N_1-p-1)T+pT^2=(1-\alpha T)(1-\beta T),$$

so that α,β are algebraic integers such that

$$N_1-p-1=-\alpha-\beta,\quad \alpha\beta=p.$$

Then

$$\log Z(E,T)=\log\frac{(1-\alpha T)(1-\beta T)}{(1-T)(1-pT)}=\sum(1+p^m-\alpha^m-\beta^m)\frac{T^m}{m},$$

and so

$$N_m(E)=1+p^m-\alpha^m-\beta^m.$$

Thus, if one knows N_1, one can find α and β, and the whole of the sequence

$$N_1(E),N_2(E),N_3(E),\ldots.$$

(b) With the notation in (a),

$$\zeta(E,s)=\frac{(1-\alpha p^{-s})(1-\beta p^{-s})}{(1-p^{-s})(1-p^{1-s})}.$$

It has simple poles at $s=0$ and $s=1$, and zeros where $p^s=\alpha$ and $p^s=\beta$. Write $s=\sigma+it$. Then $|p^s|=p^\sigma$, and the zeros of $\zeta(E,s)$ have real part $\frac{1}{2}$ if and only if α and β have absolute value $p^{\frac{1}{2}}$.

By definition α and β are the roots of a polynomial

$$1+bT+pT^2,\quad b=N_1-p-1.$$

If $b^2-4p\leq 0$, then α and β are complex conjugates. Since their product is p, this implies that they each have absolute value $p^{\frac{1}{2}}$. Conversely, if $|\alpha|=p^{\frac{1}{2}}=|\beta|$, then

$$|N_1-p-1|=|\alpha+\beta|\leq 2\sqrt{p}.$$

Thus, granted Theorem 19.6, the Riemann hypothesis for E is equivalent to the statement

$$|N_1-p-1|\leq 2\sqrt{p}.$$

Exercise 19.8. (a) Prove that the zeta function of an elliptic curve E over $\mathbb{F}_p$ satisfies the functional equation

$$\zeta(E,s)=\zeta(E,1-s).$$

(b) Compute the zeta functions for the curve

$$E:Y^2Z+YZ^2=X^3-X^2Z$$

over the fields $\mathbb{F}_2$, $\mathbb{F}_3$, $\mathbb{F}_5$, $\mathbb{F}_7$, and verify the Riemann hypothesis in each case. How many points does the curve have over the field with 625 elements?

I outline some of the ideas that go into the proof of the theorem. First consider $Z(\mathbb{A}^1, T)$ again. By definition

$$\begin{aligned} Z(\mathbb{A}^1, T) &= \prod_f \frac{1}{1 - T^{\deg f}} \\ &= \prod_f (1 + T^{\deg f} + T^{2\deg f} + T^{3\deg f} + \cdots), \end{aligned}$$

where the product is over the monic irreducible polynomials in $\mathbb{F}_p[X]$. On multiplying out we obtain a formal power series that is a sum of terms of the form

$$T^{r_1 \deg f_1} \ldots T^{r_t \deg f_t}.$$

The coefficient of T^m is the number sequences $(r_1, f_1), \ldots, (r_t, f_t)$ such that $\sum r_i \deg f_i = m$. Because unique factorization holds in $\mathbb{F}_p[X]$, we can identify such a sequence with a monic polynomial $\prod f_i^{r_i}$ of degree m. Therefore, the coefficient of T^m is the number of monic polynomials in $\mathbb{F}_p[X]$ of degree m. This is p^m, and so

$$Z(\mathbb{A}^1, T) = \sum p^m T^m = \frac{1}{1 - pT}.$$

Hence (again),

$$Z(\mathbb{P}^1, T) = \frac{1}{1 - T} Z(\mathbb{A}^1, T) = \frac{1}{(1 - T)(1 - pT)}.$$

Now consider an elliptic curve E over $\mathbb{F}_p$. Here

$$\begin{aligned} Z(E, T) &= \frac{1}{1 - T} Z(E^{\text{aff}}, T) \\ &= \frac{1}{1 - T} \prod_{\mathfrak{p}} \frac{1}{1 - T^{\deg \mathfrak{p}}}, \end{aligned}$$

where the $\mathfrak{p}$ run through the prime ideals of

$$\mathbb{F}_p[E^{\text{aff}}] = \mathbb{F}_p[X, Y]/(Y^2 - X^3 - aX - b).$$

On multiplying the product out and applying the Riemann-Roch theorem (see below), we find that

$$Z(E, T) = \sum d_m T^m$$

where $d_0 = 1$ and $d_m = N_1 \frac{p^m - 1}{p - 1}$. Therefore,

$$\begin{aligned} Z(T) &= 1 + \sum_{m > 0} N_1 \frac{p^m - 1}{p - 1} T^m \\ &= \frac{N_1}{p - 1} \left(\frac{1}{1 - pT} - \frac{1}{1 - T} \right) + 1 \\ &= \frac{N_1}{p - 1} \left(\frac{(p - 1)T}{(1 - pT)(1 - T)} \right) + 1 \\ &= \frac{1 + (N_1 - p - 1)T + pT^2}{(1 - T)(1 - pT)}. \end{aligned}$$

Application of the Riemann-Roch theorem. In this subsection, we explain how the Riemann-Roch theorem leads to the above formula for d_m. Consider an elliptic curve

$$E : Y^2Z = X^3 + aXZ^2 + bZ^3, \quad a, b \in k, \quad \Delta \neq 0,$$

over a field k. Write E^{aff} for the affine curve

$$Y^2 = X^3 + aX + b.$$

A *divisor* on E is a finite sum

$$D = \sum r_i \mathfrak{p}_i$$

with $r_i \in \mathbb{Z}$ and $\mathfrak{p}_i$ either a nonzero prime ideal in $k[E^{\text{aff}}]$ or another symbol $\mathfrak{p}_\infty$ (the "prime divisor corresponding to the point at infinity"). A divisor $D = \sum r_i \mathfrak{p}_i$ is *positive* if $r_i \geq 0$ for all i. The *degree* of $\mathfrak{p}$ is the degree of the field extension $[k[C^{\text{aff}}]/\mathfrak{p} : k]$ if $\mathfrak{p} \neq \mathfrak{p}_\infty$, and is 1 if $\mathfrak{p} = \mathfrak{p}_\infty$. We extend the definition to all divisors by linearity:

$$\deg D = \sum r_i \deg \mathfrak{p}_i.$$

The ring $k[E^{\text{aff}}]$ is a Dedekind domain, and so, for each prime ideal $\mathfrak{p}$, the localization $k[E^{\text{aff}}]_{\mathfrak{p}}$ is a principal ideal domain with a single prime element (up to associates). Therefore, corresponding to such a prime ideal, we obtain a valuation $\text{ord}_{\mathfrak{p}} : k(E)^\times \to \mathbb{Z}$—if $f \in k[E^{\text{aff}}]$, $(f) = \prod \mathfrak{p}^{\text{ord}_{\mathfrak{p}}(f)}$. Similarly, there is a valuation $\text{ord}_{\mathfrak{p}_\infty} : k(E)^\times \to \mathbb{Z}$ measuring the order of the zero or pole of f at infinity. The *divisor* of an $f \in k(E)^\times$ is

$$(f) = \sum \text{ord}_{\mathfrak{p}}(f)\mathfrak{p}.$$

For a divisor D, define

$$L(D) = \{f \in k(E)^\times \mid (f) + D \geq 0\} \cup \{0\}.$$

Theorem 19.9 (Riemann-Roch). *For any divisor D, $L(D)$ is a finite-dimensional vector space over k, and if* $\deg D \geq 1$, *then*

$$\dim L(D) = \deg D.$$

We need to restate this slightly. Fix D_0, and consider the set $P(D_0)$ of all divisors D such that

(a) $D \geq 0$,
(b) $D \sim D_0$, i.e., $D = D_0 + (f)$ for some $f \in k(E)^\times$.

Then

$$f \mapsto D_0 + (f) : L(D_0) \setminus \{0\} \to P(D_0)$$

is surjective, and two functions have the same image if and only if one is a constant multiple of the other. We therefore have a bijection

$$(L(D_0) \setminus \{0\})/k^\times \to P(D_0).$$

In the case that $k = \mathbb{F}_q$ and $\deg D_0 = m \geq 1$, the Riemann-Roch theorem implies that $P(D_0)$ has

$$\frac{q^m - 1}{q - 1}$$

elements.

Remark 19.10. The map sending (a,b) in $E^{\text{aff}}(k)$ to the ideal $\mathfrak{p}_{a,b} =_{df} (x-a, y-b) \subset k[E^{\text{aff}}]$ (ideal of all regular functions on E^{aff} zero at (a,b)) is a one-to-one correspondence

$$E^{\text{aff}}(k) \overset{1:1}{\leftrightarrow} \{\text{nonzero prime ideals in } k[E^{\text{aff}}] \text{ of degree } 1\}.$$

When k is algebraically closed, all prime ideals in $k[E^{\text{aff}}]$ have degree 1, and the definitions and statements above agree with those in Section 4.

We need two more facts. The degree of a principal divisor (f) is zero, and so the degree map factors through the quotient group $\text{Pic}(E) = \text{Div}(E)/\{(f) \mid f \in k(E)^{\times}\}$. Moreover

$$\text{Pic}E \xrightarrow{\deg} \mathbb{Z}$$

is surjective, because $\mathfrak{p}_\infty \mapsto 1$. Define

$$\text{Pic}^m E = \{\mathfrak{d} \in \text{Pic}(E) \mid \deg \mathfrak{d} = m\}.$$

Then

(a) the map $E(k) \to \text{Pic}^0(E)$,

$$\begin{cases} (a:b:1) & \mapsto & \mathfrak{p}_{a,b} \\ (0:1:0) & \mapsto & \mathfrak{p}_\infty \end{cases}$$

is an isomorphism of abelian groups (cf. 4.7; this again follows from the Riemann-Roch theorem);

(b) the map $\text{Pic}^0(E) \to \text{Pic}^m(E)$, $\mathfrak{d} \mapsto \mathfrak{d} + m\mathfrak{p}_\infty$, is a bijection (this is obvious from the definition of $\text{Pic}^m(E)$).

We are now able to derive the formula for d_m, $m \geq 1$. With the above terminology, the coefficient d_m of T^m is the set of positive divisors on E of degree m. In the discussion following the Riemann-Roch theorem, we saw that each class in $\text{Pic}^m(E)$ has $\frac{p^m-1}{p-1}$ elements, and there are

$$\#\text{Pic}^m(E) \overset{(b)}{=} \#\text{Pic}^0(E) \overset{(a)}{=} \#E(\mathbb{F}_p) = N_1$$

such classes. Therefore, altogether, there are

$$N_1 \frac{p^m - 1}{p - 1}$$

positive divisors of degree m.

Proof of the Riemann hypothesis for elliptic curves. Let

$$b = N_1 - p - 1.$$

We have to show that

$$b^2 - 4p \leq 0.$$

I'll only sketch the proof. The details of what follows can be found in [C2], Sections 24, 25, and most other books on elliptic curves.

Fix an algebraically closed field k. Consider a nonconstant regular map

$$\varphi : C \to C'$$

from one irreducible affine curve C to a second (defined over k). Then φ being regular means that it defines a homomorphism

$$f \mapsto f \circ \varphi : k[C'] \to k[C]$$

from the k-algebra of regular functions on C' to the k-algebra of regular functions on C. This map is injective, and so defines map on the fields of fractions

$$k(C') \to k(C),$$

which realizes $k(C)$ as a finite extension of $k(C')$. The degree of φ is defined to be the degree of this extension.

Important fact: If $k(C)$ is separable over $k(C')$, then $\varphi^{-1}(P)$ has $\deg\varphi$ points for all but finitely many $P \in C(k)$, i.e., if $\deg\varphi = d$, then φ is "generically" $d : 1$.

Example 19.11. Consider the map $(x, y) \mapsto x : E^{\text{aff}}(k) \to \mathbb{A}^1(k)$, where E^{aff} is the curve

$$E^{\text{aff}} : Y^2 = X^3 + aX + b.$$

The map on the rings of regular functions is

$$X \mapsto x : k[X] \to k[x, y] =_{df} k[X, Y]/(Y^2 - X^3 - aX - b).$$

Clearly $k(x, y) = k(x)[\sqrt{x^3 + ax + b}]$, and so the map has degree 2. If characteristic $k \neq 2$, then the field extension is separable, and the map is $2 : 1$ except over the roots of $X^3 + aX + b$.

A nonconstant map $\varphi : E \to E'$ of elliptic curves such that $\varphi(0) = 0$ is called an *isogeny*. An isogeny is automatically a homomorphism, and so its fibres $\varphi^{-1}(P)$ all have the same cardinality, which will be $\deg\varphi$ if $k(E)$ is separable over $k(E')$.

We need two facts about degrees of isogenies:

(a) $\deg(\varphi \circ \psi) = \deg(\varphi) \cdot \deg(\psi)$;
(b) $\deg(\varphi + \psi) + \deg(\varphi - \psi) = 2\deg(\varphi) + 2\deg(\psi)$.

The first statement is simply

$$[k(E) : k(E'')] = [k(E) : k(E')][k(E') : k(E'')]$$

(see Math 594), and the second has a proof that is similar to the proof of the parallelogram law for the height function.

The second statement implies that $\deg : \text{End}(E) \to \mathbb{Z}$ is quadratic (see 15.6), i.e., that

$$\deg(m\varphi + n\psi) = am^2 + bmn + cn^2, \quad \text{all } m, n \in \mathbb{Z},$$

for certain integers a, b, c depending on φ and ψ. In fact, on taking $(m, n) = (1, 0)$ or $(0, 1)$, we see that $a = \deg\varphi$ and $c = \deg\psi$. Now $\deg(m\varphi + n\psi) \geq 0$ for all m, n. Hence

$$ar^2 + br + c \geq 0$$

for all $r \in \mathbb{Q}$. From high school algebra, this implies that

$$b^2 - 4ac \leq 0.$$

We apply this with φ the Frobenius map $E \to E$ and ψ the identity map id_E. From the picture

$$\begin{tikzcd}
k(x,y) \arrow[r, "(x,y)\mapsto(x^p,y^p)"] \arrow[d, dash, "2"] & k(x,y) \arrow[d, dash, "2"] \\
k(X) \arrow[r, "X\mapsto X^p"] & k(X)
\end{tikzcd}$$

we see that

$$\deg\varphi = [k(x, y) : k(x^p, y^p)] = [k(X) : k(X^p)] = p,$$

and clearly $\deg \mathrm{id}_E = 1$. Therefore

$$b^2 - 4p \leq 0$$

where b is such that

$$\deg(\varphi - \mathrm{id}_E) = \deg \varphi - b + \deg \mathrm{id}_E = p - b + 1.$$

But the kernel of $\varphi - \mathrm{id}$ is the set of $P \in E(k)$ such that $\varphi(P) = P$, i.e., such that $(x^p : y^p : z^p) = (x : y : z)$, i.e., such that $P \in E(\mathbb{F}_p)$. Therefore (assuming separability),

$$\deg(\varphi - \mathrm{id}) = N_1,$$

and so

$$-b = N_1 - p - 1.$$

The above inequality becomes

$$(N_1 - p - 1)^2 \leq 4p,$$

as required.

A Brief History of Zeta. The story begins, as do most stories in number theory, with Gauss.

Gauss 1801. Consider the elliptic curve

$$E : X^3 + Y^3 + Z^3 = 0$$

over $\mathbb{F}_p$, $p \neq 3$.

If $p \not\equiv 1 \mod 3$, then 3 doesn't divide the order of $\mathbb{F}_p^\times$, and so $a \mapsto a^3$ is a bijection $\mathbb{F}_p^\times \to \mathbb{F}_p^\times$. It follows that $(x : y : z) \mapsto (x^3 : y^3 : z^3)$ is a bijection from $E(\mathbb{F}_p)$ onto $L(\mathbb{F}_p)$ where L is the line

$$L : X + Y + Z = 0.$$

Therefore

$$\#E(\mathbb{F}_p) = \#L(\mathbb{F}_p) = p + 1.$$

If $p \equiv 1 \mod 3$, then $4p = A^2 + 27B^2$ where A and B are integers, uniquely determined up to sign. Fix the sign of A by requiring that $A \equiv 1 \mod 3$. Then Gauss showed that

$$\#E(\mathbb{F}_p) = p + 1 + A.$$

Therefore,

$$|\#E(\mathbb{F}_p) - p - 1| = |A| = |4p - 27B^2|^{\frac{1}{2}} \leq 2p^{\frac{1}{2}},$$

and so the Riemann hypothesis holds for E over $\mathbb{F}_p$, all $p \neq 3$. For details, see Silverman and Tate, pp110–119.

This result is typical for elliptic curves with complex multiplication.

Emil Artin 1924. In his thesis, Artin defined the zeta function of a curve of the form

$$C : Y^2 = f(X).$$

Here $\mathbb{F}_p(C) = \mathbb{F}_p(X)[\sqrt{f}]$, and so $\mathbb{F}_p(C)$ is analogous to a quadratic extension of $\mathbb{Q}$. He proved that it is a rational function of p^{-s} and satisfies a functional equation, and he checked the Riemann hypothesis, at least for a few curves.

F.K. Schmidt 1925–1930. For a nonsingular projective curve C over $\mathbb{F}_p$, define

$$\zeta(C,s) = Z(C,p^{-s}), \quad Z(C,T) = \sum_{m\geq 1} N_m \frac{T^m}{m}, \quad N_m = \#C(\mathbb{F}_{p^m}).$$

Using the Riemann-Roch theorem, Schmidt showed (as in the above proof for elliptic curves) that

$$Z(C,T) = \frac{P(T)}{(1-T)(1-pT)}, \quad P(T) = 1 + \cdots \in \mathbb{Z}[T], \quad \deg P(T) = 2g.$$

Moreover,

$$Z(\frac{1}{pT}) = p^{1-g}T^{2-2g}Z(T),$$

and so

$$\zeta(1-s) = p^{1-g}(p^{2-2g})^{-s}\zeta(s).$$

Write

$$P(T) = \prod (1-\alpha_j T).$$

Then the formulas show that

$$N_m = 1 + p^m - \sum \alpha_j^m.$$

Thus, once one knows $\alpha_1, \ldots, \alpha_{2g}$, then one knows N_m for all m. However, unlike the elliptic curve case, N_1 doesn't determine the α_j's—one needs to know several of the N_m's.

Hasse 1933/34. Hasse proved that the Riemann hypothesis is true for elliptic curves.

Weil 1940–1948. In 1940, Weil announced a proof of the Riemann hypothesis for all curves, i.e., that $|\alpha_i| = \sqrt{p}$ for $1 \leq i \leq 2g$ where α_i is above. His proof assumed the existence of a theory of algebraic geometry over arbitrary fields, including the theory of Jacobian and Abelian varieties, which at the time was known only over $\mathbb{C}$. He spent most of the 1940's developing the algebraic geometry he needed, and gave a detailed proof in a book published in 1948.

Weil 1949/1954. Weil studied zeta functions of some special algebraic varieties, and stated his famous "Weil conjectures".

For a nonsingular projective variety V of any dimension over $\mathbb{F}_p$, one can define (as for curves)

$$\zeta(V,s) = Z(V,p^{-s}), \quad Z(V,T) = \sum_{m\geq 1} N_m \frac{T^m}{m}, \quad N_m = \#V(\mathbb{F}_{p^m}).$$

Weil conjectured that $Z(V,T)$ is a rational function of T, that a "Riemann hypothesis" holds for $\zeta(V,s)$, and that $\zeta(V,s)$ satisfies a functional equation—see below for exact statements. He even hinted that one might be able to prove the rationality by developing a cohomology theory for algebraic varieties over arbitrary fields, analogous to that provided by algebraic topology over $\mathbb{C}$, and for which a Lefschetz fixed point formula. At the time, this seemed an outlandish idea.

Dwork 1960. Dwork gave an "elementary" proof that $Z(V,T)$ is a rational function of T.

Grothendieck et al 1963/64. Grothendieck defined étale cohomology and, with the help of M. Artin and Verdier, developed it sufficiently to prove that $Z(V,T)$ is rational and satisfies a functional equation. The rationality follows from a "Lefschetz fixed point formula", and the functional equation from "Poincaré duality" theorem.

Deligne 1973. Deligne used étale cohomology to prove the remaining Weil conjecture, namely, the Riemann hypothesis. For this, he received the Fields medal.

In summary, the results of Grothendieck, Artin, Verdier, and Deligne show that, for a nonsingular projective variety V over $\mathbb{F}_p$,

$$Z(V,T)=\frac{P_1(T)P_3(T)\cdots P_{2d-1}(T)}{(1-T)P_2(T)P_4(T)\cdots P_{2d-2}(T)(1-p^dT)}$$

where $d=\dim V$ and $P_i(T)\in\mathbb{Z}[T]$. Moreover

$$Z(V,\frac{1}{p^dT})=\pm T^{\chi}p^{d\chi/2}Z(V,T)$$

where χ is the self-intersection number of the diagonal in $V\times V$. Finally (Riemann hypothesis):

$$P_i(T)=\prod(1-\alpha_{ij}T),\quad |\alpha_{ij}|=p^{i/2}.$$

This last statement says that $\zeta(V,s)$ has its zeros on the lines

$$\Re(s)=\frac{1}{2},\frac{3}{2},\ldots,\frac{2d-1}{2}$$

and its poles on the lines

$$\Re(s)=0,1,2,\ldots,d.$$

Exercise 19.12. (a) Let E be the elliptic curve

$$E:Y^2Z=X^3-4X^2Z+16Z^3.$$

Compute $N_p=_{df}\#E(\mathbb{F}_p)$ for all primes $3\leq p\leq 13$ (more if you use a computer).

(b) Let $F(q)$ be the (formal) power series given by the infinite product

$$F(q)=q\prod_{n=1}^{\infty}(1-q^n)^2(1-q^{11n})^2=q-2q^2-q^3+2q^4+\cdots.$$

Calculate the coefficient M_n of q^n in $F(q)$ for $n\leq 13$ (more if you use a computer).

(c) For each prime p, compute the sum M_p+N_p. Formulate a conjecture as to what M_p+N_p should be in general.

(d) Prove your conjecture. [This is probably very difficult, perhaps even impossible, using only the information covered in the course.]

[[Your conjecture is, or, at least, should be a special case of a theorem of Eichler and Shimura. Wiles's big result is that the theorem of Eichler and Shimura applies to virtually all elliptic curves over $\mathbb{Q}$.]]

20. The Conjecture of Birch and Swinnerton-Dyer

Introduction. We return to the problem of computing the rank of $E(\mathbb{Q})$. Our purely algebraic approach provides only an upper bound for the rank, via the Selmer group, and the difference between the upper bound and the actual rank is measured by the mysterious Tate-Shafarevich group. It is very difficult to decide whether an element of $S^{(2)}(E/\mathbb{Q})$ comes from an element of infinite order, or survives to give a nontrivial element of $\mathrm{TS}(E/\mathbb{Q})$—in fact, there is no (proven) algorithm for doing this. Clearly, it would be helpful to have another approach.

One idea is that perhaps the rank $E(\mathbb{Q})$ should be related to the orders of the groups $\bar{E}(\mathbb{F}_p)$. For any good prime p, there is a reduction map

$$E(\mathbb{Q}) \to \bar{E}(\mathbb{F}_p)$$

but, in general, this will be far from injective or surjective. For example, if $E(\mathbb{Q})$ is infinite, then so also is the kernel, and if $E(\mathbb{Q})$ is finite (and hence has order ≤ 18) then it will fail to be surjective for all large p (because $\#\bar{E}(\mathbb{F}_p) \geq p + 1 - 2\sqrt{p}$).

Let E be an elliptic curve over $\mathbb{Q}$. For each prime p where E has good reduction, I write (in contradiction to the notations of the last section)

$$N_p = \#\bar{E}(\mathbb{F}_p)$$

where $\bar{E}$ denotes the reduction of E over $\mathbb{F}_p$.

In the late fifties, Birch and Swinnerton-Dyer had the idea that if $E(\mathbb{Q})$ is large then this should force the N_p's to be "large". Since they had access to one of the few computers then in existence, they were able to test this experimentally. For P a large number (large, depending on the speed of your computer), let

$$f(P) = \prod_{p \leq P} \frac{N_p}{p}.$$

Recall that N_p is approximately p. Their calculations led them to the following conjecture.

Conjecture 20.1. *For each elliptic curve E over $\mathbb{Q}$, there exists a constant C such that*

$$\lim_{P\to\infty} f(P) = C \log(P)^r$$

where $r = \mathrm{rank}(E(\mathbb{Q}))$.

We write the conjecture more succinctly as

$$f(P) \sim C \log(P)^r \text{ as } P \to \infty.$$

Note the remarkable nature of this conjecture: it predicts that one can determine the rank of $E(\mathbb{Q})$ from the sequence of numbers N_p. Moreover, together with an estimate for the error term, it will provide an algorithm for finding r.

Birch and Swinnerton-Dyer were, in practice, able to predict r from this conjecture with fairly consistent success, but they found that $f(P)$ oscillates vigorously as P increases, and that there seemed to be little hope of finding C with an error of less than say 10%. Instead, they re-expressed their conjecture in terms of the zeta function of E.

The zeta function of a variety over $\mathbb{Q}$. Let V be a nonsingular projective variety over $\mathbb{Q}$. Such a variety is the zero set of a collection of homogeneous polynomials

$$F(X_0, \dots, X_n) \in \mathbb{Q}[X_0, \dots, X_n].$$

Scale each such polynomial so that its coefficients lie in $\mathbb{Z}$ but have no common factor, and let $\bar{F}(X_0, \dots, X_n) \in \mathbb{F}_p[X_0, \dots, X_n]$ be the reduction of the scaled polynomial modulo p. If the polynomials $\bar{F}$ define a nonsingular variety V_p over $\mathbb{F}_p$, then we say V has *good reduction* at p, or that p is *good* for V. All but finitely many primes will be good for a given variety.

For each good prime p we have a zeta function

$$\zeta(V_p, s) = Z(V_p, p^{-s}), \quad \log Z(V_p, T) = \sum \#V_p(\mathbb{F}_{p^m}) \frac{T^m}{m},$$

and we define

$$\zeta(V, s) = \prod_p \zeta(V_p, s).$$

Because the Riemann hypothesis holds for V_p, the product converges for $\Re(s) > d + 1$, $d = \dim V$ (cf. the explanation below for the L-series of an elliptic curve).

Let Pt be the point over $\mathbb{Q}$, i.e., $Pt = \mathbb{A}^0 = \mathbb{P}^0$. For this variety, all primes are good, and

$$\log Z(Pt_p, T) = \sum 1 \frac{T^m}{m} = \log \frac{1}{1-T}.$$

Therefore

$$\zeta(Pt, s) = \prod_p \frac{1}{1 - p^{-s}}$$

which is just the Riemann zeta function—already an interesting function.

Let $V = \mathbb{P}^n$. Again all primes are good, and

$$\#\mathbb{P}^n(\mathbb{F}_q) = \frac{q^{n+1} - 1}{q - 1} = 1 + q + \cdots + q^n, \quad q = p^m,$$

from which it follows that

$$\zeta(\mathbb{P}^n, s) = \zeta(s)\zeta(s-1)\cdots\zeta(s-n)$$

with $\zeta(s)$ the Riemann zeta function.

Conjecture 20.2 (Hasse-Weil). *For any nonsingular projective variety V over $\mathbb{Q}$, $\zeta(V, s)$ can be analytically continued to a meromorphic function on the whole complex plane, and satisfies a functional equation relating $\zeta(V, s)$ with $\zeta(V, d+1-s)$, $d = \dim V$ (and, of course, it is expected to satisfy a Riemann hypothesis, but let's not concern ourselves with that).*

The conjecture is widely believed to be true, but it is known in only a few cases (and the parenthetical statement is not known for any variety, not even a point). Note that the above calculations prove it for $V = \mathbb{P}^n$.

The zeta function of an elliptic curve over $\mathbb{Q}$. Let E be an elliptic curve over $\mathbb{Q}$, and let S be the set of primes where E has bad reduction. According to the above definition

$$\begin{aligned}\zeta(E,s) &= \prod_{p\notin S}\frac{1+(N_p-p-1)p^{-s}+p^{1-2s}}{(1-p^{-s})(1-p^{1-s})}\\ &= \frac{\zeta_S(s)\zeta_S(s-1)}{L_S(s)}\end{aligned}$$

where $\zeta_S(s)$ is Riemann's zeta function except that the factors corresponding to the primes in S have been omitted, and

$$L_S(E,s)=\prod_{p\notin S}\frac{1}{1+(N_p-p-1)p^{-s}+p^{1-2s}}.$$

Write

$$1+(N_p-p-1)T+pT^2=(1-\alpha_pT)(1-\beta_pT),$$

so that

$$L_S(E,s)=\prod_p\frac{1}{1-\alpha_pp^{-s}}\frac{1}{1-\beta_pp^{-s}}.$$

As we noted on p 92, the product $\prod_p\frac{1}{1-p^{-s}}$ converges for $\Re(s)>1$, and so

$$\prod_p\frac{1}{1-p^{\frac{1}{2}}p^{-s}}$$

converges for $\Re(s)>\frac{3}{2}$. Because $|\alpha_p|=p^{\frac{1}{2}}=|\beta_p|$, a similar estimate shows that $L_S(E,s)$ converges for $\Re(s)>\frac{3}{2}$.

We want to add factors to $L_S(E,s)$ for the bad primes. Define

$$L_p(T)=\begin{cases}1-a_pT+pT^2, & a_p=p+1-N_p,\quad p\text{ good}\\ 1-T, & \text{if } E \text{ has split multiplicative reduction}\\ 1+T, & \text{if } E \text{ has non-split multiplicative reduction}\\ 1, & \text{if } E \text{ has additive reduction.}\end{cases}$$

In the four cases

$$L_p(p^{-1})=\frac{N_p}{p},\frac{p-1}{p},\frac{p+1}{p},\frac{p}{p}.$$

Thus in each case, $L_p(p^{-1})=\frac{\#E^{\text{ns}}(\mathbb{F}_p)}{p}$, where E^{ns} is the nonsingular part of the reduction of the elliptic curve modulo p (see the table on p29). Define

$$L(E,s)=\prod_p\frac{1}{L_p(p^{-s})},$$

where the product is now over all prime numbers. The *conductor* $N_{E/\mathbb{Q}}$ of E is defined to be $\prod_{p\text{ bad}}p^{f_p}$ where

$$f_p=\begin{cases}f_p=1 & \text{if } E \text{ has multiplicative reduction at } p\\ f_p\geq 2 & \text{if } E \text{ has additive reduction at } p, \text{ and equals 2 if } p\neq 2,3.\end{cases}$$

A formula of Ogg (proved by painful, case by case, checking) states that

$$f_p=\operatorname{ord}_p(\Delta)+1-m_p$$

where m_p is the number of irreducible components on the Néron model (not counting multiplicities) and Δ is the discriminant of the minimum equation of E

$$Y^2 + a_1XY + a_3Y = X^3 + a_2X^2 + a_4X + a_{6.}$$

See Section 22 below.

Define $\Lambda(E,s) = N_{E/\mathbb{Q}}^{s/2}(2\pi)^{-s}\Gamma(s)L(E,s)$. The following is a more precise version of the Hasse-Weil conjecture for the case of an elliptic curve.

Conjecture 20.3. *The function $\Lambda(E,s)$ can be analytically continued to a meromorphic function on the whole of $\mathbb{C}$, and it satisfies a functional equation*

$$\Lambda(E,s) = w\Lambda(E,2-s), \quad w = \pm 1.$$

There is even a recipe for what w should be.

For curves with complex multiplication, i.e., such that $\mathrm{End}(E) \neq \mathbb{Z}$, the conjecture was proved by Deuring 1951/55. This case occurs for exactly 9 values of the j-invariant. The key point is that there is always a "formula" for N_p similar to that proved by Gauss for the curve $X^3 + Y^3 + Z^3 = 0$ (see p101) which allows one to identify the L-series of E with an L-series previously defined by Hecke and for which one knows analytic continuation and a functional equation.

A much more important result, which we'll spend most of the rest of the term discussing, is the following. Let

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \middle| c \equiv 0 \mod N \right\}.$$

Then $\Gamma_0(N)$ acts on the complex upper half-plane, and the quotient $\Gamma_0(N)\backslash\mathbb{H}$ has the structure of a Riemann surface. An elliptic curve E over $\mathbb{Q}$ is said to be *modular* if there is a nonconstant map of Riemann surfaces

$$\Gamma_0(N)\backslash\mathbb{H} \to E(\mathbb{C})$$

for some N. Eichler and Shimura (in the fifties and sixties) proved a slightly weaker form of Conjecture 20.3 for modular elliptic curves.

Recall that a curve is said to have semistable reduction at p if it has good reduction at p or multiplicative reduction at p, i.e., if the reduced curve doesn't have a cusp.

Wiles (with the help of Richard Taylor) proved that an elliptic curve with semistable reduction at all p is modular, and Diamond extended Wiles's results to elliptic curves having semistable reduction at 3 and 5.

Remark 20.4. (a) Deuring's result is valid for elliptic curves with complex multiplication over any number field. The results of Eichler, Shimura, Wiles, et al. are valid only for elliptic curves over $\mathbb{Q}$. Even today, little is known about the L-series of elliptic curves over number fields other than $\mathbb{Q}$.

(b) Both results prove much more than the simple statement of Conjecture 20.3—they succeed in identifying the function $\Lambda(E,s)$.

Statement of the Conjecture of Birch and Swinnerton-Dyer. Let E be an elliptic curve over $\mathbb{Q}$. Let

$$Y^2 + a_1XY + a_3Y = X^3 + a_2X^2 + a_4X + a_6$$

be a minimum equation for E over $\mathbb{Q}$ (see Section 22 below), and let

$$\omega = \frac{dx}{2y + a_1x + a_3}.$$

Recall that there is a canonical pairing

$$<,>: E(\mathbb{Q}) \times E(\mathbb{Q}) \to \mathbb{R}, \quad < P, Q >= \widehat{h}(P+Q) - \widehat{h}(P) - \widehat{h}(Q).$$

We define the discriminant of $<,>$ to be the determinant of the $r \times r$ matrix whose i, jth entry is $< P_i, P_j >$ where $P_1, \ldots, P_r$ is a basis for $E(\mathbb{Q})$ modulo torsion:

$$\text{disc} <,>= \det(< P_i, P_j >).$$

It is independent of the choice of the basis $\{P_i\}$.

Conjecture 20.5 (Birch and Swinnerton-Dyer). *For any elliptic curve E over $\mathbb{Q}$,*

$$L(E,s) \sim \left(\Omega \prod_{p \text{ bad}} c_p\right) \frac{[TS(E/\mathbb{Q})] \,\text{disc} <,>}{[E(\mathbb{Q})_{\text{tors}}]^2}(s-1)^r \text{ as } s \to 1,$$

where

$[*] =$ order of $*$ (elsewhere written $\#*$);

$\Omega = \int_{E(\mathbb{R})} |\omega|$;

$c_p = (E(\mathbb{Q}_p) : E^0(\mathbb{Q}_p))$.

Remark 20.6. (a) As we discuss in Section 22, for a modular elliptic curve, all terms in the conjecture are computable except for the Tate-Shafarevich group, and, in fact, can be computed by Pari.

(b) Formally, $L_p(1) = \prod \frac{p}{N_p}$, and so the conjecture has an air of compatibility with Conjecture 20.1. I don't know what (if any) is the precise mathematical relation between the two conjectures.

(c) Let $P_1, \ldots, P_r$ be linearly independent elements of $E(\mathbb{Q})$. Then

$$\frac{\det(< P_i, P_j >)}{(E(\mathbb{Q}) : \sum \mathbb{Z}P_i)^2}$$

is independent of the choice of $P_1, \ldots, P_r$, and equals

$$\frac{\text{disc} <,>}{[E(\mathbb{Q})_{\text{tors}}]^2}$$

when they form a basis.

(d) The integral $\int_{E(\mathbb{Q}_p)} |\omega|$ makes sense, and, in fact equals $(E(\mathbb{Q}_p) : E^1(\mathbb{Q}_p))/p$. The explanation for the formula is that (see 7.3) there is a bijection $E^1(\mathbb{Q}_p) \leftrightarrow p\mathbb{Z}_p$ under which ω corresponds to the Haar measure on $\mathbb{Z}_p$ for which $\mathbb{Z}_p$ has measure 1 and (therefore) $p\mathbb{Z}_p$ has measure $1/(\mathbb{Z}_p : p\mathbb{Z}_p) = 1/p$. Hence,

$$\int_{E(\mathbb{Q}_p)} |\omega| = (E(\mathbb{Q}_p) : E^1(\mathbb{Q}_p)) \int_{E^1(\mathbb{Q}_p)} = |\omega| = \frac{(E(\mathbb{Q}_p) : E^1(\mathbb{Q}_p))}{p} = \frac{c_pN_p}{p}.$$

For any finite set S of prime numbers including all those for which E has bad reduction, define

$$L_S^*(s) = \left(\prod_{p\in S\cup\{\infty\}} \int_{E(\mathbb{Q}_p)} |\omega|\right)^{-1} \prod_{p\notin S} \frac{1}{L_p(p^{-s})}.$$

In this, $\mathbb{Q}_p = \mathbb{R}$ when $p = \infty$. When p is good,

$$L_p(p^{-1}) = \frac{N_p}{p} = \left(\int_{E(\mathbb{Q}_p)} |\omega|\right),$$

and so the behaviour of $L_S(s)$ near s is independent[23] of S satisfying the condition, and the conjecture of Birch and Swinnerton-Dyer can be stated as:

$$L_S^*(E,s) \sim \frac{[\mathrm{TS}(E/\mathbb{Q})]\,\mathrm{disc} <,>}{[E(\mathbb{Q})_{\mathrm{tors}}]^2}(s-1)^r \text{ as } s \to 1.$$

This is how Birch and Swinnerton-Dyer stated their conjecture.

What's known about the conjecture of B-S/D. Birch and Swinnerton-Dyer, Stephens, and many others, have computed all the terms in the conjecture except TS for several thousand curves. The predicted value of [TS] turns out to be a square, and, when computed, the 2 and 3 primary components have the correct order.

Cassels proved that [TS] is a square if finite. Thus, if $[\mathrm{TS}_p]$ has order not equal to a square for some p, then TS is infinite.

Let E and E' be two elliptic curves over $\mathbb{Q}$, and suppose there is an isogeny $E \to E'$. Most of the terms in Conjecture 20.5 differ for the two curves, but nevertheless Cassels was able to show that if the conjecture if true for one curve, then it is true for the other, i.e., that the conjecture is compatible with isogenies.

These results of Cassels were interesting applications of Galois cohomology.

For certain elliptic curves over function fields, the conjecture was proved in 1967 (see the next section).

Thus, by the mid-seventies, the little progress had been made toward proving the conjecture over $\mathbb{Q}$. The Tate-Shafarevich group was not known to be finite for a single curve. In 1974, Tate said:

> This remarkable conjecture relates the behaviour of a function L at a point where it is not at present known to be defined to the order of a group TS which is not known to be finite.

Coates and Wiles (1977): If E has complex multiplication, and $E(\mathbb{Q})$ is infinite, then $L(E,1) = 0$.

Birch: For a modular elliptic curve $E/\mathbb{Q}$ and a complex quadratic extension K of $\mathbb{Q}$, he defined a "Heegner point" $P_K \in E(K)$, and suggested that it should often be of infinite order.

Gross-Zagier (1983): Proved the formula

$$\widehat{h}(P_K) = (\text{nonzero})L'(E/K,1).$$

Thus P_K has infinite order if and only if $L'(E/K,1) \neq 0$.

[23] More precisely, $\lim_{s\to 1} L_S(s)/L_{S'}(s) = 1$ for any two such sets S, S'.

Let $K = \mathbb{Q}[\sqrt{D}]$, $D < 0$, be a complex quadratic extension of $\mathbb{Q}$. If E is the curve

$$E : Y^2 Z = X^3 + aXZ^2 + bZ^3,$$

define E^K to be the curve

$$E^K : DY^2 Z = X^3 + aXZ^2 + bZ^3$$

—thus E^K becomes isomorphic to E over K. There is an elementary formula:

$$L(E/K, s) = L(E/\mathbb{Q}, s) \cdot L(E^K, s).$$

Bump-Friedberg-Hoffstein (1989): Showed that, given a modular elliptic curve E over $\mathbb{Q}$, there exists a complex quadratic field K such that $L'(E^K, 1) \neq 0$ (and hence the formula of Gross and Zagier proves that P_K has infinite order if $L(E/\mathbb{Q}, 1) \neq 0$).

Kolyvagin (1988): For a modular elliptic curve $E/\mathbb{Q}$, if P_K has infinite order for some complex quadratic extension K of $\mathbb{Q}$, then $E(\mathbb{Q})$ and $\mathrm{TS}(E/\mathbb{Q})$ are both finite.

On combining these results, we find that $L(E/\mathbb{Q}, 1) \neq 0 \implies E(\mathbb{Q})$ and $\mathrm{TS}(E/\mathbb{Q})$ are finite.

In fact, Kolyvagin proves much more. For example, he shows that $[\mathrm{TS}(E/\mathbb{Q})]$ divides its conjectured order. To complete the proof of the conjecture of Birch and Swinnerton-Dyer, it suffices to check the its p-primary component has the correct order for a finite set of primes.

Roughly speaking, this is what was known by 1990.

21. Elliptic Curves and Sphere Packings

The conjecture of Birch and Swinnerton-Dyer is expected to hold, not just for elliptic curves over $\mathbb{Q}$, but also for elliptic curves over number fields and over certain function fields. In the second case, the full conjecture has been proved in some important cases, and Elkies and Shioda have shown that it can be used to recover (at least in dimensions ≤ 1000) most of the known lattices that give very dense sphere packings, and in certain dimensions, for example, 33, 54, 64, 80,..., to discover new denser sphere packings.

Let q be a power of the prime p, and let $\mathbb{F}_q(T)$ be the field of fractions of $\mathbb{F}_q[T]$. The height of a point P of $\mathbb{P}^1(\mathbb{F}_q(T))$ can be defined as for a point of $\mathbb{P}^1(\mathbb{Q})$: represent the point as $(f(T) : g(T))$ where f and g have been chosen to lie in $\mathbb{F}_q[T]$ and be relatively prime, and define

$$H(P) = \max(\mathbb{F}_q[T] : (f)), (\mathbb{F}_q[T] : (g)) = \max q^{\deg f}, q^{\deg g}.$$

The logarithmic height is

$$h(P) = \log q \cdot \max\{\deg f, \deg g\}$$

If $p \neq 2, 3$, an elliptic curve E over $\mathbb{F}_q(T)$ can be written

$$Y^2 Z = X^3 + a(T)X^2 Z + b(T)Z^3, \quad a, b \in \mathbb{F}_q[T], \quad \Delta(T) = 4a^3 + 27b^2 \neq 0.$$

For each monic irreducible polynomial $p(T)$ in $\mathbb{F}_q[T]$, we have a homomorphism $a \mapsto \bar{a} : \mathbb{F}_q[T] \to \mathbb{F}_q[T]/(p(T))$, and so we obtain a curve

$$\bar{E} : Y^2 Z = X^3 + \bar{a}X^2 Z + \bar{b}Z^3, \quad \bar{a}, \bar{b} \in \mathbb{F}_q[T]/(p(T)),$$

over the field $\mathbb{F}_q[T]/(p(T))$. All the terms that go into the conjecture of Birch and Swinnerton-Dyer in the number field case can be defined here.

More generally, let K be a finite extension of $\mathbb{F}_q(T)$. There will exist a nonsingular projective curve C such that $\mathbb{F}_q(T) = \mathbb{F}_q(C)$. As we discussed on p102,

$$Z(C,T) = \frac{\prod_{i=1}^{2g}(1-\omega_i T)}{(1-T)(1-qT)}, \quad |\omega_i| = q^{\frac{1}{2}}, \quad g = \mathrm{genus}(C).$$

Now consider a *constant* elliptic curve E over K, i.e., a curve defined by an equation

$$E : Y^2 Z = X^3 + aXZ^2 + bZ^3$$

with the $a, b \in \mathbb{F}_q \subset K$. Let

$$Z(E,T) = \frac{(1-\alpha_1 T)(1-\alpha_2 T)}{(1-T)(1-qT)},$$

$|\alpha_1| = q^{\frac{1}{2}} = |\alpha_2|$.

Proposition 21.1. *For a constant elliptic curve E over $K = \mathbb{F}_q(C)$ (as above), the conjecture of Birch and Swinnerton-Dyer is equivalent to the following statements:*

(a) *the rank of $E(K)$ is equal to the number of pairs (i,j) such that $\alpha_i = \omega_j$;*
(b) $[TS(E/K)]\,\mathrm{disc} <,>= q^g \prod_{\alpha_i \neq \omega_j}(1 - \frac{\omega_j}{\alpha_i})$.

Proof. Elementary, but omitted. □

Theorem 21.2. *In the situation of the proposition, the conjecture of Birch and Swinnerton-Dyer is true.*

Proof. Tate (1966) proved statement (a) of the Proposition, and I proved statement (b) in my thesis (1967). □

In fact, the conjecture of Birch and Swinnerton-Dyer is true under the weaker hypothesis that $j(E) \in \mathbb{F}_q$ (Milne 1975), for example, for all curves of the form

$$Y^2 Z = X^3 + bZ^3, \quad b \in K.$$

Sphere packings. [24] As we noted in (16.6) pairs consisting of a free $\mathbb{Z}$-module of finite rank L and a positive definite quadratic form q on $V =_{df} L \otimes \mathbb{R}$ are of great interest. By Sylvester's theorem, we can choose a basis for V that identifies (V, q) with $(\mathbb{R}^n, X_1^2 + \cdots + X_n^2)$. The bilinear form associated with q is

$$<x, y>= q(x+y) - q(x) - q(y).$$

Given such a pair (L, q), the numbers one needs to compute are

(a) the rank r of L;
(b) the square of the length of the shortest vector
$$m(L) = \inf_{v \in L, v \neq 0} <v, v>;$$
(c) the discriminant of L,
$$\mathrm{disc}\, L = \det(<e_i, e_j>)$$
where $e_1, \ldots, e_r$ is a basis for L.

[24] The best reference for this is Oesteré's Bourbaki talk (Astérisque, 189/190, 1990). There are some uncorrected misprints in the next two pages.

The discriminant is independent of the choice of a basis for L. Let

$$\gamma(L) = m(L)/\operatorname{disc}(L)^{\frac{1}{r}}.$$

The volume a fundamental parallelopiped for L is $\sqrt{\operatorname{disc} L}$. The sphere packing associated with L is formed of spheres of radius $\frac{1}{2}\sqrt{m(L)}$, and therefore its density is

$$d(L) = 2^{-r} b_r \gamma(L)^{\frac{r}{2}}$$

where $b_r = \pi^{r/2}/\Gamma(\frac{r+2}{2})$ is the volume of the r-dimensional unit ball. To maximize $d(L)$, we need to maximise $\gamma(L)$.

Let E be a constant elliptic curve over a field $\mathbb{F}_q(C)$ as above, and let $L = E(\mathbb{Q})/E(\mathbb{Q})_{\text{tors}}$ with the quadratic form $q = 2\widehat{h}$. If we know the ω_i and α_j, part (a) of Theorem 21.2 gives r, and part (b) gives an upper bound for $\operatorname{disc} L$:

$$\operatorname{disc} <,>= q^g \prod_{\alpha_i \neq \omega_j} (1 - \frac{\omega_j}{\alpha_i})/[\mathrm{TS}] \leq q^g \prod_{\alpha_i \neq \omega_j} (1 - \frac{\omega_j}{\alpha_i}).$$

Finally, an easy, but nonelementary argument, shows that

$$m(L) \geq 2[C(k)]/[E(k)]$$

for all finite $k \supset \mathbb{F}_q$ (and $[*] = \operatorname{Card}(*)$). The point is that an element P of $E(K)$ defines a map $u : C \to E$, and $\widehat{h}(P)$ is related to the degree of u. Thus, we get a lower bound for $m(L)$ in terms of the ω_i and α_j.

Example. Consider the curve

$$C : X^{q+1} + Y^{q+1} + Z^{q+1} = 0$$

over $\mathbb{F}_{q^2}$ (note, not over $\mathbb{F}_q$).

Lemma 21.3.

(a) *The curve C is nonsingular, of genus $g = \frac{q(q-1)}{2}$.*
(b) $\#C(\mathbb{F}_{q^2}) = q^3 + 1$.
(c) $Z(C, T) = \frac{(1+qT)^{q(q-1)}}{(1-T)(1-q^2T)}$.

Proof. (a) The partial derivatives of the defining equation are X^q, Y^q, Z^q, and these have no common zero in $\mathbb{P}^2$. Therefore, the curve is nonsingular, and so the formula on p9 shows that it has genus $q(q-1)/2$.

(b) The group $\mathbb{F}_{q^2}^\times$ is cyclic of order $q^2 - 1 = (q+1)(q-1)$, and $\mathbb{F}_q^\times$ is its subgroup of order $q-1$. Therefore, as x runs through $\mathbb{F}_{q^2}$, x^{q+1} takes the value 0 once, and each value in $\mathbb{F}_q^\times$ $q+1$ times. A similar remark applies to y^{q+1} and z^{q+1}. We can scale each solution of $X^{q+1} + Y^{q+1} + Z^{q+1} = 0$ so that $x = 0$ or 1.

Case 1: $x = 1$, $1 + y^{q+1} \neq 0$. There are $q^2 - q - 1$ possibilities for y, and then $q+1$ possibilities for z. Hence $(q^2 - q - 1)(q+1) = q^3 - 2q - 1$ solutions.

Case 2: $x = 1$, $1 + y^{q+1} = 0$. There are $q+1$ possiblities for y and one for z. Hence $q+1$ solutions.

Case 3: $x = 0$. We can take $y = 1$, and then there are $q+1$ possibilities for z.

In sum, there are $q^3 + 1$ solutions.

(c) We know that

$$\#C(\mathbb{F}_q) = 1 + q^2 - \sum_{i=1}^{2g} \omega_i.$$

Therefore $\sum_{i=1}^{2g} \omega_i = q^2 - q^3 = -2gq$. Because $|\omega_i| = q$, this forces $\omega_i = -q$. □

For all q, it is known that there is an elliptic curve E over $\mathbb{F}_{q^2}$, such that $E(\mathbb{F}_{q^2})$ has $q^2 + 2q + 1$ elements (the maximum allowed by the Riemann hypothesis). For such a curve

$$Z(E,T) = \frac{(1+qT)^2}{(1-T)(1-q^2T)}.$$

Proposition 21.4. *Let $L = E(K)/E(K)_{\mathrm{tors}}$ with E and $K = \mathbb{F}_q(C)$ as above. Then:*

(a) *The rank r of L is $2q(q-1)$;*
(b) *$m(L) \geq 2(q-1)$;*
(c) *$[TS(E/K)]\,\mathrm{disc}(L) = q^{q(q-1)}$;*
(d) *$\gamma(L) \geq 2(q-1)/\sqrt{q}$.*

Proof. (a) Since all α_i and ω_j equal $-q$, if follows from (21.2a) that the rank is $2 \times 2g = 2q(q-1)$.

(b) We have

$$m(L) \geq \frac{[C(\mathbb{F}_{q^2})]}{[E(\mathbb{F}_{q^2})]} = \frac{q^3+1}{(q^2+1)^2} > q - 2.$$

(c) This is a special case of (21.2), taking count that our field is $\mathbb{F}_{q^2}$ (not $\mathbb{F}_q$) and $g = q(q-1)/2$.

(d) Follows immediately from the preceding. □

Remark 21.5. (a) Gross and Dummigan have obtained information on the Tate-Shafarevich group in the above, and a closely related, situation. For example, $\mathrm{TS}(E/K)$ is zero if $q = p$ or p^2, and has cardinality at least $p^{p^3(p-1)^3/2}$ if $q = p^3$.

(b) For $q = 2$, L is isomorphic to the lattice denoted D_4, for $q = 3$, to the Coxeter-Todd lattice K_{12}, and for $q = 3$ it is similar to the Leech lattice.

The best description of the work of Elkies and Shioda on the application of elliptic curves to sphere packings is Oesterlé's Séminaire Bourbaki talk, 1989/90, no. 727 (published in Asterisque).

Exercise 21.6. Consider $E : Y^2Z + YZ^2 = X^3$.

(a) Show that E is a nonsingular curve over $\mathbb{F}_2$.

(b) Compute $\#E(\mathbb{F}_4)$, $\mathbb{F}_4$ being the field with 4 elements.

(c) Let K be the field of fractions of the integral domain $\mathbb{F}_4[X,Y]/(X^5 + Y^5 + 1)$, and let $L = E(K)/E(K)_{\mathrm{tors}}$ considered as a lattice in $V = L \otimes \mathbb{R}$ endowed with the height pairing. Compute the rank of L, $m(L)$, and $\gamma(L)$.

[[This exercise is becomes more interesting when $X^5 + Y^5 + 1$ is replaced by $X^3 + Y^3 + 1$.]]

Solution to Exercise 16.10. There are detailed solutions in Knapp, p110–114. Consider the curve $Y^2 = X^3 - 4X$, and take $\alpha = -2$, $\beta = 0$, $\gamma = 2$. Suppose P is a point of infinite order on E—we may assume $P \notin 2E(\mathbb{Q})$. The images of the 2-torsion points $(-2,0)$, $(0,0)$, $(2,0)$ under φ_2 are $(1,1)$, $(1,0)$, and $(0,1)$. Since these fill out all possible nonzero values of φ_2, after possibly replacing a point P of infinite order by $P+Q$, $2Q=0$, $\varphi_2(P)$ will be $(0,0)$. If $\varphi_\infty(P)\neq 0$ (i.e., $\varphi_\infty(P)\neq(+,+)$) then $\varphi_\infty(P)=(+,-)$, which means that

$$x+2=\square,\quad x=-\square,\quad x-2=-\square$$

where $P=(x:y:1)$ and $\square$ denotes a square. Subtracting the first two equations gives $2=\square+\square$. If these squares have even denominators, one finds that

$$0\equiv\square+\square \mod 8$$

with both squares odd integers, which is impossible. Thus the squares $x+2$ and x have odd denominators. Hence

$$2=x-(x-2)=-\square+\square$$

where the first square (hence also the second) has odd denominator. On clearing denominators, one finds that

$$2m^2\equiv-\square+\square \mod 8$$

with m odd and all terms integers. This is impossible. Hence $\varphi_\infty(P)=0$, and so $P\in 2E(\mathbb{Q})$—contradiction.

22. Algorithms for Elliptic Curves

The general Weierstrass equation of an elliptic curve E over a field k is

$$Y^2Z+a_1XYZ+a_3YZ^2=X^3+a_2X^2Z+a_4XZ^2+a_6Z^3.$$

One attaches to the curve the following quantities:

$$\begin{array}{rcl@{\qquad}rcl}
b_2 & = & a_1^2+4a_2 & c_4 & = & b_2^2-24b_4\\
b_4 & = & a_1a_3+2a_4 & c_6 & = & -b_2^3+36b_2b_4-216b_6\\
b_6 & = & a_3^2+4a_6 & & & \text{[Silverman (1st printing) p46}\\
b_8 & = & b_2a_6-a_1a_3a_4+a_2a_3^2-a_4^2 & & & \text{has } c_6=+b_2^3+\cdots]\\
\Delta & = & -b_2^2b_8-8b_4^3-27b_6^2+9b_2b_4b_6 & j & = & c_4^3/\Delta.
\end{array}$$

The curve is nonsingular if and only if $\Delta\neq 0$. The differential $\omega=\frac{dx}{2y+a_1x+a_3}$ is invariant under translation. A Weierstrass equation for an elliptic curve E is unique up to a coordinate transformation of the form

$$x=u^2x'+r\qquad y=u^3y'+su^2x'+t\ ,\quad u,r,s,t\in k,\quad u\neq 0.$$

The quantities Δ, j, ω transform according to the rules:

$$u^{12}\Delta'=\Delta,\quad j'=j,\quad \omega'=u\omega.$$

Two curves become isomorphic over the algebraic closure of k if and only if they have the same j-invariant. When k has characteristic $\neq 2,3$, the terms involving a_1,a_3,a_2 can be eliminated from the Weierstrass equation, and the above equations become those of (5.3).

A *minimum* Weierstrass equation for an elliptic curve E over $\mathbb{Q}$ is an equation of the above form with the $a_i\in\mathbb{Z}$ and Δ minimal. It is unique up to a coordinate transformation of the above form with $r,s,t,u\in\mathbb{Z}$ and $u\in\mathbb{Z}^\times=\{\pm1\}$.

There is an algorithm (due to Tate) for computing the minimum Weierstrass equation, discriminant, conductor, j-invariant, the fibres of its Néron model, etc. of an elliptic curve over $\mathbb{Q}$, which has been implemented in computer programs, for example, in the program Pari, which is specifically designed for calculations in algebraic number theory (including elliptic curves). In the following, I explain how to use Pari as a supercalculator. You can also program it, but for that you will have to read the manual.

To start Pari on the Suns, type:

`gp` (why, I don't know).[25]

An elliptic curve is specified by giving a vector `e=[a1,a2,a3,a4,a6]`

`smallinitell(e)` Computes the 13-component vector

$$[a_1, a_2, a_3, a_4, a_5, b_1, b_4, b_6, b_8, c_4, c_6, \Delta, j]$$

`addell(e,z1,z2)` Computes the sum of the points `z1=[x1,y2]` and `z2=[x2,y2]`.

In the following operations, `e` is usually required to be the output of `smallinitell`.

`globalred(e)` Computes the vector `[N,v]` where `N` is the conductor of the curve and `v=[u,r,s,t]` is the coordinate transformation giving the Weierstrass minimum model with $a_1 = 0$ or 1, $a_2 = 0, 1, -1$, and $a_3 = 0, 1$. Such a model is unique.

`chell(e,v)` Changes e to e', where e' is the 13-component vector corresponding to the curve obtained by the change of coordinates `v=[u,r,s,t]`.

Some of the remaining functions require the curve e to be in minimal Weierstrass form.

`anell(e,k)` Computes the first k of the a_n's for the curve (the coefficients of n^{-s} in the Dirichlet series, e.g., for a good p, $N_p = p + 1 - a_p$).

`apell(e,p)` Computes a_p.

`hell(e,z)` Computes the Néron-Tate canonical height of the point z on e.

`localred(e,p)` Computes the type of the reduction at p using Kodaira's notation ([S1, p359]. It produces `[f,n,...]` where `f` is the exponent of p in the conductor of e, $n = 1$ means good reduction (type I_0), $n = 2, 3, 4$ means reduction of type II,III,IV, $n = 4 + \nu$ means type I_ν, and -1, -2 etc. mean I^* II^* etc..

`lseriesell(e,s,N,A)` Computes the L-series of e at s. Here N is ± the conductor depending on the sign of the functional equation (i.e., the w), and A is a cutoff point for the integral, which must be close to 1 for best speed (see the reference below).

`pointell(e,z)` Computes the coordinates [x,y] where $x = \wp(z)$ and $y = \wp'(z)$ (I think).

`powell(e,n,z)` Computes n times the point z on e.

To quit, type $\backslash q$ (supporting my conjecture that no two programs written by Unixphiles terminate with the same command).

EXAMPLE: `gp`

`?  e=[0,-4,0,0,16]` Defines the elliptic curve $Y^2 = X^3 - 4X^2 + 16$ (see Exercise 19.12).

`%1=[0,-4,0,0,16]`

`?  smallinitell(e)`

`%2=[0,-4,0,0,16,-16,0,64,-256,256,-9728,-45056,-4096/11]` For example, $\Delta = -45056$.

`?  globalred(%2)`

`%3=[11, [2,0,0,4],1]` Computes the minimum conductor and the change of coordinates required to give the minimal equation.

[25] Maybe Go Pari?

```
? chell(%2,[2,0,0,4])
%4=[0,-1,1,0,0,-4,0,1,-1,16,-152,-11,-4096/11]
```

Computes the minimal Weierstrass equation for E, $Y^2 + Y = X^3 - X^2$, which now has discriminant -11 but (of course) the same j-invariant.

```
? anell(%4,13)
%5=[1,-2,-1,2,1,2,-2,0,-2,-2,1,-2,4]
```

In particular,

$$a_p =_{df} p + 1 - N_p = -1, 1, -2, 1, 4 \text{ for } p = 3, 5, 7, 11, 13.$$

```
? localred(%4,2)
%6 = [0,1,...]
```

So E now has good reduction at 2.

```
? localred(%4,11)
%6 = [1,5,...]
```

So E has bad reduction at 11, with conductor 11^1 (hence the singularity is a node), and the Kodaira type of the special fibre of the Néron model is I_1.

Pari is available (free!) by anonymous ftp from `math.ucla.edu`—it runs on PC's and Macs. Henri Cohen, the main author of Pari, has also written the best book on computational algebraic number theory "A Course in Computational Algebraic Number Theory", which explains most of the algorithms incorporated into Pari.

23. The Riemann Surfaces $X_0(N)$

We wish to understand the L-series of an elliptic curve E over $\mathbb{Q}$, i.e., we wish to understand the sequence of numbers

$$N_2, N_3, N_5, N_7, \ldots, N_p, \ldots \qquad N_p = \#E(\mathbb{F}_p).$$

There is no direct way of doing this. Instead, we shall see how the study of modular curves and modular forms leads to functions that are candidates for being the L-series of an elliptic curve over $\mathbb{Q}$, and then we shall see how Wiles showed that the L-series of (almost all) elliptic curves over $\mathbb{Q}$ do arise from modular forms.

The notion of a Riemann surface. Let X be a connected Hausdorff topological space. A *coordinate neighbourhood* for X is a pair (U, z) with U an open subset of X and z a homeomorphism of U onto an open subset of the complex plane $\mathbb{C}$. Two coordinate neighbourhoods (U_1, z_1) and (U_2, z_2) are *compatible* if the function

$$z_1 \circ z_2^{-1} : z_2(U_1 \cap U_2) \to z_1(U_1 \cap U_2)$$

is holomorphic with nowhere vanishing derivative. A family of coordinate neighbourhoods $(U_i, z_i)_{i \in I}$ is a *coordinate covering* if $X = \cup U_i$ and (U_i, z_i) is compatible with (U_j, z_j) for all pairs $(i, j) \in I \times I$. Two coordinate coverings are said to be *equivalent* if the their union is also a coordinate covering. This defines an equivalence relation on the set of coordinate coverings, and we call an equivalence class a *complex structure* on X. A Hausdorff topological space X together with a complex structure is a *Riemann surface.*

Let $\mathcal{U} = (U_i, z_i)$ be a coordinate covering of X. A function $f : U \to \mathbb{C}$ on an open subset U of X is said to be *holomorphic* relative to $\mathcal{U}$ if $f \circ z_i^{-1} : z_i(U \cap U_i) \to \mathbb{C}$ is holomorphic for all $i \in I$. If f is holomorphic relative to one coordinate covering, then it is holomorphic relative to every equivalent covering, and so it will be said to be holomorphic for the complex structure on X.

Recall that a *meromorphic function* on an open subset U of $\mathbb{C}$ is a holomorphic function f on $U - \Xi$ for some discrete subset $\Xi \subset U$ that has at worst a pole at each point of Ξ, i.e., such that for each $a \in \Xi$, there exists an m such that $(z-a)^m f(z)$ is holomorphic in some neighbourhood of a. A *meromorphic function* on an open subset of a Riemann surface is defined similarly.

A map $f : X \to X'$ from one Riemann surface to a second is *holomorphic* if $g \circ f$ is holomorphic whenever g is a holomorphic function on an open subset of X'. For this, it suffices to check that for every point P in X, there are coordinate neighbourhoods (U, z) of P and (U', z') of $f(P)$ such that $z' \circ f \circ z^{-1} : z(U) \to z'(U')$ is holomorphic. An *isomorphism* of Riemann surfaces is a bijective holomorphic map whose inverse is also holomorphic.

Example 23.1. Any open subset of $\mathbb{C}$ is a Riemann surface with a single coordinate neighbourhood—U itself with the identity function z.

Example 23.2. Let X be the unit sphere

$$S_2 : X^2 + Y^2 + Z^2 = 1$$

in $\mathbb{R}^3$, and let P be the north pole $(0,0,1)$. Stereographic projection from P is a map

$$(x, y, z) \mapsto \frac{x+iy}{1-z} : X - P \to \mathbb{C}.$$

Take this to be a coordinate neighbourhood for X. Stereographic projection from the south pole S gives a second coordinate neighbourhood. These two coordinate neighbourhoods define a complex structure on X, and X together with the complex structure is called the *Riemann sphere.*

Example 23.3. Let $X = \mathbb{R}^2/\mathbb{Z}^2$. For any $\tau \in \mathbb{H}$, the homeomorphism

$$(x, y) \mapsto x\tau + y : \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{C}/\mathbb{Z}\tau + \mathbb{Z}$$

defines a complex structure on X. The Riemann surfaces corresponding to τ and τ' are isomorphic if and only $j(\tau) = j(\tau')$ (see Section 10). In particular, this shows that there are uncountably many nonisomorphic complex structures on the topological space X.

Quotients of Riemann surfaces by group actions. We shall need to define Riemann surfaces as the quotients of other (simpler) Riemann surfaces by group actions. This can be quite complicated. The following examples will help.

Example 23.4. Let $n \in \mathbb{Z}$ act on $\mathbb{C}$ by $z \mapsto z + n$. Topologically $\mathbb{C}/\mathbb{Z}$ is a cylinder. We can give it a complex structure as follows: let $\pi : \mathbb{C} \to \mathbb{C}/\mathbb{Z}$ be the quotient map; for any $P \in \mathbb{C}/\mathbb{Z}$ and $Q \in f^{-1}(P)$ we can find open neighbourhoods U of P and V of Q such that $\pi : U \to V$ is a homeomorphism; take any such pair $(U, \pi^{-1} : U \to V)$ to be a coordinate function.

For any open $U \subset \mathbb{C}/\mathbb{Z}$, a function $f : U \to \mathbb{C}$ is holomorphic for this complex structure if and only if $f \circ \pi$ is holomorphic. Thus the holomorphic functions f on $U \subset \mathbb{C}/\mathbb{Z}$ can be identified with the holomorphic functions g on $\pi^{-1}(U)$ invariant under $\mathbb{Z}$, i.e., such that $g(z+1) = g(z)$.

For example, $q(z) = e^{2\pi i z}$ defines a holomorphic function on $\mathbb{C}/\mathbb{Z}$. In fact, it gives an isomorphism $\mathbb{C}/\mathbb{Z} \to \mathbb{C}^\times$ whose in inverse $\mathbb{C}^\times \to \mathbb{C}/\mathbb{Z}$ is (by definition) $(2\pi i)^{-1} \cdot \log$.

Example 23.5. Let D be the open unit disk $\{z \mid |z| < 1\}$, and let Δ be a finite group acting on D. The Schwarz lemma implies that $\mathrm{Aut}(D) = \{z \in \mathbb{C} \mid |z| = 1\} \approx \mathbb{R}/\mathbb{Z}$, and it follows that Δ is a finite cyclic group. Let $z \mapsto \zeta z$ be its generator and suppose that ζ has order m, i.e., $\zeta^m = 1$. Then z^m is invariant under Δ, and so defines a function on $\Delta\backslash D$, which in fact is a homeomorphism $\Delta\backslash D \to D$, and therefore defines a complex structure on $\Delta\backslash D$.

Let $\pi : D \to \Delta\backslash D$ be the quotient map. Then $f \mapsto f \circ \pi$ identifies the space of holomorphic functions on $U \subset \Delta\backslash D$ with the space of holomorphic functions on $\pi^{-1}(U)$ such that $f(\zeta z) = f(z)$, i.e., which are of the form $f(z) = h(z^m)$ with h holomorphic. Note that if $\pi(Q) = P$, then $\mathrm{ord}_P(f) = \frac{1}{m}\mathrm{ord}_Q(f \circ \pi)$.

Let Γ be a group acting on a Riemann surface X. A *fundamental domain* for Γ is a connected open subset D of X such that

(a) no two points of D lie in the same orbit of Γ;
(b) the closure $\bar{D}$ of D contains at least one element from each orbit.

For example,

$$D = \{z \in \mathbb{C} \mid 0 < \Re(z) < 1\}$$

is a fundamental domain for $\mathbb{Z}$ acting on $\mathbb{C}$ (as in 23.4), and

$$D_0 = \{z \in D \mid 0 < \arg z < \pi/n\}$$

is a fundamental domain for $\mathbb{Z}/n\mathbb{Z}$ acting on the unit disk (as in 23.5).

The Riemann surfaces $X(\Gamma)$. Let Γ be a subgroup of finite index in $\mathrm{SL}_2(\mathbb{Z})$. We want to define the structure of a Riemann surface on the quotient $\Gamma\backslash\mathbb{H}$. This we can do, but the resulting surface will not be compact. Instead, we need to form a quotient $\Gamma\backslash\mathbb{H}^*$ where $\mathbb{H}^*$ properly contains $\mathbb{H}$.

The action of $\mathrm{SL}_2(\mathbb{Z})$ *on the upper half plane.* Recall that $\mathrm{SL}_2(\mathbb{Z})$ acts on $\mathbb{H} = \{z \mid \Im(z) > 0\}$ according to

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} z = \frac{az+b}{cz+d}.$$

Note that $-I = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ acts trivially on $\mathbb{H}$, and so the action factors through

$$\mathrm{PSL}_2(\mathbb{Z}) = \mathrm{SL}_2(\mathbb{Z})/\{\pm I\}.$$

Let

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \text{ so } Sz = \frac{-1}{z},$$

and

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \text{ so } Tz = z+1.$$

Then

$$S^2 = 1, \quad (ST)^3 = 1 \text{ in } \mathrm{PSL}_2(\mathbb{Z}).$$

Proposition 23.6. *Let*

$$D = \left\{ z \in \mathbb{H} \mid |z| > 1, \quad -\frac{1}{2} < \Re(z) < \frac{1}{2} \right\}.$$

(a) *D is a fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$; moreover, two elements z and z' of $\bar{D}$ are in the same orbit if and only if*
   (i) *$\Re(z) = \pm\frac{1}{2}$ and $z' = z \pm 1$ (so $z' = Tz$ or $z = Tz'$);*
   (ii) *$|z| = 1$ and $z' = -1/z$ $(= Sz)$.*

(b) *For $z \in \bar{D}$, the stabilizer of z is $\neq \{\pm I\}$ if and only if $z = i$, in which case the stabilizer is $<S>$, or $\rho = e^{2\pi i/6}$, in which case the stabilizer is $<TS>$, or ρ^2, in which case it is $<ST>$.*

(c) *The group $\mathrm{PSL}_2(\mathbb{Z})$ is generated by S and T.*

Proof. Let $\Gamma' =< S, T >$. One first shows that $\Gamma'\bar{D} = \mathbb{H}$, from which (a) and (b) follow easily. For (c), let $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, and choose a point z_0 in D. There exists a γ' in Γ' such that $\gamma z_0 = \gamma' z_0$, and it follows from (b) that $\gamma'\gamma^{-1} = \pm I$. For the details, see (Serre, Course on Arithmetic, VII.1.2). □

Remark 23.7. Let Γ be a subgroup of finite index in $\mathrm{SL}_2(\mathbb{Z})$, and write

$$\mathrm{SL}_2(\mathbb{Z}) = \Gamma\gamma_1 \cup \ldots \cup \Gamma\gamma_m \quad \text{(disjoint union)}.$$

Then $D' = \cup\gamma_i D$ satisfies the conditions to be a fundamental domain for Γ, except that it won't be connected. However, it is possible to choose the γ_i so that the closure of D' is connected, in which case the interior of the closure will be a fundamental domain. for Γ.

The extended upper half plane. The elements of $\mathrm{SL}_2(\mathbb{Z})$ act on $\mathbb{P}^1(\mathbb{C})$ by projective linear transformations,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(z_1 : z_2) = (az_1 + bz_2 : cz_1 + dz_2).$$

Identify $\mathbb{H}$, $\mathbb{Q}$, and $\{\infty\}$ with subsets of $\mathbb{P}^1(\mathbb{C})$ according to

$$\begin{array}{rcll} z & \leftrightarrow & (z:1) & z \in \mathbb{H} \\ r & \leftrightarrow & (r:1) & r \in \mathbb{Q} \\ \infty & \leftrightarrow & (1:0) & \end{array}.$$

The action of $\mathrm{SL}_2(\mathbb{Z})$ stabilizes $\mathbb{H}^* =_{df} \mathbb{H} \cup \mathbb{Q} \cup \{\infty\}$. For example, for $z \in \mathbb{H}$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(z : 1) = (az + b : cz + d) = (\frac{az+b}{cz+d} : 1)$$

as usual, and for $r \in \mathbb{Q}$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(r : 1) = (ar + b : cr + d) = \begin{cases} (\frac{ar+b}{cr+d} : 1) & r \neq -\frac{d}{c} \\ \infty & r = -\frac{d}{c} \end{cases},$$

and, finally,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\infty = (a : c) = \begin{cases} (\frac{a}{c} : 1) & c \neq 0, \\ \infty & c = 0 \end{cases}.$$

Thus in passing from $\mathbb{H}$ to $\mathbb{H}^*$, we have added one additional $\mathrm{SL}_2(\mathbb{Z})$ orbit. The points in $\mathbb{H}^*$ not in $\mathbb{H}$ are often called the *cusps.*

We make $\mathbb{H}^*$ into a topological space as follows: the topology on $\mathbb{H}$ is that inherited from $\mathbb{C}$; the sets

$$\{z \mid \Im(z) > M\}, \quad M > 0$$

form a fundamental system of neighbourhoods of ∞; the sets

$$\{z \mid |z-(a+ir)| < r\} \cup \{a\}$$

form a fundamental system of neighbourhoods of $a \in \mathbb{Q}$. One shows that $\mathbb{H}^*$ is Hausdorff, and that the action of $\mathrm{SL}_2(\mathbb{Z})$ is continuous.

The topology on $\Gamma\backslash\mathbb{H}^*$. Recall that if $\pi : X \to Y$ is a surjective map and X is a topological space, then the *quotient topology* on Y is that for which a set U is open if and only of $\pi^{-1}(U)$ is open. In general the quotient of a Hausdorff space by a group action will not be Hausdorff, even if the orbits are closed—one needs that distinct orbits have disjoint open neighbourhoods.

Let Γ be a subgroup of finite index in $\mathrm{SL}_2(\mathbb{Z})$. One can show that such a Γ acts *properly discontinuously* on $\mathbb{H}$, i.e., that for any pair of points $x, y \in \mathbb{H}$, there exist neighbourhoods U of x and V of y such that

$$\{\gamma \in \Gamma \mid \gamma U \cap V \neq \emptyset\}$$

is finite. In particular, this implies that the stabilizer of any point in $\mathbb{H}$ is finite (which we knew anyway).

Proposition 23.8. (a) *For any compact sets A and B of $\mathbb{H}$, $\{\gamma \in \Gamma \mid \gamma A \cap B \neq \emptyset\}$ is finite.*

(b) *Any $z \in \mathbb{H}$ has a neighbourhood U such that*

$$\gamma U \cap U \neq \emptyset$$

only if $\gamma z = z$.

(c) *For any points x, y of $\mathbb{H}$ not in the same Γ-orbit, there exist neighbourhoods U of x and V of y such that $\gamma U \cap V = \emptyset$ for all $\gamma \in \Gamma$*

Proof. (a) This follows easily from the fact that Γ acts properly discontinuously.

(b) Let V be compact neighbourhood of z. From (a) we know that there is only a finite set $\{\gamma_1, \ldots, \gamma_n\}$ of Γ such that $V \cap \gamma_i V \neq \emptyset$. Let $\gamma_1, \ldots, \gamma_s$ be the γ_i's fixing z, and for each $i > s$, choose disjoint neighbourhoods V_i of z and W_i of $\gamma_i z$, and set

$$U = V \cap (\cap_{i>s} V_i \cap \gamma_i^{-1} W_i).$$

For $i > s$, $\gamma_i U \subset W_i$, which is disjoint from V_i, which contains U.

(c) Choose compact neighbourhoods A of x and B of y, and let $\gamma_1, \ldots, \gamma_n$ be the elements of Γ such that $\gamma_i A \cap B \neq \emptyset$. We know $\gamma_i x \neq y$, and so we can find disjoint neighbourhoods U_i and V_i of $\gamma_i x$ and y. Take

$$U = A \cap \gamma_1^{-1} U_1 \cap \ldots \cap \gamma_n^{-1} U_n, \quad V = B \cap V_1 \cap \ldots \cap V_n.$$

□

Corollary 23.9. *The space $\Gamma\backslash\mathbb{H}$ is Hausdorff.*

Proof. Let x and y be points of $\mathbb{H}$ not in the same Γ-orbit, and choose neighbourhoods U and V of x and y as in (c) of the last proposition. Then ΓU and ΓV are disjoint neighbourhoods of Γx and Γy. □

In fact, $\Gamma\backslash\mathbb{H}^*$ will be Hausdorff, and compact.

The complex structure on $\Gamma_0(N)\backslash\mathbb{H}^*$. The subgroups of $\mathrm{SL}_2(\mathbb{Z})$ that we shall be especially interested in are

$$\Gamma_0(N) = \left\{ \left(\begin{array}{cc} a & b \\ c & d \end{array} \right) \middle| c \equiv 0 \mod N \right\}.$$

We let $\Gamma_0(1) = \mathrm{SL}_2(\mathbb{Z})$.

For $z_0 \in \mathbb{H}$, choose a neighbourhood V of z_0 such that

$$\gamma V \cap V \neq \emptyset \implies \gamma z_0 = z_0,$$

and let $U = \pi(V)$—it is open because $\pi^{-1}U = \cup\gamma V$ is open.

If the stabilizer of z_0 in $\Gamma_0(N)$ is $\pm I$, then $\pi : V \to U$ is a homeomorphism, with inverse φ say, and we require (U, φ) to be a coordinate neighbourhood.

If the stabilizer of z_0 in $\Gamma_0(N)$ is $\neq \{\pm I\}$, then it is a cyclic group of order $2m$ with $m = 2$ or 3 (and its stabilizer in $\Gamma_0(N)/\{\pm I\}$ has order 2 or 3)—see (23.6b). The fractional linear transformation

$$\lambda : \mathbb{H} \to D, \quad z \mapsto \frac{z - z_0}{z - \bar{z}_0},$$

carries z_0 to 0 in the unit disk D. There is a well-defined map $\varphi : U \to \mathbb{C}$ such that $\varphi(\pi(z)) = \lambda(z)^n$, and we require (U, φ) to be a coordinate neighbourhood (cf. Example 23.5).

Next consider $z_0 = \infty$. Choose V to be the neighbourhood $\{z \mid \Im(z) > 2\}$ of ∞, and let $U = \pi(V)$. If

$$z \in V \cap \gamma V, \quad \gamma = \left(\begin{array}{cc} a & b \\ c & d \end{array} \right) \in \Gamma_0(N),$$

then

$$2 \leq \Im(\gamma z) = \frac{\Im(z)}{|cz+d|^2} \leq \frac{1}{|c|^2 \Im(z)} \leq \frac{1}{2|c|^2}$$

and so $c = 0$. Therefore

$$\gamma = \pm \left(\begin{array}{cc} 1 & m \\ 0 & 1 \end{array} \right),$$

and so there is a well-defined map $\varphi : U \to \mathbb{C}$ such that $\varphi(\pi(z)) = e^{2\pi i z}$, and we require (U, φ) to be a coordinate neighbourhood (cf. Example 23.4).

For $z_0 \in \mathbb{Q}$, we choose a $\beta \in \mathrm{SL}_2(\mathbb{Z})$ such that $\beta(z_0) = \infty$, and proceed similarly.

Proposition 23.10. *The coordinate neighbourhoods defined above are compatible, and therefore define on $\Gamma_0(N)\backslash\mathbb{H}^*$ the structure of a Riemann surface.*

Proof. Omitted. □

Write $X_0(N)$ for the Riemann surface $\Gamma_0(N)\backslash\mathbb{H}^*$, and $Y_0(N)$ for its open subsurface $\Gamma_0(N)\backslash\mathbb{H}$.

The genus of $X_0(N)$. The genus of a Riemann surface can be computed by "triangulating" it, and using the formula

$$2-2g=V-E+F$$

where V is the number of vertices, E is the number of edges, and F is the number of faces. This, presumably, is the original definition of the genus. For example, the sphere may be triangulated by projecting out from a regular tetrahedron. Then $V=4$, $E=6$, and $F=4$, so that $g=0$ as expected.

Proposition 23.11. *The Riemann surface $X_0(1)$ has genus zero.*

Proof. One gets a fake triangulation of the sphere by taking taking as vertices three points on the equator, and the upper and lower hemispheres as the faces. This gives the correct genus

$$2=3-3+2$$

but it violates the usual definition of a triangulation, which requires that any two triangles intersect in a single side, a single vertex, or not at all. It can be made into a valid triangulation by adding the north pole as a vertex, and joining it to the three vertices on the equator.

One gets a fake triangulation of $X_0(1)$ by taking the three vertices ρ, i, and ∞ and the obvious curves joining them (two on the boundary of D and one the nimaginary axis from i to ∞). It can be turned into a valid triangulation by adding a fourth point not on any of these curves, and joining it to ρ, i, and ∞. □

For a finite mapping $\pi : Y \to X$ of compact Riemann surfaces, the Hurwitz genus formula relates the two genuses:

$$2g_Y-2=(2g_X-2)m+\sum_{Q\in Y}(e_Q-1).$$

Here m is the degree of the mapping, so that $\pi^{-1}(P)$ has m elements except for finitely many P, and e_Q is the ramification index, so that $e_Q=1$ unless at least two sheets come together at Q above $\pi(Q)$ in which case it is the number of such sheets.

For example, if E is the elliptic curve

$$E: Y^2Z=X^3+aXZ^2+bZ^3, \quad a,b\in\mathbb{C}, \quad \Delta\neq 0,$$

and π is the map

$$\infty\mapsto\infty,\ (x:y:z)\mapsto(x:z): E(\mathbb{C})\to\mathbb{P}^1(\mathbb{C})$$

then $m=2$ and $e_Q=1$ except for $Q=\infty$ or one of the points of order 2 on E, in which case $e_Q=2$. This is consistent with $E(\mathbb{C})$ having genus 1 and $\mathbb{P}^1(\mathbb{C})$ (the Riemann sphere) having genus 0.

The Hurwitz genus formula can be proved without too much difficulty by triangulating Y in such a way that the ramification points are vertices and such that the triangulation of Y lies over a triangulation of X.

Now one can compute the genus of $X_0(N)$ by studying the quotient map $X_0(N)\to X_0(1)$. The only (possible) ramification points are those $\Gamma_0(1)$-equivalent to one of i, ρ, or ∞.

Explicit formulas can be found in Shimura, Arithmetic Theory of Automorphic Functions, pp23-25. For example, one finds that, for p a prime > 3,

$$\text{genus}(X_0(p)) = \begin{cases} n-1 & \text{if } p = 12n+1 \\ n & \text{if } p = 12n+5, 12n+7 \\ n+1 & \text{if } p = 12n+11. \end{cases}$$

Moreover,

$$\begin{aligned} g = 0 \quad &\text{if} \quad N = 1, \dots, 10, 12, 13, 16, 18, 25; \\ g = 1 \quad &\text{if} \quad N = 11, 14, 15, 17, 19, 20, 21, 24, 27, 32, 36, 49 \\ g = 2 \quad &\text{if} \quad N = 22, 23, 26, 28, 29, 31, 37, 50. \end{aligned}$$

Exercise 23.12. (a) For a prime p, show that the natural action of $\Gamma_0(p)$ on $\mathbb{P}^1(\mathbb{Q})$ has only two orbits, represented by 0 and $\infty = (1 : 0)$. Deduce that $X_0(p) \setminus Y_0(p)$ has exactly two elements.

(b) Define $\Delta(z) = \Delta(\mathbb{Z}z + \mathbb{Z})$ (see p51), so that Δ is a basis for the $\mathbb{C}$-vector space of cusp forms of weight 12 for $\Gamma_0(1)$. Define $\Delta_{11}(z) = \Delta(11z)$, and show that it is a cusp form of weight 12 for $\Gamma_0(11)$. Deduce that $\Delta \cdot \Delta_{11}$ is a cusp form of weight 24 for $\Gamma_0(11)$.

(c) Assume Jacobi's formula:

$$\Delta(z) = (2\pi)^{12} q \prod_{n=1}^{\infty} (1 - q^n)^{24},$$

($q = e^{2\pi i z}$), and that $\mathcal{S}_2(\Gamma_0(11))$ has dimension 1. Show that

$$F(z) = q \prod_{n=1}^{\infty} (1 - q^n)^2 (1 - q^{11n})^2,$$

is a cusp form of weight 2 for $\Gamma_0(11)$. [Hint: Let f be a nonzero element of $\mathcal{S}_2(\Gamma_0(11))$, and let $g = \Delta \cdot \Delta_{11}$. Show that f^{12}/g is holomorphic on $\mathbb{H}^*$ and invariant under $\Gamma_0(1)$, and is therefore constant (because the only holomorphic functions on a compact Riemann surface are the constant functions). The only real difficulty is in handling the cusp 0, since I have more-or-less ignored cusps other than ∞.]

24. $X_0(N)$ as an Algebraic Curve over $\mathbb{Q}$

In the last section, we defined compact Riemann surfaces $X_0(N)$. A general theorem states that any compact Riemann surface X can be identified with the set of complex points of a unique nonsingular projective algebraic curve[26] C over $\mathbb{C}$. However, in general C can't be defined over $\mathbb{Q}$ (or even $\mathbb{Q}^{\text{al}}$)—consider for example a Riemann surface $\mathbb{C}/\Lambda$ whose j-invariant is transcendental—and when C can be defined over $\mathbb{Q}$, in general, it can't be defined in any canonical way—consider an elliptic curve E over $\mathbb{C}$ with $j(E) \in \mathbb{Q}$.

In this section, we'll see that $X_0(N)$ has the remarkable property that it *is* the set of complex points of a *canonical* curve over $\mathbb{Q}$.

[26] The inconsistency between "surface" and "curve" is due to the analysts inability to count.

Modular functions. For a connected compact Riemann surface X, the meromorphic functions on X form a field of transcendence degree 1 over $\mathbb{C}$.

For a subgroup Γ of finite index in $\mathrm{SL}_2(\mathbb{Z})$, the meromorphic functions on $\Gamma\backslash\mathbb{H}^*$ are called the *modular functions* for Γ. If $\pi : \mathbb{H} \to \Gamma\backslash\mathbb{H}^*$ is the quotient map, then $g \mapsto \pi \circ g$ identifies the modular functions for Γ with the functions f on $\mathbb{H}$ such that

(a) f is meromorphic on $\mathbb{H}$;
(b) for any $\gamma \in \Gamma$, $f(\gamma z) = f(z)$;
(c) f is meromorphic at the cusps (i.e., at the points of $\mathbb{H}^* \setminus \mathbb{H}$).

The meromorphic functions on $X_0(1)$. Let S be the Riemann sphere $S = \mathbb{C} \cup \{\infty\}$ (better, $S = \mathbb{P}^1(\mathbb{C}) = \mathbb{A}^1(\mathbb{C}) \cup \{(1:0)\}$. The meromorphic functions on S are the rational functions of z, and the automorphisms of S are the fractional-linear transformations,

$$z \mapsto \frac{az+b}{cz+d}, \quad a,b,c,d \in \mathbb{C}, \quad ad-bc \neq 0.$$

In fact, $\mathrm{Aut}(S) = \mathrm{PGL}_2(\mathbb{C}) =_{df} \mathrm{GL}_2(\mathbb{C})/\mathbb{C}^\times$. Moreover, given two sets $\{P_1, P_2, P_3\}$ and $\{Q_1, Q_2, Q_3\}$ of distinct points on S, there is a unique fractional-linear transformation sending each P_i to Q_i. (The proof of the last statement is an easy exercise in linear algebra: given two sets $\{L_1, L_2, L_3\}$ and $\{M_1, M_2, M_3\}$ of distinct lines through the origin in $\mathbb{C}^2$, there is a linear transformation carrying each L_i to M_i, and the linear transformation is unique up to multiplication by a nonzero constant.)

We use ∞, i, and ρ to denote also the images of these points on $X_0(1)$.

Proposition 24.1. *There exists a unique meromorphic function J on $X_0(1)$ that is holomorphic except at ∞, where it has a simple pole, and takes the values*

$$J(i) = 1, \quad J(\rho) = 0.$$

Moreover, the meromorphic functions on $X_0(1)$ are the rational functions of J.

Proof. We saw in the last section that $X_0(1)$ is isomorphic (as a Riemann surface) to the Riemann sphere S. Let $f : X_0(1) \to S$ be an isomorphism, and let P, Q, R be the images of ρ, i, ∞. There is a unique fractional-linear transformation L sending P, Q, R to $0, 1, \infty$, and the composite $L \circ f$ has the required properties. If J' is a second such function, then the composite $J' \circ J^{-1}$ is an automorphism of S fixing $0, 1, \infty$, and so is the identity map. Under this isomorphism, the function z on S corresponds to the function J on $X_0(1)$. □

In minor disagreement with the notation in Section 10, I write

$$G_{2k}(\Lambda) = \sum_{\omega \in \Lambda, \omega \neq 0} \frac{1}{\omega^{2k}},$$

for a lattice $\Lambda \subset \mathbb{C}$, and

$$G_{2k}(z) = G_{2k}(\mathbb{Z}z + \mathbb{Z}), \quad g_4(z) = 60G_4(z), \quad g_6(z) = 140G_6(z), \quad z \in \mathbb{H}.$$

Then $(\wp, \wp')$ maps $\mathbb{C}/\mathbb{Z}z + \mathbb{Z}$ onto the elliptic curve

$$Y^2Z = 4X^3 - g_4(z)XZ^2 - g_6(z)Z^3, \quad \Delta = g_4(z)^3 - 27g_6(z)^2 \neq 0,$$

whose j-invariant is

$$j(z) = \frac{1728 g_4(z)^3}{\Delta}.$$

From their definitions, it is clear that $G_{2k}(z)$, $\Delta(z)$, and $j(z)$ are invariant under $T : z \mapsto z+1$, and so can be expressed in terms of the variable $q = e^{2\pi i z}$. In Serre, Cours d'Arithmétique, VII, one can find the following expansions:

$$\begin{aligned} G_{2k}(z) &= 2\zeta(2k) + \frac{2(2\pi i)^{2k}}{(2k-1)!}\sum_{n=1}^{\infty}\sigma_{2k-1}(n)q^n, \quad \sigma_k(n) = \sum_{d|n} d^k, \\ \Delta &= (2\pi)^{12}(q - 24q^2 + 252q^3 - 1472q^4 + \cdots), \\ j &= \frac{1}{q} + 744 + 196884q + 21493760q^2 + \sum_{n=3}^{\infty} c(n)q^n, \quad c(n) \in \mathbb{Z}. \end{aligned}$$

The proof of the formula for $G_{2k}(z)$ is elementary, and the others follow from it together with elementary results on $\zeta(2k)$. The factor 1728 was traditionally included in the formula for j so that it has residue 1 at infinity.

The function j is invariant under $\mathrm{SL}_2(\mathbb{Z})$, because $j(z)$ depends only on the lattice $\mathbb{Z}z + \mathbb{Z}$. Moreover:

> $j(\rho) = 0$, because $\mathbb{C}/\mathbb{Z}\rho + \mathbb{Z}$ has complex multiplication by $\rho^2 = \sqrt[3]{1}$, and therefore is of the form $Y^2 = X^3 + b$, which has j-invariant 0.
> $j(i) = 1728$, because $\mathbb{C}/\mathbb{Z}i + \mathbb{Z}$ has complex multiplication by i, and therefore is of the form $Y^2 = X^3 + aX$.

Consequently $j = 1728J$, and the field of meromorphic functions on $X_0(N)$ is $\mathbb{C}(j)$.

The meromorphic functions on $X_0(N)$. Define j_N to be the function on $\mathbb{H}$ such that $j_N(z) = j(Nz)$. For $\gamma \in \Gamma_0(1)$, one is tempted to say

$$j_N(\gamma z) = j(N\gamma z) = j(\gamma N z) = j(Nz) = j_N(z),$$

but, this is false in general, because $N\gamma z \neq \gamma N z$. However, it is true that $j_N(\gamma z) = j_N(z)$ if $\gamma \in \Gamma_0(N)$. In fact, let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$, so that $c = Nc'$ with $c' \in \mathbb{Z}$. Then

$$j_N(\gamma z) = j(\frac{Naz + Nb}{cz + d}) = j(\frac{a(Nz) + Nb}{c'(Nz) + d}) = j(\gamma' N z)$$

where $\gamma' = \begin{pmatrix} a & Nb \\ c' & b \end{pmatrix} \in \Gamma_0(1)$, so

$$j(\gamma' N z) = j(Nz) = j_N(z).$$

Thus, we see that j_N is invariant under $\Gamma_0(N)$, and therefore defines a meromorphic function on $X_0(N)$.

Theorem 24.2. *The field of meromorphic functions on $X_0(N)$ is $\mathbb{C}(j, j_N)$.*

Proof. The curve $X_0(N)$ is a covering of $X_0(1)$ of degree $m = (\Gamma_0(1) : \Gamma_0(N))$. The general theory implies that the field of meromorphic functions on $X_0(N)$ has degree m over $\mathbb{C}(j)$, but we shall prove this again. Let $\{\gamma_1 = 1, ..., \gamma_m\}$ be a set of representatives for the right cosets of $\Gamma_0(N)$ in $\Gamma_0(1)$, so that,

$$\Gamma_0(1) = \bigcup_{i=1}^{m} \Gamma_0(N)\gamma_i \quad \text{(disjoint union)}.$$

For any $\gamma \in \Gamma_0(1)$, $\{\gamma_1\gamma, ..., \gamma_m\gamma\}$ is also a set of representatives for the right cosets of $\Gamma_0(N)$ in $\Gamma_0(1)$—the family $(\Gamma_0(N)\gamma_i\gamma)$ is just a permutation of the family $(\Gamma_0(N)\gamma_i)$.

If $f(z)$ is a modular function for $\Gamma_0(N)$, then $f(\gamma_i z)$ depends only on the coset $\Gamma_0(N)\gamma_i$. Hence the functions $\{f(\gamma_i\gamma z)\}$ are a permutation of the functions $\{f(\gamma_i z)\}$, and any symmetric polynomial in the $f(\gamma_i z)$ is invariant under $\Gamma_0(1)$; since such a polynomial obviously satisfies the other conditions, it is a modular function for $\Gamma_0(1)$, and hence a rational function of j. Therefore $f(z)$ satisfies a polynomial of degree m with coefficients in $\mathbb{C}(j)$, namely, $\prod(Y - f(\gamma_i z))$. Since this holds for every meromorphic function on $X_0(N)$, we see that the field of such functions has degree at most m over $\mathbb{C}(j)$.

Next I claim that all the $f(\gamma_i z)$ are conjugate to $f(z)$ over $\mathbb{C}(j)$: for let $F(j, Y)$ be the minimum polynomial of $f(z)$ over $\mathbb{C}(j)$, so that $F(j, Y)$ is monic and irreducible when regarded as a polynomial in Y with coefficients in $\mathbb{C}(j)$; on replacing z with $\gamma_i z$ and remembering that $j(\gamma_i z) = j(z)$, we find that $F(j(z), f(\gamma_i z)) = 0$, which proves the claim.

If we can show that the functions $j(N\gamma_i z)$ are distinct, then it will follow that the minimum polynomial of j_N over $\mathbb{C}(j)$ has degree m, and that the field of meromorphic functions on $X_0(N)$ has degree m over $\mathbb{C}(j)$, and is generated by j_N.

Suppose $j(N\gamma_i z) = j(N\gamma_j z)$ for some $i \neq j$. Recall that j defines an isomorphism $\Gamma_0(1)\backslash\mathbb{H}^* \to S$ (Riemann sphere), and so

$$j(N\gamma_i z) = j(N\gamma_j z)\text{all } z \implies \exists\gamma \in \Gamma_0(1) \text{ such that } N\gamma_i z = \gamma N\gamma_j z \text{ all } z,$$

and this implies that

$$\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}\gamma_i = \pm\gamma\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}\gamma_j.$$

Hence $\gamma_i\gamma_j^{-1} \in \Gamma_0(1) \cap \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}^{-1}\Gamma_0(1)\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} = \Gamma_0(N)$, which contradicts the fact that γ_i and γ_j lie in different cosets. □

We saw in the proof that the minimum polynomial of j_N over $\mathbb{C}(j)$ is

$$F(j, Y) = \prod_{i=1}^{m}(Y - j(N\gamma_i z)).$$

The symmetric polynomials in the $j(N\gamma_i z)$ are holomorphic on $\mathbb{H}$. As they are rational functions of $j(z)$, they must in fact be polynomials in $j(z)$, and so $F_N(j, Y) \in \mathbb{C}[j, Y]$ (rather than $\mathbb{C}(j)[Y]$).

On replacing j with the variable X, we obtain a polynomial $F_N(X, Y) \in \mathbb{C}[X, Y]$,

$$F_N(X, Y) = \sum c_{r,s}X^rY^s, \quad c_{r,s} \in \mathbb{C}, \quad c_{0,m} = 1.$$

I claim that $F_N(X, Y)$ is the unique polynomial of degree $\leq m$ in Y, with $c_{0,m} = 1$, such that

$$F_N(j, j_N) = 0.$$

In fact, $F_N(X, Y)$ generates the ideal in $\mathbb{C}[X, Y]$ of all polynomials $G(X, Y)$ such that $G(j, j_N) = 0$, from which the claim follows.

Proposition 24.3. *The polynomial $F_N(X, Y)$ has coefficients in $\mathbb{Q}$.*

Proof. We know that

$$j(z) = q^{-1} + \sum_{n=0}^{\infty} c(n)q^n, \quad c(n) \in \mathbb{Z}.$$

When we substitute this into the equation

$$F(j(z), j(Nz)) = 0,$$

and equate coefficients of powers of q, we obtain a set of linear equations for the $c_{r,s}$ with coefficients in $\mathbb{Q}$, and when we adjoin the equation

$$c_{0,m} = 1,$$

then the system determines the $c_{r,s}$ uniquely. Because the system of linear equations has a solution in $\mathbb{C}$, it also has a solution in $\mathbb{Q}$ (look at ranks of matrices); because the solution is unique, the solution in $\mathbb{C}$ must in fact lie in $\mathbb{Q}$. Therefore $c_{r,s} \in \mathbb{Q}$. □

The polynomial $F_N(X, Y)$ was introduced by Kronecker more than 100 years ago. It is known to be symmetric in X and Y. For $N = 2$, it is

$$X^3 + Y^3 - X^2Y^2 + 1488XY(X+Y) - 162000(X^2+Y^2)+$$

$$40773375XY + 8748000000(X+Y) - 157464000000000.$$

It was computed for $N = 3, 5, 7$ by Smith (1878), Berwick (1916), and Herrmann (1974). At this point the humans gave up, and left it to MACSYMA to compute F_{11} (1984). This last computation took about 20 hours on a VAX-780, and the result is a polynomial with coefficients up to 10^{60} that takes 5 pages to write out. It is important to know that the polynomial exists; fortunately, it is not important to know what it is.

The curve $X_0(N)$ over $\mathbb{Q}$. Let C_N be the affine curve over $\mathbb{Q}$ with equation $F_N(X, Y) = 0$, and let $\bar{C}_N$ be the projective curve defined by F_N made homogeneous. Then $z \mapsto (j(z), j(Nz))$ is a map $X_0(N) \setminus \Xi \to C_N(\mathbb{C})$, where Ξ is the set where j or j_N has a pole. This map extends uniquely to a map $X_0(N) \to \bar{C}_N(\mathbb{C})$, which is an isomorphism except over the singular points of $\bar{C}_N$, and the pair $(X_0(N), X_0(N) \to \bar{C}_N(\mathbb{C}))$ is uniquely determined by $\bar{C}_N$ (up to a unique isomorphism): it is the canonical "desingularization" of $\bar{C}_N$ over $\mathbb{C}$.

Now consider $\bar{C}_N$ over $\mathbb{Q}$. There is a canonical desingularization $X \to \bar{C}_N$ over $\mathbb{Q}$, i.e., a projective nonsingular curve X over $\mathbb{Q}$, and a regular map $X \to \bar{C}_N$ that is an isomorphism except over the singular points of $\bar{C}_N$, and the pair $(X, X \to \bar{C}_N)$ is uniquely determined by $\bar{C}_N$ (up to unique isomorphism). When we pass to the $\mathbb{C}$-points, we see that $(X(\mathbb{C}), X(\mathbb{C}) \to \bar{C}_N(\mathbb{C}))$ has the property characterizing $(X_0(N), X_0(N) \to \bar{C}_N(\mathbb{C}))$, and so there is a unique isomorphism of Riemann surfaces $X_0(N) \to X(\mathbb{C})$ compatible with the maps to $\bar{C}_N(\mathbb{C})$.

In summary, we have a well-defined curve X over $\mathbb{Q}$, a regular map $\gamma : X \to \bar{C}_N$ over $\mathbb{Q}$, and an isomorphism $X_0(N) \to X(\mathbb{C})$ whose composite with $\gamma(\mathbb{C})$ is (outside a finite set) $z \mapsto (j(z), j(Nz))$.

In future, we'll often use $X_0(N)$ to denote the curve X over $\mathbb{Q}$—it should be clear from the context whether we mean the curve over $\mathbb{Q}$ or the Riemann surface. The affine curve $X_0(N) \setminus \{\text{cusps}\} \subset X_0(N)$ is denoted $Y_0(N)$; thus $Y_0(N)(\mathbb{C}) = \Gamma_0(1)\backslash\mathbb{H}$.

Remark 24.4. It is known that the curve $F_N(X, Y) = 0$ is highly singular, because, in the absence of singularities, the formula on p9 would predict much too high a genus.

The points on the curve $X_0(N)$. Since we can't write down an equation for $X_0(N)$ as a projective curve over $\mathbb{Q}$, we would at least like to know what its points are in any field containing $\mathbb{Q}$. This we can do.

We first look at the complex points of $X_0(N)$, i.e., at the Riemann surface $X_0(N)$. Consider the diagram:

$$\begin{tikzcd}
\{(E,S)\}/\approx \arrow[r, leftrightarrow] \arrow[d] & \{(\Lambda,S)\}/\mathbb{C}^\times \arrow[r, leftrightarrow] \arrow[d] & \Gamma_0(N)\backslash M/\mathbb{C}^\times \arrow[r, leftrightarrow] \arrow[d] & \Gamma_0(N)\backslash\mathbb{H} \arrow[d] \\
\{E\}/\approx \arrow[r, leftrightarrow] & \mathcal{L}/\mathbb{C}^\times \arrow[r, leftrightarrow] & \Gamma_0(1)\backslash M/\mathbb{C}^\times \arrow[r, leftrightarrow] & \Gamma_0(1)\backslash\mathbb{H}
\end{tikzcd}$$

The bottom row combines maps in Section 10. All the symbols $\leftrightarrow$ are natural bijections.

Recall that M is the subset of $\mathbb{C}\times\mathbb{C}$ of pairs (ω_1,ω_2) such that $\Im(\omega_1/\omega_2)>0$ (so $M/\mathbb{C}^\times\subset\mathbb{P}^1(\mathbb{C})$), and that the bijection $M/\mathbb{C}\to\mathbb{H}$ sends (ω_1,ω_2) to ω_1/ω_2. The rest of the right hand square is now obvious.

Recall that the $\mathcal{L}$ is the set of lattices in $\mathbb{C}$, and that the lattices defined by two pairs in M are equal if and only if the pairs lie in the same $\Gamma_0(1)$-orbit. Thus in passing from an element of M to its $\Gamma_0(1)$-orbit we are forgetting the basis and remembering only the lattice. In passing from an element of M to its $\Gamma_0(N)$-orbit, we remember a little of the basis, for suppose

$$\begin{pmatrix}\omega_1'\\ \omega_2'\end{pmatrix}=\begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}\omega_1\\ \omega_2\end{pmatrix},\quad \begin{pmatrix}a & b\\ c & d\end{pmatrix}\in\Gamma_0(N).$$

Then

$$\begin{aligned}\omega_1' &= a\omega_1+b\omega_2\\ \omega_2' &= c\omega_1+d\omega_2\equiv d\omega_2 \mod N\Lambda.\end{aligned}$$

Hence

$$\frac{1}{N}\omega_2'\equiv\frac{d}{N}\omega_2 \quad \mod \Lambda.$$

Note that because $\begin{pmatrix}a & b\\ c & d\end{pmatrix}$ has determinant 1, $\gcd(d,N)=1$, and so $\frac{1}{N}\omega_2'$ and $\frac{1}{N}\omega_2$ generate the same cyclic subgroup S of order N in $\mathbb{C}/\Lambda$. We see that the map

$$(\omega_1,\omega_2)\mapsto(\Lambda(\omega_1,\omega_2),<\frac{1}{N}\omega_2>)$$

defines a bijection from $\Gamma_0(N)\backslash M$ to the set of pairs consisting of a lattice Λ in $\mathbb{C}$ and a cyclic subgroup S of $\mathbb{C}/\Lambda$ of order N. Now $(\Lambda,S)\mapsto(\mathbb{C}/\Lambda,S)$ defines a one-to-one correspondence between this last set and the set of isomorphism classes of pairs (E,S) consisting of an elliptic curve over $\mathbb{C}$ and a cyclic subgroup S of $E(\mathbb{C})$ of order N. An isomorphism $(E,S)\to(E',S')$ is an isomorphism $E\to E'$ carrying S into S'.

Note that $E/S=\mathbb{C}/\Lambda(\omega_1,\frac{1}{N}\omega_2)\leftrightarrow N\frac{\omega_1}{\omega_2}$, and so, if $j(E)=j(z)$, then $j(E/S)=j(Nz)$.

Now, for any field $k\supset\mathbb{Q}$, define $\mathcal{E}_0(N)(k)$ to be the set of isomorphism classes of pairs E consisting of an elliptic curve E over k and a cyclic subgroup $S\subset E(k^{\mathrm{al}})$ of order N stable under $\mathrm{Gal}(k^{\mathrm{al}}/k)$—thus the subgroup S is defined over k, but not necessarily its elements. The above remarks show that there is a canonical bijection

$$\mathcal{E}_0(N)(\mathbb{C})/\approx\to Y_0(N)$$

whose composite with the map $Y_0(N)\to C_N(\mathbb{C})$ is $(E,S)\mapsto(j(E),j(E/S))$. Here $Y_0(N)$ denotes the Riemann surface $\Gamma_0(N)\backslash\mathbb{H}$.

Theorem 24.5. *For any field $k \supset \mathbb{Q}$, there is a map*

$$\mathcal{E}_0(N)(k) \to Y_0(N)(k),$$

functorial in k, such that

(a) *the composite $\mathcal{E}_0(N)(k) \to Y_0(N)(k) \to C_N(k)$ is $(E, S) \mapsto (j(E), j(E/S))$;*
(b) *for all k, $\mathcal{E}_0(N)(k)/\approx \to Y_0(N)(k)$ is surjective, and for all algebraically closed k it is bijective.*

The map being functorial in k means that for every homomorphism $\sigma : k \to k'$ of fields, the diagram

$$\begin{tikzcd}
\mathcal{E}_0(N)(k') \arrow[r] & Y_0(N)(k') \\
\mathcal{E}_0(N)(k) \arrow[r] \arrow[u, "\sigma"] & Y_0(N)(k) \arrow[u, "\sigma"]
\end{tikzcd}$$

commutes. In particular, $\mathcal{E}_0(N)(k^{\mathrm{al}}) \to Y_0(N)(k^{\mathrm{al}})$ commutes with the actions of $\mathrm{Gal}(k^{\mathrm{al}}/k)$. Since $Y_0(N)(k^{\mathrm{al}})^{\mathrm{Gal}(k^{\mathrm{al}}/k)} = Y_0(N)(k)$, this implies that

$$\boxed{Y_0(N)(k) = (\mathcal{E}_0(N)(k^{\mathrm{al}})/\approx)^{\mathrm{Gal}(k^{\mathrm{al}}/k)}}$$

for any field $k \supset \mathbb{Q}$.

This description of the points can be extended to $X_0(N)$ by adding to $\mathcal{E}_0(N)$ certain "degenerate" elliptic curves.

Variants. For our applications to elliptic curves, we shall only need to use the quotients of $\mathbb{H}^*$ by the subgroups $\Gamma_0(N)$, but quotients by other subgroups are also of interest. For example, let

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \middle| \, a \equiv 1 \equiv d \mod N, \quad c \equiv 0 \mod N \right\}.$$

The quotient $X_1(N) = \Gamma_1(N)\backslash\mathbb{H}^*$ again defines a curve, also denoted $X_1(N)$, over $\mathbb{Q}$, and there is a theorem similar to (24.5) but with $\mathcal{E}_1(N)(k)$ the set of pairs (E, P) consisting of an elliptic curve E over k and a point $P \in E(k)$ of order N.

In this case, the map

$$\mathcal{E}_1(N)(k)/\approx \to Y_1(N)(k)$$

is a bijection whenever $4|N$. The curve $X_1(N)$ has genus 0 exactly for $N = 1, 2, \ldots, 10, 12$. Since $X_1(N)$ has a point with coordinates in $\mathbb{Q}$ for each of these N (there does exist an elliptic curve over $\mathbb{Q}$ with a point of that order), $X_1(N) \approx \mathbb{P}^1$, and so $X_1(N)$ has infinitely many rational points. Therefore, for $N = 1, 2, \ldots, 10, 12$, there are infinitely many elliptic curves over $\mathbb{Q}$ with a point of order N (rational over $\mathbb{Q}$). Mazur showed, that for all other N, $Y_0(N)$ is empty, and so these are the only possible orders for a point on an elliptic curve over $\mathbb{Q}$ (Conjecture of Beppo Levi).

25. Modular Forms

It is difficult to construct functions on $\mathbb{H}$ invariant under a subgroup Γ of $\mathrm{SL}_2(\mathbb{Z})$ of finite index. One strategy is to construct functions, not invariant under Γ, but transforming in a certain fixed manner. Two functions transforming in the same manner will be invariant under Γ. This idea suggests the notion of a modular form.

Definition of a modular form.

Definition 25.1. Let Γ be a subgroup of finite index in $\mathrm{SL}_2(\mathbb{Z})$. A *modular form* for Γ of weight[27] $2k$ is a function $f:\mathbb{H}\to\mathbb{C}$ such that

(a) f is holomorphic on $\mathbb{H}$;
(b) for any $\gamma=\begin{pmatrix} a & b\\ c & d\end{pmatrix}\in\Gamma$, $f(\gamma z)=(cz+d)^{2k}f(z)$;
(c) f is holomorphic at the cusps.

Recall that the cusps are the points in $\mathbb{H}^*$ not in $\mathbb{H}$. Since Γ is of finite index in $\mathrm{SL}_2(\mathbb{Z})$, $T^h=\begin{pmatrix} 1 & h\\ 0 & 1\end{pmatrix}$ is in Γ for some integer $h>0$, which we may take to be as small as possible. Then condition (b) implies that $f(T^hz)=f(z)$, i.e., that $f(z+h)=f(z)$, and so

$$f(z)=f^*(q),\quad q=e^{2\pi iz/h},$$

and f^* is a function on a neighbourhood of $0\in\mathbb{C}$, with 0 removed. To say that f is holomorphic at ∞ means that f^* is holomorphic at 0, and so

$$f(z)=\sum_{n\geq 0}c(n)q^n,\quad q=e^{2\pi iz/h}.$$

For a cusp $r\neq\infty$, choose a $\gamma\in\mathrm{SL}_2(\mathbb{Z})$ such that $\gamma(\infty)=r$, and then the requirement is that $f\circ\gamma$ be holomorphic at ∞. It suffices to check the condition for only one cusp in each Γ-orbit.

A modular form is called a *cusp form* if it is zero at the cusps. For example, for the cusp ∞ this means that

$$f(z)=\sum_{n\geq 1}c(n)q^n,\quad q=e^{2\pi iz/h}.$$

Remark 25.2. Note that, for $\gamma=\begin{pmatrix} a & b\\ c & d\end{pmatrix}\in\mathrm{SL}_2(\mathbb{Z})$,

$$d\gamma z=d\frac{az+b}{cz+d}=\frac{a(cz+d)-c(az+b)}{(cz+d)^2}dz=(cz+d)^{-2}dz.$$

Thus condition (25.1b) says that $f(z)(dz)^k$ is invariant under the action of Γ.

Write $\mathcal{M}_{2k}(\Gamma)$ for the vector space of modular forms of weight $2k$, and $\mathcal{S}_{2k}(\Gamma)$ for the subspace[28] of cusp forms. A modular form of weight 0 is a holomorphic modular function (i.e., a holomorphic function on the compact Riemann surface $X(\Gamma)$), and is therefore constant: $\mathcal{M}_0(\Gamma)=\mathbb{C}$. The product of modular forms of weight $2k$ and $2k'$ is a modular form of weight $2(k+k')$, which is a cusp form if one of the two forms is a cusp form. Therefore $\oplus_{k\geq 0}\mathcal{M}_{2k}(\Gamma)$ is a graded $\mathbb{C}$-algebra.

Proposition 25.3. *Let π be the quotient map $\mathbb{H}^*\to\Gamma_0(N)\backslash\mathbb{H}^*$, and for any holomorphic differential ω on $\Gamma_0(N)\backslash\mathbb{H}^*$, set $\pi^*\omega=fdz$. Then $\omega\mapsto f$ is an isomorphism from the space of holomorphic differentials on $\Gamma_0(N)\backslash\mathbb{H}^*$ to $\mathcal{S}_2(\Gamma_0(N))$.*

[27] k and $-k$ are also used.

[28] The $\mathcal{S}$ is for "Spitzenform", the German name for cusp form. The French call them "forme parabolique".

Proof. The only surprise is that f is necessarily a cusp form rather than just a modular form. I explain what happens at ∞. Recall (p122) that there is a neighbourhood U of ∞ in $\Gamma_0(N)\backslash\mathbb{H}^*$ and an isomorphism $q : U \to D$ (some disk) such that $q \circ \pi = e^{2\pi iz}$. Consider the differential $g(q)dq$ on U. Its inverse image on $\mathbb{H}$ is

$$g(e^{2\pi iz})d(e^{2\pi iz}) = 2\pi i \cdot g(e^{2\pi iz}) \cdot e^{2\pi iz}dz = 2\pi i f dz$$

where $f(z) = g(e^{2\pi iz}) \cdot e^{2\pi iz}$. If g is holomorphic at 0, then $g(q) = \sum_{n\geq 0} c(n)q^n$, and so the q-expansion of f is $q\sum_{n\geq 0} c(n)q^n$, which is zero at ∞. □

Corollary 25.4. *The $\mathbb{C}$-vector space $\mathcal{S}_2(\Gamma_0(N))$ has dimension equal to the genus of $X_0(N)$.*

Proof. It is part of the theory surrounding the Riemann-Roch theorem that the holomorphic differential forms on a compact Riemann surface form a vector space equal to the genus of the surface. □

Hence, there are explicit formulas for the dimension of $\mathcal{S}_2(\Gamma_0(N))$—see p123. For example, it is zero for $N \leq 10$, and has dimension 1 for $N = 11$. In fact, the Riemann-Roch theorem gives formulas for the dimension of $\mathcal{S}_{2k}(\Gamma_0(N))$ for all N.

The modular forms for $\Gamma_0(1)$. In this section, we find the $\mathbb{C}$-algebra $\oplus_{k\geq 0}\mathcal{M}_{2k}(\Gamma_0(1))$.

We first explain a method of constructing functions satisfying (25.1b). As before, let $\mathcal{L}$ be the set of lattices in $\mathbb{C}$, and let $F : \mathcal{L} \to \mathbb{C}$ be a function such that

$$F(\lambda\Lambda) = \lambda^{-2k}F(\Lambda), \quad \lambda \in \mathbb{C}, \quad \Lambda \in \mathcal{L}.$$

Then

$$\omega_2^{2k}F(\Lambda(\omega_1, \omega_2))$$

depends only on the ratio $\omega_1 : \omega_2$, and so there is a function $f(z)$ defined on $\mathbb{H}$ such that

$$\omega_2^{2k}F(\Lambda(\omega_1, \omega_2)) = f(\omega_1/\omega_2), \text{ whenever } \Im(\omega_1/\omega_2) > 0.$$

For $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z})$, $\Lambda(a\omega_1 + b\omega_2, c\omega_1 + d\omega_2) = \Lambda(\omega_1, \omega_2)$ and so

$$f(\frac{az+b}{cz+d}) = (cz+d)^{-2k}F(\Lambda(z,1)) = (cz+d)^{-2k}f(z).$$

When we apply this remark to the Eisenstein series

$$G_{2k}(\Lambda) = \sum_{\omega\in\Lambda,\omega\neq 0} \frac{1}{\omega^{2k}},$$

we find that the function $G_{2k}(z) =_{df} G_{2k}(\Lambda(z,1))$ satisfies (25.1b). In fact:

Proposition 25.5. *For all $k > 1$, $G_{2k}(z)$ is a modular form of weight $2k$ for $\Gamma_0(1)$, and Δ is a cusp form of weight 12.*

Proof. We know that $G_{2k}(z)$ is holomorphic on $\mathbb{H}$, and the formula on p125 shows that it is holomorphic at ∞, which is the only cusp for $\Gamma_0(1)$ (up to $\Gamma_0(1)$-equivalence). The statement for Δ is obvious from its definition $\Delta = g_4(z)^3 - 27g_4(z)^2$, and its q-expansion (p125). □

Theorem 25.6. *The $\mathbb{C}$-algebra $\oplus_{k\geq 0}\mathcal{M}_{2k}(\Gamma_0(1))$ is generated by G_4 and G_6, and G_4 and G_6 are algebraically independent over $\mathbb{C}$. Therefore*

$$\mathbb{C}[G_4, G_6] \xrightarrow{\approx} \oplus_{k\geq 0}\mathcal{M}_{2k}(\Gamma_0(1)), \quad \mathbb{C}[G_4, G_6] \approx \mathbb{C}[X, Y]$$

(isomorphisms of graded $\mathbb{C}$-algebras if X and Y are given weights 4 and 6 respectively). Moreover,

$$f \mapsto f \cdot \Delta : \mathcal{M}_{2k-12}(\Gamma_0(1)) \to \mathcal{S}_{2k}(\Gamma_0(1))$$

is a bijection.

Proof. Straightforward—see Serre, Cours..., VII.3.2. □

Therefore, for $k \geq 0$,

$$\dim \mathcal{M}_{2k}(\Gamma_0(N)) = \begin{cases} [k/6] & \text{if } k \equiv 1 \mod 6 \\ [k/6] + 1 & \text{otherwise.} \end{cases}$$

Here $[x]$ is the largest integer $\leq x$.

Theorem 25.7 (Jacobi). *There is the following formula:*

$$\Delta = (2\pi)^{12} q \prod_{n=1}^{\infty} (1 - q^n)^{24}, \quad q = e^{2\pi i z}.$$

Proof. Let

$$F(z) = q \prod_{n=1}^{\infty} (1 - q^n)^{24}.$$

From the theorem, we know that the space of cusp forms of weight 12 has dimension 1, and therefore if we can show that $F(z)$ is such a form, then we'll know it is a multiple of Δ, and it will be follow from the formula on p125 that the multiple is $(2\pi)^{12}$.

Because $\mathrm{SL}_2(\mathbb{Z})/\{\pm I\}$ is generated by $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, to verify the conditions in (25.1), it suffices to verify that F transforms correctly under T and S. For T this is obvious from the way we have defined F, and for S it amounts to checking that

$$F(-\frac{1}{z}) = z^{12} F(z).$$

This is trickier than it looks, but there are short (2 page) elementary proofs—see for example, Serre, ibid., VII.4.4. □

26. Modular Forms and the L-series of Elliptic Curves

In this section, I'll discuss how the L-series classify the elliptic curves over $\mathbb{Q}$ up to isogeny, and then I'll explain how the work of Hecke, Petersson, and Atkin-Lehner leads to a list of candidates for the L-series of such curves, and hence suggests a classification of the isogeny classes.

Dirichlet Series. A *Dirichlet series* is a series of the form

$$f(s)=\sum_{n\geq 1} a(n)n^{-s}, \quad a(n)\in\mathbb{C}, \quad s\in\mathbb{C}.$$

The simplest example of such a series is, of course, the Riemann zeta function $\sum_{n\geq 1} n^{-s}$. If there exist positive constants A and b such that $|\sum_{n\leq x} a(n)|\leq Ax^b$ for all large x, then the series for $f(s)$ converges to an analytic function on the half-plane $\Re(s)>b$.

It is important to note that the function $f(s)$ determines the $a(n)$'s, i.e., if $\sum a(n)n^{-s}$ and $\sum b(n)n^{-s}$ are equal as functions of s on some half-plane, then $a(n)=b(n)$ for all n. In fact, by means of the Mellin transform and its inverse (see 26.4 below), f determines, and is determined by, a function $g(q)$ convergent on some disk about 0, and $g(q)=\sum a(n)q^n$.

We shall be especially interested in Dirichlet series that are equal to Euler products, i.e., those that can be expressed as

$$f(s)=\prod_p \frac{1}{1-P_p(p^{-s})}$$

where each P_p is a polynomial.

Dirichlet series arise in two essentially different ways: from analysis and from geometry and number theory. One of *big* problems mathematics is to show that the second set of Dirichlet series is a subset of the first, and to identify the subset. This is a major theme in Langlands's philosophy, and the rest of the course will be concerned with explaining how Wiles was able to identify the L-series of (almost all) elliptic curves over $\mathbb{Q}$ with certain L-series attached to modular forms.

The L-series of an elliptic curve. Recall that for an elliptic curve E over $\mathbb{Q}$, we define

$$L(E,s)=\prod_{p \text{ good}} \frac{1}{1-a_p p^{-s}+p^{1-s}}\cdot\prod_{p \text{ bad}}\frac{1}{1-a_p p^{-s}}$$

where

$$a_p=\begin{cases} p+1-N_p & p \text{ good;}\\ 1 & p \text{ split nodal;}\\ -1 & p \text{ nonsplit nodal;}\\ 0 & p \text{ cuspidal.}\end{cases}$$

Recall also that the conductor $N=N_{E/\mathbb{Q}}$ of $\mathbb{Q}$ is $\prod_p p^{f_p}$ where $f_p=0$ if E has good reduction at p, $f_p=1$ if E has nodal reduction at p, and $f_p\geq 2$ otherwise (and $=2$ unless $p=2,3$).

On expanding out the product (cf. below), we obtain a Dirichlet series

$$L(E,s)=\sum a_n n^{-s}.$$

This series has, among others, the following properties:

(a) (Rationality) Its coefficients a_n lie in $\mathbb{Q}$.
(b) (Euler product) It can be expressed as an "Euler product"; in fact, that's how it is defined.
(c) (Functional equation) *Conjecturally* it can be extended analytically to a meromorphic function on the whole complex plane that satisfies the functional equation

$$\Lambda(E,s)=w\Lambda(E,2-s), \quad w=\pm 1,$$

where $\Lambda(E,s)=N_{E/\mathbb{Q}}^{s/2}(2\pi)^{-s}\Gamma(s)L(E,s)$.

L-series and isogeny classes. Recall that two elliptic curves E and E' are said to be *isogenous* if there is a nonconstant regular map from one to the other. By composing the map with a translation, we will then get a map sending 0 to 0, in which case it will also be a homomorphism for the group structures on E and E'. A nonconstant map $\varphi : E \to E'$ such that $\varphi(0) = 0$ is called an *isogeny.*

Lemma 26.1. *Isogeny is an equivalence relation.*

Proof. The identity map is an isogeny, so it is reflexive, and the composite of two isogenies is an isogeny, so it is transitive. Let $\varphi : E \to E'$ be an isogeny, and let S be its kernel. Since S is finite, it will be contained in E_n for some n, and there are isogenies

$$\begin{tikzcd} E \arrow[r] \arrow[d, equal] & E/S \arrow[r] \arrow[d, equal] & E/E_n \arrow[d, equal] \\ E \arrow[r] & E' \arrow[r] & E \end{tikzcd}$$

—the isomorphism $E \to E/E_n$ is induced by multiplication by n in

$$0 \to E_n \to E \xrightarrow{n} E \to 0.$$

Here I'm assuming facts about elliptic curves and their quotients by finite subgroups ([S1] III.4). □

An isogeny $E \to E'$ induces a homomorphism $E(\mathbb{Q}) \to E'(\mathbb{Q})$ which, in general, will be neither injective nor surjective. The ranks of $E(\mathbb{Q})$ and $E'(\mathbb{Q})$ will be the same, but their torsion subgroups will in general be different. Surprisingly, isogenous curves over a finite field do have the same number of points.

Theorem 26.2. *Let E and E' be elliptic curves over $\mathbb{Q}$. If E and E' are isogenous, then $N_p(E) = N_p(E')$ for all good p. Conversely, if $N_p(E) = N_p(E')$ for sufficiently many good p, then E is isogenous to E'.*

Proof. The fact that allows us to show that $N_p(E) = N_p(E')$ when E and E' are isogenous is that $N_p(E)$ is the degree of a map $E \to E$, in fact, it is the degree of $\varphi - 1$ where φ is the Frobenius map (see p101). An isogeny $\alpha : E \to E'$ induces and isogeny $\alpha_p : E_p \to E'_p$ on the reductions of the curves modulo p, which commutes with the Frobenius map: if $\alpha(x : y : z) = (P(x, y, z) : Q(x : y : z), R(x : y : z)$, $P, Q, R \in \mathbb{F}_p[X, Y, Z]$, then

$$\alpha\varphi(x : y : z) = (P(x^p, y^p, z^p), \dots)$$

whereas

$$\varphi\alpha(x : y : z) = (P(x, y, z)^p, \dots),$$

which the characteristic p binomial theorem shows to be equal. Because the diagram

$$\begin{tikzcd} E \arrow[r, "\varphi-1"] \arrow[d, "\alpha"] & E \arrow[d, "\alpha"] \\ E' \arrow[r, "\varphi-1"] & E' \end{tikzcd}$$

commutes, we see that

$$\deg \alpha \cdot \deg(\varphi - 1) = \deg(\varphi - 1) \cdot \deg \alpha,$$

so,

$$\deg \alpha \cdot N_p(E) = N_p(E') \cdot \deg \alpha,$$

and we can cancel $\deg\alpha$.

The converse is much more difficult. It was conjectured by Tate about 1963, and proved under various hypotheses by Serre. It was proved in general by Faltings in his paper on Mordell's conjecture (1983). □

Faltings's result gives an effective procedure for deciding whether two elliptic curves over $\mathbb{Q}$ are isogenous: there is a constant P such that if $N_p(E) = N_p(E')$ for all good $p \leq P$, then E and E' are isogenous. Unfortunately, P is impossibly large, but, in practice, if your computer fails to find a p with $N_p(E) \neq N_p(E')$ in a few minutes you can be very confident that the curves are isogenous.

It is not quite obvious, but it follows from the theory of Néron models, that isogenous elliptic curves have the same type of reduction at every prime. Therefore, isogenous curves have exactly the same L-series and the same conductor. Because the L-series is determined by, and determines the N_p, we have the following corollary.

Corollary 26.3. *Two elliptic curves E and E' are isogenous if and if $L(E,s) = L(E',s)$.*

We therefore have a one-to-one correspondence between

$$\{\text{isogeny classes of elliptic curves over } \mathbb{Q}\} \leftrightarrow \{\text{certain } L\text{-series}\}$$

In the remainder of this section we shall identify (using only complex analysis) the L-series arising from elliptic curves over $\mathbb{Q}$ (in fact, we'll identify the L-series of the elliptic curves with a fixed conductor).

Since we shall be classifying elliptic curves only up to isogeny, it is worth noting that a theorem of Shafarevich implies there are only finitely many isomorphism classes of elliptic curves over $\mathbb{Q}$ with a given conductor, hence only finitely many in each isogeny class—see [S1], IX.6.

The L-series of a modular form. Let f be a modular form of weight $2k$ for $\Gamma_0(N)$. By definition, it is invariant under $z \mapsto z+1$ and is zero at the cusp ∞, and so can be expressed

$$f(s) = \sum_{n\geq 1} c(n)q^n, \quad q = e^{2\pi iz}, \quad c(n) \in \mathbb{C}.$$

The *L-series* of f is the Dirichlet series

$$L(f,s) = \sum c(n)n^{-s}, \quad s \in \mathbb{C}.$$

A rather rough estimate shows that $|c(n)| \leq Cn^k$ for some constant C, and so this Dirichlet series is convergent for $\Re(s) > k+1$.

Remark 26.4. Let f be cusp form. The *Mellin transform* of f (more accurately, of the function $y \mapsto f(iy) : \mathbb{R}_{>0} \to \mathbb{C}$) is defined to be

$$g(s) = \int_0^\infty f(iy)y^s \frac{dy}{y}.$$

Ignoring (as usual) questions of convergence, we find that

$$\begin{aligned} g(s) &= \int_0^\infty \sum_{n=1}^\infty c_n e^{-2\pi n y} y^s \frac{dy}{y} \\ &= \sum_{n=1}^\infty c_n \int_0^\infty e^{-t}(2\pi n)^{-s} t^s \frac{dt}{t} \quad (t = 2\pi n y) \\ &= (2\pi)^{-s}\Gamma(s) \sum_{n=1}^\infty c(n) n^{-s} \\ &= (2\pi)^{-s}\Gamma(s) L(f,s). \end{aligned}$$

For the experts, the Mellin transform is the version of the Fourier transform appropriate for the multiplicative group $\mathbb{R}_{>0}$.

Modular forms whose L-series have a functional equations. Let $\alpha_N = \begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$. Then

$$\alpha_N \begin{pmatrix} a & b \\ c & d \end{pmatrix} \alpha_N^{-1} = \begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & 1/N \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} d & -c/N \\ -Nb & a \end{pmatrix},$$

and so conjugation by α_N preserves $\Gamma_0(N)$. Define

$$(w_N f)(z) = (\sqrt{N} z)^{2k} f(-1/z).$$

Then w_N preserves $\mathcal{S}_{2k}(\Gamma_0(N))$ and has order 2, $w_N^2 = 1$. Therefore the eigenvalues of w_N are ± 1 (or perhaps just $+1$), and $\mathcal{S}_{2k}(\Gamma_0(N))$ is a direct sum of the corresponding eigenspaces $\mathcal{S}_{2k} = \mathcal{S}_{2k}^{+1} \oplus \mathcal{S}_{2k}^{-1}$.

Theorem 26.5 (Hecke). *Let $f \in \mathcal{S}_{2k}(\Gamma_0(N))$ be a cusp form in the ε-eigenspace, $\varepsilon = 1$ or -1. Then f extends analytically to a holomorphic function on the whole complex plane, and satisfies the functional equation*

$$\Lambda(f,s) = \varepsilon(-1)^k \Lambda(f, k-s),$$

where

$$\Lambda(f,s) = N^{s/2}(2\pi)^{-s}\Gamma(s)L(f,s).$$

Proof. We omit the proof—it involves only fairly straightforward analysis (see Knapp, p270). □

Thus we see that, for $k = 2$, $L(f,s)$ has exactly the functional equation we hope for the $L(E,s)$.

Modular forms whose L-functions are Euler products. Write

$$q \prod_1^\infty (1-q^n)^{24} = \sum \tau(n) q^n.$$

The function $n \mapsto \tau(n)$ is called the *Ramanujan τ-function.* Ramanujan conjectured that it had the following properties:

(a) $|\tau(p)| \leq 2p^{11/2}$;

(b) $\begin{cases} \tau(mn) = \tau(m)\tau(n) & \text{if } \gcd(m,n) = 1; \\ \tau(p)\cdot\tau(p^n) = \tau(p^{n+1}) + p^{11}\tau(p^{n-1}) & \text{if } p \text{ is prime and } n \geq 1. \end{cases}$

Conjecture (a) was proved by Deligne: he first showed that $\tau(p) = \alpha + \beta$ where α and β occur as the reciprocal roots of a "$P_{11}(T)$" (see p103), and so (a) became a consequence of his proof of the Riemann hypothesis.

Conjecture (b) was proved by Mordell in 1917 in a paper in which he introduced the first examples of Hecke operators. Consider a modular form f of weight $2k$ for $\Gamma_0(N)$ (e.g., $\Delta = (2\pi)^{12}q\prod(1-q^n)^{24}$, which is a modular form of weight 12 for $\Gamma_0(1)$), and write

$$L(f,s) = \sum_{n\geq 0} c(n)n^{-s}.$$

Proposition 26.6. *The Dirichlet series $L(f,s)$ has an Euler product expansion of the form*

$$L(f,s) = \prod_{p|N} \frac{1}{1-c(p)p^{-s}} \prod_{\gcd(p,N)=1} \frac{1}{1-c(p)p^{-s}+p^{2k-1-s}}$$

if (and only if)

$$(*)\left\{\begin{array}{rcl} c(mn) & = & c(m)c(n) \quad \text{if } \gcd(m,n)=1; \\ c(p)\cdot c(p^r) & = & c(p^{r+1}) + p^{2k-1}c(p^{r-1}),\ r\geq 1,\ \text{if } p \text{ does not divide } N; \\ c(p^r) & = & c(p)^r,\ r\geq 1, \quad \text{if } p|N. \end{array}\right.$$

Proof. For a prime p not dividing N, define

$$L_p(s) = \sum c(p^m)p^{-ms} = 1 + c(p)p^{-s} + c(p^2)(p^{-s})^2 + \cdots.$$

By inspection, the coefficient of $(p^{-s})^r$ in the product

$$(1-c(p)p^{-s}+p^{2k-1}p^{-s})L_p(s)$$

is

$$\begin{array}{cl} 1 & \text{for} \quad r=0 \\ 0 & \text{for} \quad r=1 \\ & \quad\cdots \\ c(p^{r+1}) - c(p)c(p^r) + p^{2k-1}c(p^{r-1}) & \text{for} \quad r+1. \end{array}$$

Therefore

$$L_p(s) = \frac{1}{1-c(p)p^{-s}+p^{2k-1-s}}$$

if and only if the second equation in (*) holds.

Similarly,

$$L_p(s) =_{df} \sum c(p^r)p^{-rs} = \frac{1}{1-c(p)p^{-s}}$$

if and only if the third equation in (*) holds.

If $n\in\mathbb{N}$ factors as $n=\prod p_i^{r_i}$, then the coefficient of $(p^{-s})^n$ in $\prod L_p(s)$ is $\prod c(p_i^{r_i})$, which equals $c(n)$ if (*) holds. □

Remark 26.7. The proposition says that $L(f,s)$ is equal to an Euler product of the above form if and only if $n\mapsto c(n)$ is weakly multiplicative and if the $c(p^m)$ satisfy a suitable recurrence relation. Note that $(*)$, together with the normalization $c(1)=1$, shows that the $c(n)$ are determined by the $c(p)$ for p prime.

Hecke defined linear maps

$$T(n): \mathcal{S}_{2k}(\Gamma_0(N)) \to \mathcal{S}_{2k}(\Gamma_0(N)), \quad n \geq 1,$$

and proved the following theorems.

Theorem 26.8. *The maps $T(n)$ have the having the following properties:*

(a) $T(mn) = T(m)T(n) \quad \text{if } \gcd(m,n) = 1$*;*
(b) $T(p) \cdot T(p^r) = T(p^{r+1}) + p^{2k-1}T(p^{r-1})$ *if p doesn't divide N;*
(c) $T(p^r) = T(p)^r$, $r \geq 1$, $p|N$;
(d) *all $T(n)$ commute.*

Theorem 26.9. *Let f be a cusp form of weight $2k$ for $\Gamma_0(N)$ that is simultaneously an eigenvector for all $T(n)$, say $T(n)f = \lambda(n)f$, and let*

$$f(z) = \sum_{n=1}^{\infty} c(n)q^n, \quad q = e^{2\pi iz}.$$

Then

$$c(n) = \lambda(n)c(1).$$

Note that $c(1) \neq 0$, because otherwise $c(n) = 0$ for all n, and so $f = 0$.

Corollary 26.10. *Let f be as in Theorem 26.9, and normalize f so that $c(1) = 1$. Then*

$$L(f,s) = \prod_{p|N} \frac{1}{1 - c(p)p^{-s}} \prod_{\gcd(p,N)=1} \frac{1}{1 - c(p)p^{-s} + p^{2k-1-s}}$$

Example 26.11. Since $\mathcal{S}_{12}(\Gamma_0(1))$ has dimension 1, Δ must be an eigenform for all $T(n)$, which implies (b) of Ramanujan's conjecture.

Definition of the Hecke operators. I first explain the definition of the Hecke operators for the full group $\Gamma_0(1) = \mathrm{SL}_2(\mathbb{Z})$.

Recall that we have canonical bijections

$$\mathcal{L}/\mathbb{C}^\times \leftrightarrow \Gamma_0(1)\backslash M/\mathbb{C}^\times \leftrightarrow \Gamma_0(1)\backslash\mathbb{H}.$$

Moreover, the equation

$$f(z) = F(\Lambda(z,1))$$

defines a one-to-one correspondence between

(a) functions $F: \mathcal{L} \to \mathbb{C}$ such that $F(\lambda\Lambda) = \lambda^{-2k}F(\Lambda)$, $\lambda \in \mathbb{C}^\times$;
(b) functions $f: \mathbb{H} \to \mathbb{C}$ such that $f(\gamma z) = (cz+d)^{2k}f(z)$, $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

We'll work first with $\mathcal{L}$.

Let $\mathcal{D}$ be the free abelian group generated by the $\Lambda \in \mathcal{L}$; thus an element of $\mathcal{D}$ is a finite sum

$$\sum n_\Lambda[\Lambda], \quad n_\Lambda \in \mathbb{Z}, \quad \Lambda \in \mathcal{L},$$

and two such sums $\sum n_\Lambda[\Lambda]$ and $\sum n'_\Lambda[\Lambda]$ are equal if and only if $n_\Lambda = n'_\Lambda$ for all Λ.

For $n \geq 1$, define

$$T(n): \mathcal{D} \to \mathcal{D}, \quad T(n) = \sum_{(\Lambda:\Lambda')=n} [\Lambda']$$

and

$$R(n):\mathcal{D}\to\mathcal{D},\quad R(n)=[n\Lambda].$$

Proposition 26.12. *(a)* $T(mn)=T(m)\circ T(n)$ *if* $\gcd(m,n)=1$*;*
(b) $T(p^r)\circ T(p)=T(p^{r+1})+pR(p)\circ T(p^{r-1})$*.*

Proof. (a) For a lattice Λ,

$$\begin{aligned} T(mn)[\Lambda] &= \textstyle\sum[\Lambda''] \text{ sum over } \Lambda'', (\Lambda:\Lambda'')=mn),\\ T(m)\circ T(n)[\Lambda] &= \textstyle\sum[\Lambda''] \text{ (sum over pairs } (\Lambda',\Lambda'') \text{ with } (\Lambda:\Lambda')=n,\ (\Lambda':\Lambda'')=m).\end{aligned}$$

But if Λ'' is a lattice of index mn, then Λ/Λ'' is an abelian group of order mn with $\gcd(m,n)=1$, and so has a unique subgroup of order m. The inverse image of this subgroup in Λ will be the unique lattice $\Lambda'\supset\Lambda''$ such that $(\Lambda':\Lambda'')=m$. Thus the two sums are the same.

(b) For a lattice Λ,

$$\begin{aligned} T(p^r)\circ T(p)[\Lambda] &= \textstyle\sum[\Lambda''] \text{ (sum over pairs } (\Lambda',\Lambda'') \text{ with } (\Lambda:\Lambda')=p,\ (\Lambda':\Lambda'')=p^r),\\ T(p^{r+1})[\Lambda] &= \textstyle\sum[\Lambda''] \text{ (sum of } \Lambda'' \text{ with } (\Lambda:\Lambda'')=p^{r+1});\\ pR(p)\circ T(p^{n-1})[\Lambda] &= p\cdot\textstyle\sum R(p)[\Lambda'] \text{ (sum over } \Lambda' \text{ with } (\Lambda:\Lambda')=p^{r-1})\\ &= p\cdot\textstyle\sum[\Lambda''] \text{ (sum over } \Lambda''\subset p\Lambda \text{ with } (p\Lambda:\Lambda'')=p^{r-1}).\end{aligned}$$

Each of these is a sum of lattices Λ'' of index p^{r+1} in Λ. Fix such a lattice Λ'', and let a be the number of times that $[\Lambda'']$ occurs in the first sum, and b the number of times it occurs in the third sum. It occurs exactly once in the second sum, and so we have to prove that

$$a=1+pb.$$

There are two cases to consider.

The lattice Λ'' is not contained in $p\Lambda$. In this case, $b=0$, and a is the number of lattices Λ' such that $(\Lambda:\Lambda')=p$ and $\Lambda'\supset\Lambda''$. Such lattices are in one-to-one correspondence with the subgroups of $\Lambda/p\Lambda$ of index p containing the image $\bar{\Lambda}''$ of Λ'' in $\Lambda/p\Lambda$. But $(\Lambda:p\Lambda)=p^2$ and $\Lambda/p\Lambda\neq\bar{\Lambda}''\neq0$, and so there is only one such subgroup, namely $\bar{\Lambda}''$ itself. Therefore there is only one possible Λ', namely $p\Lambda+\Lambda''$, and so $a=1$.

The lattice $\Lambda''\supset p\Lambda$. Here $b=1$. Every lattice Λ' of index p in Λ contains $p\Lambda$, hence also Λ'', and the number of such Λ''s is the number of lines through the origin in $\Lambda/p\Lambda\approx\mathbb{F}_p^2$, i.e., the number of points in $\mathbb{P}^1(\mathbb{F}_p)$, which is $p+1$ as required. □

Corollary 26.13. *For any m and n,*

$$T(m)\circ T(n)=\sum d\cdot R(d)\circ T(mn/d^2)$$

(sum is over the positive divisors d of $\gcd(m,n)$*).*

Proof. Prove by induction on s that

$$T(p^r)T(p^s)=\sum_{i\leq r,s}p^i\cdot R(p^i)\circ T(p^{r+s-2i}),$$

and then apply (a) of the theorem. □

Corollary 26.14. *Let $\mathcal{H}$ be the $\mathbb{Z}$-subalgebra of* $\operatorname{End}(\mathcal{D})$ *generated by $T(p)$ and $R(p)$ for p prime; then $\mathcal{H}$ is commutative, and it contains $T(n)$ for all n.*

Proof. Obvious from the theorem. □

Let F be a function $\mathcal{L} \to \mathbb{C}$. We can extend F by linearity to a function $F : \mathcal{D} \to \mathbb{C}$,

$$F(\sum n_\Lambda[\Lambda]) = \sum n_\Lambda F(\Lambda).$$

For any linear map $T : \mathcal{D} \to \mathcal{D}$, we define $T \cdot F$ to be the function $\mathcal{L} \to \mathbb{C}$ such that $T \cdot F(\Lambda) = F(T\Lambda)$. For example,

$$(T(n) \cdot F)(\Lambda) = \sum_{(\Lambda:\Lambda')=n} F(\Lambda'),$$

and if $F(\lambda\Lambda) = \lambda^{-2k}F(\Lambda)$, then

$$R(n) \cdot F = n^{-2k}F.$$

Proposition 26.15. *If $F : \mathcal{L} \to \mathbb{C}$ has the property that $F(\lambda\Lambda) = \lambda^{-2k}F(\Lambda)$ for all λ, Λ, then so also does $T(n) \cdot F$, and*

(a) $T(mn) \cdot F = T(m) \cdot T(n) \cdot F \quad \text{if } \gcd(m,n) = 1$;
(b) $T(p) \cdot T(p^r) \cdot F = T(p^{r+1}) \cdot F + p^{1-2k}T(p^{r-1}) \cdot F$ *if p doesn't divide N;*
(c) $T(p^r) \cdot F = T(p)^r \cdot F$, $r \geq 1$, $p|N$.

Now let $f(z)$ be a modular form of weight $2k$, and let F be the associated function on $\mathcal{L}$. We define $T(n) \cdot f$ to be the function on $\mathbb{H}$ associated with $n^{2k-1} \cdot T(n) \cdot F$. Thus

$$(T(n) \cdot f)(z) = n^{2k-1}(T(n) \cdot F)(\Lambda(z,1)).$$

Theorem 26.8 in the case $N = 1$ follows easily from the Proposition. To prove Theorem 26.9 we need an explicit description of the lattices of index n in a fixed lattice.

Write $M_2(\mathbb{Z})$ for the ring of 2×2 matrices with coefficients in $\mathbb{Z}$.

Lemma 26.16. *For any $A \in M_2(\mathbb{Z})$, there exists a $U \in M_2(\mathbb{Z})^\times$ such that*

$$UA = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}, \quad ad = n, \quad a \geq 1, \quad 0 \leq b < d.$$

Moreover, the integers a, b, d are uniquely determined.

Proof. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and suppose $ra+sc = a'$ where $a' = \gcd(a,c)$. Then $\gcd(r,s) = 1$, and so there exist e, f such that $re + sf = 1$. Now

$$\begin{pmatrix} r & s \\ -f & e \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$$

and $\det \begin{pmatrix} r & s \\ -f & e \end{pmatrix} = 1$. Now apply the appropriate elementary row operations. For the uniqueness, note that multiplication by such a U doesn't change the greatest common divisor of the entries in any column, and so a is uniquely determined. Now d is uniquely determined by the equation $ad = n$, and b is obviously uniquely determined modulo d. □

For the lattice $\Lambda(z,1)$, the sublattices of index n are exactly the lattices $\Lambda(az+b, d)$ where (a,b,d) runs through the triples in the lemma. Therefore

$$(T(n) \cdot f)(z) = n^{2k-1} \sum_{a,b,d} d^{-2k} f(\frac{az+b}{d})$$

where the sum is over the same triples. On substituting this into the q-expansion

$$f = \sum_{m \geq 1} c(m) q^m$$

one finds (after a little work) that

$$T(n) \cdot f = c(n)q + \cdots .$$

Therefore, if $T(n) \cdot f = \lambda(n) f$, then

$$\lambda(n)c(1) = c(n).$$

This proves Theorem 26.9 in the case $N = 1$.

When $N \neq 1$, the theory of the Hecke operators is much the same, only a little more complicated. For example, instead of $\mathcal{L}$, one must work with the set of pairs (Λ, S) where $\Lambda \in \mathcal{L}$ and S is a cyclic subgroup of order N in $\mathbb{C}/\Lambda$. This is no problem for the $T(n)$'s with $\gcd(n, N) = 1$, but the $T(p)$'s with $p|N$ have to be treated differently.

Thus the problem of finding cusp forms f whose L-series have Euler products becomes a problem of finding simultaneous eigenforms for the linear map $T(n) : \mathcal{S}_{2k}(\Gamma_0(N)) \to \mathcal{S}_{2k}(\Gamma_0(N))$. Hecke had trouble doing this because he didn't know some linear algebra, which we now review.

Linear algebra: the spectral theorem. Recall that a *Hermitian form* on a vector space V is a mapping $<,>$: $V \times V \to \mathbb{C}$ such that $<v, w> = \overline{<w, v>}$ and $<,>$ is linear in one variable and conjugate-linear in the second. Such a form is said to be *positive-definite* if $<v, v> > 0$ whenever $v \neq 0$. A linear map $\alpha : V \to V$ is *Hermitian* or *self-adjoint* relative to $<,>$ if

$$<\alpha v, w> = <v, \alpha w>, \quad \text{all } v, w.$$

Theorem 26.17 (Spectral Theorem). *Let V be a finite-dimensional complex vector space with a positive-definite Hermitian form $<,>$.*

(a) *Any self-adjoint linear map $\alpha : V \to V$ is diagonalizable, i.e., V is a direct sum of eigenspaces for α.*

(b) *Let $\alpha_1, \alpha_2, \ldots$ be a sequence of commuting self-adjoint linear maps $V \to V$; then V has a basis of vectors that are eigenvectors for all α_i.*

Proof. (a) Because $\mathbb{C}$ is algebraically closed, α has an eigenvector e_1. Let V_1 be $(\mathbb{C}e_1)^\perp$. Then V_1 is stable under α, and so contains an eigenvector e_2. Let $V_2 = (\mathbb{C}e_1 \oplus \mathbb{C}e_2)^\perp$ etc.

(b) Now suppose $V = \oplus V(\lambda_i)$ where the λ_i are the distinct eigenvalues of α_1. Because α_2 commutes with α_1, it stabilizes each $V(\lambda_i)$, and so each $V(\lambda_i)$ can be decomposed into a direct sum of eigenspaces for α_2. Continuing in this fashion, we arrive at a decomposition $V = \oplus V_j$ such that each α_i acts as a scalar on each V_j. Choose bases for each V_j, and take their union. □

This suggests that we should look for a Hermitian form on $\mathcal{S}_{2k}(\Gamma_0(N))$ for which the $T(n)$'s are self-adjoint.

The Petersson inner product. As Poincaré pointed out, the unit disk forms a model for hyperbolic geometry[29]: if one defines a "line" to be a segment of a circle orthogonal to the circumference of the disk, angles to be the usual angles, and distances in terms of cross-ratios, one obtains a geometry that satisfies all the axioms for Euclidean geometry except that given a point P and a line ℓ, there exist *more* than one line through P not meeting ℓ. The map $z \mapsto \frac{z-i}{z+i}$ sends the upper-half plane onto the unit disk, and, being fractional-linear, maps circles and lines to circles and lines (collectively, not separately) and preserves angles. Therefore the upper half-plane is also a model for hyperbolic geometry. The group $\mathrm{PSL}_2(\mathbb{R}) =_{df} \mathrm{SL}_2(\mathbb{R})/\{\pm I\}$ is the group of transformations preserving distances and orientation, and therefore plays the same role as the group of orientation preserving affine transformations of the Euclidean plane. The next proposition shows that the measure $\mu(U) = \iint_U \frac{dxdy}{y^2}$ plays the same role as the measure $\iint_U dxdy$ on sets in the Euclidean plane—it is invariant under transformations in $\mathrm{PGL}_2(\mathbb{R})$.

Proposition 26.18. *Define $\mu(U) = \iint_U \frac{dxdy}{y^2}$; then $\mu(\gamma U) = \mu(U)$ for all $\gamma \in \mathrm{SL}_2(\mathbb{R})$.*

Proof. If $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then

$$\frac{d\gamma}{dz} = \frac{1}{(cz+d)^2}, \quad \Im(\gamma z) = \frac{\Im(z)}{|cz+d|^2}.$$

The next lemma shows that

$$\gamma^*(dxdy) = \frac{dxdy}{|cz+d|^4} \quad (z = x+iy),$$

and so $\frac{dxdy}{y^2}$, $y = \Im(z)$, is invariant under γ. □

Lemma 26.19. *For any holomorphic function $w(z)$, the map $z \mapsto w(z)$ multiplies areas by $|w'(z)|^2$.*

Proof. Write $w(z) = u(x,y) + iv(x,y)$, so that $z \mapsto w(z)$ is the map

$$(x,y) \mapsto (u(x,y), v(x,y)),$$

whose Jacobian is

$$\begin{vmatrix} u_x & v_x \\ u_y & v_y \end{vmatrix} = u_x v_y - v_x u_y.$$

On the otherhand, $w'(z) = u_x + iv_x$, so that

$$|w'(z)|^2 = u_x^2 + v_x^2.$$

The Cauchy-Riemann equations state that $u_x = v_y$ and $v_x = -u_y$, and so the two expressions agree. □

[29] Apparently Bolyai showed that it is possible to square the circle in hyperbolic geometry. A recent popular (shoddy) book on Fermat's Last Theorem contains the following mystifying statement (in italics): *If no one believes that it is possible to square the circle despite Bolyai's proof, why should we believe Wiles's proof of Fermat's last theorem, which also uses hyperbolic geometry.*

If f, g are modular forms of weight $2k$ for $\Gamma_0(N)$, then

$$f(z)\cdot\overline{g(z)}y^{2k}$$

is invariant under $\mathrm{SL}_2(\mathbb{R})$, which suggests defining

$$<f,g>=\iint_D f\bar{g}y^{2k}\frac{dxdy}{y^2}$$

for D a fundamental domain for $\Gamma_0(N)$—the above discussion shows that (assuming the integral converges) $<f,g>$ will be independent of the choice of D.

Theorem 26.20 (Petersson). *The above integral converges provided at least one of f or g is a cusp form. It therefore defines a positive-definite Hermitian form on the vector space $\mathcal{S}_{2k}(\Gamma_0(N))$ of cusp forms. The Hecke operators $T(n)$ are self-adjoint for all n relatively prime to N.*

Proof. Fairly straightforward calculus—see Knapp, p280. □

On putting the theorems of Hecke and Petersson together, we find that there exists a decomposition

$$\mathcal{S}_{2k}(\Gamma_0(N))=\oplus V_i$$

of $\mathcal{S}_{2k}$ into a direct sum of orthogonal subspaces V_i, each of which is a simultaneous eigenspace for all $T(n)$ with $\gcd(n,N)=1$. The $T(p)$ for $p|N$ stabilize each V_i and commute, and so there does exist at least one f in each V_i that is also an eigenform for the $T(p)$ with $p|N$. If we scale f so that $f=q+\sum_{n\geq 2}c(n)q^n$, then

$$L(f,s)=\prod_p\frac{1}{1-c(p)p^{-s}+p^{2k-1-2s}}\prod_{p|N}\frac{1}{1-c_pp^{-s}}$$

where the first product is over the primes not dividing N, and the second is over those dividing N.

The operator w_N is self-adjoint for the Petersson product, and does commute with the $T(n)$'s with $\gcd(n,N)=1$, and so each V_i decomposes into orthogonal eigenspaces

$$V_i=V_i^{+1}\oplus V_i^{-1}$$

for w_N. Unfortunately, w_N doesn't commute with the $T(p)$'s, $p|N$, and so the decompostion is not necessarily stable under these $T(p)$'s. Thus, the results above do not imply that there is a single f that is simultaneously an eigenvector for w_N (and hence has a functional equation) and for *all* $T(n)$ (and hence is equal to an Euler product).

New forms: the theorem of Atkin and Lehner. The problem left by the last subsection has a simple remedy. If $M|N$, then $\Gamma_0(M)\supset\Gamma_0(N)$, and so $\mathcal{S}_{2k}(\Gamma_0(M))\subset\mathcal{S}_{2k}(\Gamma_0(N))$. Recall that the N turns up in the functional equation for $L(f,s)$, and so it is not surprising that we run into trouble when we mix f's of "level" N with f's that are really of level $M|N$, $M<N$.

The way out of the problem is to define a cusp form that it in some subspace $\mathcal{S}_{2k}(\Gamma_0(M))$, $M|N$, $M<N$, to be *old*. The old forms form a subspace $\mathcal{S}_{2k}^{\mathrm{old}}(\Gamma_0(N))$ of $\mathcal{S}_{2k}(\Gamma_0(N))$, and the orthogonal complement $\mathcal{S}_{2k}^{\mathrm{new}}(\Gamma_0(N))$ is called the space of new forms. It is stable under all the operators $T(n)$ and w_N, and so $\mathcal{S}_{2k}^{\mathrm{new}}$ decomposes into a direct sum of orthogonal subspaces W_i,

$$\mathcal{S}_{2k}^{\mathrm{new}}(\Gamma_0(N))=\oplus W_i$$

each of which is a simultaneous eigenspace for all $T(n)$ with $\gcd(n,N)=1$. The $T(p)$ for $p|N$ and w_N stabilize each W_i.

Theorem 26.21 (Atkin-Lehner (1970)). *The spaces W_i in the above decomposition all have dimension* 1.

It follows that each W_i is also an eigenspace for w_N and $T(p)$, $p|N$. Each W_i contains (exactly) one cusp form f whose q-expansion is of the form $q+\sum_{n\geq 2}c(n)q^n$. For this form, $L(f,s)$ is equal to an Euler product, and $\Lambda(f,s)$ satisfies a functional equation

$$\Lambda(f,s)=\varepsilon\Lambda(f,2-s)$$

where $\varepsilon=\pm 1$ is the eigenvalue of w_N acting on W_i. If the $c(n)\in\mathbb{Z}$, then $\Lambda(f,s)$ is a candidate for being the L-function of an elliptic curve E over $\mathbb{Q}$.

Exercise 26.22. Let α,β,γ be integers, relatively prime in pairs, such that

$$\alpha^\ell+\beta^\ell=\gamma^\ell,$$

where ℓ is a prime $\neq 2,3$, and consider the elliptic curve

$$E: Y^2Z=X(X-\alpha^\ell Z)(X-\gamma^\ell Z).$$

(a) Show that E has discriminant $\Delta=16\alpha^{2\ell}\beta^{2\ell}\gamma^{2\ell}$.

(b) Show that if p does not divide $\alpha\beta\gamma$, then E has good reduction at p.

(c) Show that if p is an odd prime dividing $\alpha\beta\gamma$, then E has at worst nodal reduction at p.

(d) Show that (the minimal equation for) E has at worst nodal reduction at 2. [[After possibly re-ordering α,β,γ, we may suppose, first that γ is even, and then that $\alpha^\ell\equiv 1$ mod 4. Make the change of variables $x=4X$, $y=8Y+4X$, and verify that the resulting equation has integer coefficients.]]

(b),(c),(d) show that the conductor N of E divides $\prod_{p|\alpha\beta\gamma}p$, and hence is much smaller than Δ. This is enough to show that E doesn't exist, but the enthusiasts may wish to verify that $N=\prod_{p|\alpha\beta\gamma}p$. [Hint: First show that if p doesn't divide c_4, then the equation is minimal at p.]

27. Statement of the Main Theorems

Recall that to an elliptic curve E over $\mathbb{Q}$, we have attached an L-series $L(E,s)=\sum a_n n^{-s}$ that has coefficients $a_n\in\mathbb{Z}$, can be expressed as an Euler product, and (conjecturally) satisfies a functional equation (involving $N_{E/\mathbb{Q}}$, the conductor on E). Moreover, isogenous elliptic curves have the same L-series. We therefore have a map

$$E\mapsto L(E,s):\{\text{elliptic curves}/\mathbb{Q}\}/\!\sim\;\to\{\text{Dirichlet series}\}.$$

An important theorem of Faltings (1983) shows that the map is injective: two elliptic curves are isogenous if they have the same L-function.

On the other hand, the theory of Hecke and Petersson, together with the theorem of Atkin and Lehner, shows that the subspace $\mathcal{S}_2^{\text{new}}(\Gamma_0(N))\subset\mathcal{S}_2(\Gamma_0(N))$ of new forms decomposes into a direct sum

$$\mathcal{S}_2^{\text{new}}(\Gamma_0(N))=\oplus W_i$$

of one-dimensional subspaces W_i that are simultaneous eigenspaces for all the $T(n)$'s with $\gcd(n,N)=1$. Because they have dimension 1, each W_i is also an eigenspace for w_N and for the $T(p)$ with $p|N$. An element of one of the subspaces W_i, i.e., a simultaneous eigenforms in $\mathcal{S}_2^{\text{new}}(\Gamma_0(N))$, is traditionally called a *newform*, and I'll adopt this terminology.

In each W_i there is exactly one form $f_i=\sum c(n)q^n$ with $c(1)=1$ (said to be normalized). Because f_i is an eigenform for all the Hecke operators, it has an Euler product, and because it is an eigenform for w_N, it satisfies a functional equation. If the $c(n)$'s are[30] in $\mathbb{Z}$, then $L(f_i,s)$ is a candidate for being the L-function of an elliptic curve over $\mathbb{Q}$.

Conjecture 27.1. *The following sets are equal:*

$$\{L(E,s)\mid E \text{ an elliptic curve over } \mathbb{Q} \text{ with conductor } N\}$$

$$\{L(f,s)\mid f \text{ a normalized newform for } \Gamma_0(N)\text{, i.e., } f=f_i \text{ some } i\}.$$

The following theorem of Eichler and Shimura (and others) (1954/1958/...) shows that the second set is contained in the first.

Theorem 27.2 (Eichler-Shimura). *Let $f=\sum c(n)q^n$ be a normalized newform for $\Gamma_0(N)$. If all $c(n)\in\mathbb{Z}$, then there exists an elliptic curve E_f of conductor N such that $L(E_f,s)=L(f,s)$.*

The early forms of the theorem were less precise—in particular, they predate the work of Atkin and Lehner in which newforms were defined.

The theorem of Eichler-Shimura has two parts: given f, construct the curve E_f (up to isogeny); having constructed E_f, prove that $L(E_f,s)=L(f,s)$. I'll discuss the two parts in Sections 28 and 29.

After the theorem of Eichler-Shimura, to prove Conjecture 27.1, it remains to show that every elliptic curve E arises from a modular form f—such an elliptic curve is said to be *modular*.

In a set of problems circulated (in Japanese) to the members of a conference in 1955, Taniyama asked (in rather vague form) whether every elliptic curve was modular. In the ensuing years, this question was apparently discussed by various people, including Shimura, who however published nothing about it.

One can ask whether *every* Dirichlet L-series $L(s)=\sum a_n n^{-s}$, $a_n\in\mathbb{Z}$, equal to an Euler product (of the same type as $L(E,s)$), and satisfying a functional equation (of the same type as $L(E,s)$) must automatically be of the form $L(f,s)$. This is not so, but Weil (1967) proved something only a little weaker. Let $\chi:(\mathbb{Z}/n\mathbb{Z})^\times\to\mathbb{C}^\times$, $\gcd(n,N)=1$, be a homomorphism, and extend χ to a map $\mathbb{Z}\to\mathbb{C}$ by setting $\chi(m)=\chi(m \mod n)$ if m and n are relatively prime and $=0$ otherwise. Define

$$L_\chi(s)=\sum\chi(n)a_n n^{-s},\quad \Lambda_\chi(s)=\left(\frac{m}{2\pi}\right)^{-s}\Gamma(s)L_\chi(s).$$

Weil showed that if all the functions $\Lambda_\chi(s)$ satisfy a functional equation relating $\Lambda_\chi(s)$ and $\Lambda_\chi(2k-s)$ (and some other mild conditions), then $L(s)=L(f,s)$ for some cusp form f of

[30] In the next section, we shall see that the $c(n)$'s automatically lie in some finite extension of $\mathbb{Q}$, and that if they lie in $\mathbb{Q}$ then they lie in $\mathbb{Z}$

weight $2k$ for $\Gamma_0(N)$. Weil also stated Conjecture 27.1 (as an exercise!)—this was its first appearance in print.

Weil's result showed that if $L(E, s)$ and its twists satisfy a functional equation of the correct form, then E is modular. Since the Hasse-Weil conjecture was widely believed, Weil's paper (for the first time) gave a strong reason for believing Conjecture 27.1, i.e., it made (27.1) into a conjecture rather than a question. Also, for the first time it related the level N of f to the conductor of E, and so made it possible to test the conjecture numerically: list all the f's for $\Gamma_0(N)$, list all isogeny classes of elliptic curves over $\mathbb{Q}$ with conductor N, and see whether they match. A small industry grew up to do just that.

For several years, the conjecture was referred to as Weil's conjecture. Then, after Taniyama's question was rediscovered, it was called the Taniyama-Weil conjecture. Finally, after Lang adopted it as one of his pet projects[31], it became unsafe to call it anything other than the Shimura-Taniyama conjecture—see Lang's scurrilous article in the Notices of the AMS, November 1995, pp 1301–1307.

In a lecture in 1985, Frey suggested that the curve in Exercise 26.22, defined by a counterexample to Fermats' Last Theorem, should not be modular. This encouraged Serre to rethink some old conjectures of his, and formulate two conjectures, one of which implies that Frey's curve is indeed not modular. In 1986, Ribet proved sufficient of Serre's conjectures to be able to show that Frey's curve can't be modular. I'll discuss this work in Section 31.

Thus, at this stage (1986) it was known that Conjecture 27.1 for semistable elliptic curves over $\mathbb{Q}$ implies Fermat's Last Theorem, which inspired Wiles to attempt to prove Conjecture 27.1. After a premature announcement in 1993, Wiles proved in 1994 (with the help of R. Taylor) that all semistable elliptic curves over $\mathbb{Q}$ are modular. Recall that semistable just means that the curve doesn't have cuspidal reduction at any prime. Diamond improved the theorem so that it now says that an elliptic curve E over $\mathbb{Q}$ is modular provided it doesn't have additive reduction at 3 or 5. In other words, the image of the map

$$f \mapsto E_f : \{f\} \to \{E \text{ over } \mathbb{Q}\}/\sim$$

contains (at least) all E's with at worst nodal reduction at 3 and 5. Needless to say, efforts are being made to remove this last condition. I'll discuss the strategy of Wiles's proof in Section 30.

28. How to get an Elliptic Curve from a Cusp Form

Not long after Newton and Leibniz invented calculus, mathematicians discovered that they couldn't evaluate integrals of the form

$$\int \frac{dx}{\sqrt{f(x)}}$$

where $f(x) \in \mathbb{R}[x]$ is a cubic polynomial without a repeated factor. In fact, such an integral can't be evaluated in terms of elementary functions. Thus, they were forced to treat them

[31] To the great benefit of the Xerox Co., as Weil put it—I once made some of the points in the above paragraph to Lang and received a 40 page response.

as new functions and to study their properties. For example, Euler showed that

$$\int_a^{t_1} \frac{dx}{\sqrt{f(x)}} + \int_a^{t_2} \frac{dx}{\sqrt{f(x)}} = \int_a^{t_3} \frac{dx}{\sqrt{f(x)}}$$

where t_3 is a rational function of t_1, t_2. The explanation for this lies with elliptic curves.

Consider the elliptic curve $Y^2 = f(X)$ over $\mathbb{R}$, and the differential one-form $\omega = \frac{1}{y}dx + 0dy$ on $\mathbb{R}^2$. According to Math 215, to integrate ω over a segment of the elliptic curve, we should parametrize the curve. We assume that the segment $\gamma(a,t)$ of the elliptic curve over $[a,t]$ can be smoothly parametrized by x. Thus the segment is

$$x \mapsto (x, \sqrt{f(x)}), \quad x \in [a,t].$$

Then, again according to Math 215,

$$\int_{\gamma(a,t)} \frac{dx}{y} = \int_a^t \frac{dx}{\sqrt{f(x)}}.$$

Thus, the elliptic integral can be regarded as an integral over a segment of an elliptic curve.

A key point, which we'll discuss later, is that the restriction of ω to E is translation invariant, i.e., if t_Q denotes the map $P \mapsto P + Q$ on E, then $t_Q^*\omega = \omega$ (on E). Hence

$$\int_{\gamma(a,t)} \omega = \int_{\gamma(a+x(Q),t+x(Q))} \omega$$

for any $Q \in E(\mathbb{R})$ (here $x(Q)$ is the x-coordinate of Q). Now Euler's theorem becomes the statement

$$\int_{\gamma(a,t_1)} \omega + \int_{\gamma(a,t_2)} = \int_{\gamma(a,t_1)} \omega + \int_{\gamma(t_1,t_3)} \omega = \int_{\gamma(a,t_3)} \omega$$

where t_3 is determined by

$$(t_2, \sqrt{f(t_2)}) - (a, \sqrt{f(a)}) + (t_1, \sqrt{f(t_1)}) = (t_3, \sqrt{f(t_3)})$$

(difference and sum for the group structure on $E(\mathbb{R})$).

Thus the study of elliptic integrals leads to the study of elliptic curves. Jacobi and Abel showed that the study of more complicated integrals leads to other interesting varieties.

Differentials on Riemann surfaces. A differential one-form on an open subset of $\mathbb{C}$ is simply an expression $\omega = fdz$, with f a meromorphic function. Given a smooth curve γ

$$t \mapsto z(t) : [a,b] \to \mathbb{C}, \quad [a,b] = \{t \in \mathbb{R} \mid a \leq t \leq b\},$$

we can form the integral

$$\int_\gamma \omega = \int_a^b f(z(t)) \cdot z'(t) \cdot dt \in \mathbb{C}.$$

Now consider a compact Riemann surface X. If ω is a differential one-form on X and (U_i, z_i) is a coordinate neighbourhood for X, then $\omega|U_i = f_i(z_i)dz_i$. If (U_j, z_j) is a second coordinate neighbourhood, so that $z_j = w(z_i)$ on $U_i \cap U_j$, then

$$f_i(z_i)dz_i = f_j(w(z_i))w'(z_i)dz_i$$

on $U_i \cap U_j$. Thus, to give a differential one-form on X is to give differential one-forms $f_i dz_i$ on each U_i, satisfying the above equation on the overlaps. For any (real) curve $\gamma : I \to X$ and differential one-form ω on X, the integral $\int_\gamma \omega$ makes sense.

A differential one-form is *holomorphic* if it is represented on the coordinated neighbourhoods by forms $f dz$ with f holomorphic.

It is an important fact (already noted) that the holomorphic differential one-forms on a Riemann surface of genus g form a complex vector space $\Omega^1(X)$ of dimension g.

For example, the Riemann sphere has genus 0 and so should have no nonzero holomorphic differential one-forms. Note that dz is holomorphic on $\mathbb{C} = S \setminus \{\text{north pole}\}$, but that $z = 1/z'$ on $S \setminus \{\text{poles}\}$, and so $dz = -\frac{1}{z'^2}dz'$, which has a pole at the north pole. Hence dz does not extend to a holomorphic differential one-form on the whole of S.

An elliptic curve has genus 1, and so the holomorphic differential one-forms on it form a vector space of dimension 1. It is generated by $\omega = \frac{dx}{2y}$ (more accurately, the restriction of $\frac{1}{2y}dx + 0dy$ to $E(\mathbb{C}) \subset \mathbb{C}^2$). Here I'm assuming that E has equation

$$Y^2Z = X^3 + aXZ^2 + bZ^3, \quad \Delta \neq 0.$$

Note that, on E^{aff},

$$2ydy = (3x^2 + a)dx,$$

and so

$$\frac{dx}{2y} = \frac{dy}{3x^2 + a}$$

where both are defined. Thus it is holomorphic on E^{aff}, and one can check that it also holomorphic at the point at infinity.

For any $Q \in E(\mathbb{C})$, $t_Q^*\omega$ is also holomorphic, and so $t_Q^*\omega = c\omega$ for some $c \in \mathbb{C}$. Now $Q \mapsto c : E(\mathbb{C}) \to \mathbb{C}$ is a holomorphic function on $\mathbb{C}$, and all such functions are constant (see 10.3). Since the function takes the value 1 when $Q = 0$, it is 1 for all Q, and so ω is invariant under translation. Alternatively, one can simply note that the inverse image of ω under the map

$$(x, y) = (\wp(z), \wp'(z)), \quad \mathbb{C} \setminus \Lambda \to E^{\text{aff}}(\mathbb{C})$$

is

$$\frac{d\wp(z)}{2\wp'(z)} = \frac{dz}{2},$$

which is clearly translation invariant on $\mathbb{C}$—$d(z + c) = dz$.

The Jacobian variety of a Riemann surface. Consider an elliptic curve over E and a nonzero holomorphic differential one-form ω. We choose a point $P_0 \in E(\mathbb{C})$ and try to define a map

$$P \mapsto \int_{P_0}^{P} \omega : E(\mathbb{C}) \to \mathbb{C}.$$

This is not well-defined because the value of the integral depends on the path we choose from P_0 to P—nonhomotopic paths may give different answers. However, if we choose a basis $\{\gamma_1, \gamma_2\}$ for $H_1(E(\mathbb{C}), \mathbb{Z})(= \pi_1(E(\mathbb{C}), P_0))$, then the integral is well-defined modulo the lattice Λ in $\mathbb{C}$ generated by

$$\int_{\gamma_1} \omega, \quad \int_{\gamma_2} \omega.$$

In this way, we obtain an isomorphism

$$P \mapsto \int_{P_0}^{P} \omega : E(\mathbb{C}) \to \mathbb{C}/\Lambda.$$

Note that this construction is inverse to that in Section 10.

Jacobi and Abel made a similar construction for any compact Riemann surface X. Suppose X has genus g, and let $\omega_1, \ldots, \omega_g$ be a basis for the vector space $\Omega^1(X)$ of holomorphic one-forms on X. Choose a point $P_0 \in X$. Then there is a smallest lattice Λ in $\mathbb{C}^g$ such that the map

$$P \mapsto \left(\int_{P_0}^{P} \omega_1, \ldots, \int_{P_0}^{P} \omega_g\right) : X \to \mathbb{C}^g/\Lambda$$

is well-defined. By a lattice in $\mathbb{C}^g$, I mean the free $\mathbb{Z}$-submodule of rank $2g$ generated by a basis for $\mathbb{C}^g$ regarded as a real vector space (strictly, this is a full lattice). The quotient $\mathbb{C}^g/\Lambda$ is a complex manifold, called the *Jacobian variety* $Jac(X)$ of X, which can be considered to be a higher-dimensional analogue of $\mathbb{C}/\Lambda$. Note that it is a commutative group.

We can make the definition of $Jac(X)$ more canonical. Let $\Omega^1(X)^\vee$ be the dual of $\Omega^1(X)$ as a complex vector space. For any $\gamma \in H_1(X, \mathbb{Z})$,

$$\omega \mapsto \int_\gamma \omega$$

is an element of $\Omega^1(X)^\vee$, and in this way we obtain an injective homomorphism

$$H_1(X, \mathbb{Z}) \hookrightarrow \Omega^1(X)^\vee,$$

which (one can prove) identifies $H_1(X, \mathbb{Z})$ with a lattice in $\Omega^1(X)^\vee$. Define

$$Jac(X) = \Omega^1(X)^\vee / H_1(X, \mathbb{Z}).$$

When we fix a $P_0 \in X$, any $P \in X$ defines an element

$$\omega \mapsto \int_{P_0}^{P} \omega \mod H_1(X, \mathbb{Z})$$

of $Jac(X)$, and so we get a map $X \to Jac(X)$. The choice of a different P_0 gives a map that differs from the first only by a translation.

Construction of the elliptic curve over $\mathbb{C}$. We apply the above theory to the Riemann surface $X_0(N)$. Let π be the map $\pi : \mathbb{H} \to X_0(N)$ (not quite onto). For any $\omega \in \Omega^1(X)$, $\pi^*\omega = f dz$ where $f \in \mathcal{S}_2(X_0(N))$, and the map $\omega \mapsto f$ is a bijection

$$\Omega^1(X) \to \mathcal{S}_2(X_0(N))$$

(see 25.3). The Hecke operator $T(n)$ acts on $\mathcal{S}_2(X_0(N))$, and hence on $\Omega^1(X)$ and its dual.

Proposition 28.1. *There is a canonical action of $T(n)$ on $H_1(X_0(N), \mathbb{Z})$, which is compatible with the map $H_1(X_0(N), \mathbb{Z}) \to \Omega^1(X_0(N))^\vee$. In other words, the action of $T(n)$ on $\Omega^1(X)^\vee$ stabilizes its sublattice $H_1(X_0(N), \mathbb{Z})$, and therefore induces an action on the quotient $Jac(X_0(N))$.*

Proof. One can give an explicit set of generators for $H_1(X_0(N), \mathbb{Z})$, explicitly describe an action of $T(n)$ on them, and then explicitly verify that this action is compatible with the map $H_1(X_0(N), \mathbb{Z}) \to \Omega^1(X_0(N))^\vee$. Alternatively, as we discuss in the next section, there are more geometric reasons why the $T(n)$ should act on $Jac(X)$. □

Remark 28.2. From the action of $T(n)$ on $H_1(X, \mathbb{Z}) \approx \mathbb{Z}^{2g}$ we get a characteristic polynomial $P(Y) \in \mathbb{Z}[Y]$ of degree $2g$. What is its relation to the characteristic polynomial $Q(Y) \in \mathbb{C}[Y]$ of $T(n)$ acting on $\Omega^1(X)^\vee \approx \mathbb{C}^g$? The obvious guess is that $P(Y) = Q(Y)\overline{Q(Y)}$. The proof that this is so is an exercise in linear algebra. See the next section.

Now let $f = \sum c(n)q^n$ be a normalized newform for $\Gamma_0(N)$ with $c(n) \in \mathbb{Z}$. The map

$$\alpha \mapsto \alpha(f) : \Omega^1(X_0(N))^\vee \to \mathbb{C}$$

identifies $\mathbb{C}$ with the largest quotient of $\Omega^1(X)^\vee$ on which each $T(n)$ acts as multiplication by $c(n)$. The image of $H_1(X_0(N), \mathbb{Z})$ is a lattice Λ_f, and we set $E_f = \mathbb{C}/\Lambda_f$—it is an elliptic curve over $\mathbb{C}$. Note that we have constructed maps

$$X_0(N) \to Jac(X_0(N)) \to E_f.$$

The inverse image of the differential on E_f represented by dz is the differential on $X_0(N)$ represented by fdz.

Construction of the elliptic curve over $\mathbb{Q}$. We briefly explain why the above construction in fact gives an elliptic curve over $\mathbb{Q}$. There will be a few more details in the next section.

For a compact Riemann surface X, we defined

$$Jac(X) = \Omega^1(X)^\vee / H_1(X, \mathbb{Z}) \approx \mathbb{C}^g/\Lambda, \quad g = \text{genus}X.$$

This is a complex manifold, but as in the case of an elliptic curve, it is possible to construct enough functions on it to embed it into projective space, and so realize it as a projective algebraic variety.

Now suppose X is a nonsingular projective curve over an field k. Weil showed (as part of the work mentioned on p102) that it is possible to attach to X a projective algebraic variety $Jac(X)$ over k, which, in the case $k = \mathbb{C}$ becomes the variety defined in the last paragraph. There is again a map $X \to Jac(X)$, well-defined up to translation by the choice of a point $P_0 \in X(k)$. The variety $Jac(X)$ is an *abelian variety*, i.e., not only is it projective, but it also has a group structure. (An abelian variety of dimension 1 is an elliptic curve.)

In particular, there is such a variety attached to the curve $X_0(N)$ defined in Section 24. Moreover (see the next section), the Hecke operators $T(n)$ define endomorphisms of $Jac(X_0(N))$. Because it has an abelian group structure, any integer m defines an endomorphism of $Jac(X_0(N))$, and we define E_f to be the largest "quotient" of $Jac(X_0(N))$ on which $T(n)$ and $c(n)$ agree for all n relatively prime to N. One can prove that this operation of "passing to the quotient" commutes with change of the ground field, and so in this way we obtain an elliptic curve over $\mathbb{Q}$ that becomes equal over $\mathbb{C}$ to the curve defined in the last subsection. On composing $X_0(N) \to Jac(X_0(N))$ with $Jac(X_0(N)) \to E_f$ we obtain a map $X_0(N) \to E_f$. In summary:

Theorem 28.3. *Let $f = \sum c(n)q^n$ be a newform in $\mathcal{S}_2(\Gamma_0(N))$, normalized to have $c(1) = 1$, and assume that all $c(n) \in \mathbb{Z}$. Then there exists an elliptic curve E_f and a map $\alpha : X_0(N) \to E_f$ with the following properties:*

(a) *α factors uniquely through $Jac(X_0(N))$,*

$$X_0(N) \to Jac(X_0(N)) \to E_f,$$

and the second map realizes E_f as the largest quotient of $Jac(X_0(N))$ on which the endomorphisms $T(n)$ and $c(n)$ of $Jac(X_0(N))$ agree.

(b) *The inverse image of an invariant differential ω on E_f under $\mathbb{H} \to X_0(N) \to E_f$ is a nonzero rational multiple of fdz.*

29. Why the L-Series of E Agrees with the L-Series of f

In this section we sketch a proof of the identity of Eichler and Shimura relating the Hecke correspondence $T(p)$ to the Frobenius map, and hence the L-series of E_f to that of f.

The ring of correspondences of a curve. Let X and X' be projective nonsingular curves over a field k which, for simplicity, we take to be algebraically closed.

A *correspondence* T between X and X', written $T : X \vdash X'$, is a pair of finite surjective regular maps

$$X \xleftarrow{\alpha} Y \xrightarrow{\beta} X'.$$

It can be thought of as a many-valued map $X \to X'$ sending a point $P \in X(k)$ to the set $\{\beta(Q_i)\}$ where the Q_i run through the elements of $\alpha^{-1}(P)$ (the Q_i need not be distinct). Better, define $\mathrm{Div}(X)$ to be the free abelian group on the set of points of X; thus an element of $\mathrm{Div}(X)$ is a finite formal sum

$$D = \sum n_P P, \quad n_P \in \mathbb{Z}, \quad P \in X(k).$$

A correspondence T then defines a map

$$\mathrm{Div}(X) \to \mathrm{Div}(X'), \quad P \mapsto \sum \beta(Q_i).$$

(notations as above). This map multiplies the degree of a divisor by $\deg(\alpha)$. It therefore sends the divisors of degree zero on X into the divisors of degree zero on X', and one can show that it sends principal divisors to principal divisors. Hence it defines a map $T : J(X) \to J(X')$ where

$$J(X) =_{df} \mathrm{Div}^0(X)/\{\text{principal divisors}\}.$$

We define the *ring of correspondences* $\mathcal{A}(X)$ on X to be the subring of $\mathrm{End}(J(X))$ generated by the maps defined by correspondences.

If T is the correspondence

$$X \xleftarrow{\beta} Y \xrightarrow{\alpha} X,$$

then the transpose T^{tr} of T is the correspondence

$$X \xleftarrow{\alpha} Y \xrightarrow{\beta} X.$$

A morphism $\alpha : X \to X'$ can be thought of as a correspondence

$$X \leftarrow \Gamma \to X'$$

where $\Gamma \subset X \times X'$ is the graph of α and the maps are the projections. The transpose of a morphism α is the many valued map $P \mapsto \alpha^{-1}(P)$.

Remark 29.1. Let U and U' be the curves obtained from X and X' by removing a finite number of points. Then, it follows from the theory of algebraic curves, that a regular map $\alpha : U \to U'$ extends *uniquely* to a regular map $\bar{\alpha} : X \to X'$: take $\bar{\alpha}$ to be the regular map whose graph is the Zariski closure of the graph of α. On applying this remark twice, we see that a correspondence $U \vdash U'$ extends uniquely to a correspondence $X \vdash X'$.

Remark 29.2. Let

$$X \xleftarrow{\alpha} Y \xrightarrow{\beta} X'.$$

be a correspondence $T : X \vdash X'$. For any regular function f on X', we define $T(f)$ to be the regular function $P \mapsto \sum f(\beta Q_i)$ on X (notation as above). Similarly, T will define a homomorphism $\Omega^1(X') \to \Omega^1(X)$.

The Hecke correspondence. For $p \nmid N$, the Hecke correspondence $T(p) : Y_0(N) \to Y_0(N)$ is defined to be

$$Y_0(N) \xleftarrow{\alpha} Y_0(pN) \xrightarrow{\beta} Y_0(N)$$

where α is the obvious projection map and β is the map induced by $z \mapsto pz : \mathbb{H} \to \mathbb{H}$.

On points, it has the following description. Recall that a point of $Y_0(pN)$ is represented by a pair (E, S) where E is an elliptic curve and S is a cyclic subgroup of E of order pN. Because $p \nmid N$, any such subgroup decomposes uniquely into subgroups of order N and p, $S = S_N \times S_p$. The map α sends the point represented by (E, S) to the point represented by (E, S_N), and β sends it to the point represented by $(E/S_p, S/S_p)$. Since E_p has $p+1$ cyclic subgroups, the correspondence is $1 : p+1$.

The unique extension of $T(p)$ to a correspondence $X_0(N) \to X_0(N)$ acts on $\Omega^1(X_0(N)) = \mathcal{S}_2(X_0(N))$ as the Hecke correspondence defined in Section 26. This description of $T(p)$, $p \nmid N$, makes sense, and is defined on, the curve $X_0(N)$ over $\mathbb{Q}$. Similar remarks apply[32] to the $T(p)$ for $p|N$.

The Frobenius map. Let C be a curve defined over the algebraic closure $\mathbb{F}$ of $\mathbb{F}_p$. If C is defined by equations

$$\sum a_{i_0 i_1 \cdots} X_0^{i_0} X_1^{i_1} \cdots = 0,$$

then $C^{(p)}$ is defined by equations

$$\sum a_{i_0 i_1 \cdots}^p X_0^{i_0} X_1^{i_1} \cdots = 0,$$

and the *Frobenius map* $\varphi_p : C \to C^{(p)}$ sends the point $(b_0 : b_1 : b_2 : \ldots)$ to $(b_0^p : b_1^p : b_2^p : \ldots)$. If C is defined over $\mathbb{F}_p$, then $C = C^{(p)}$ and φ_p is the Frobenious map defined earlier.

Recall that a nonconstant morphism $\alpha : C \to C'$ of curves defines an inclusion $\alpha^* : k(C') \hookrightarrow k(C)$ of function fields, and that the degree of α is defined to be $[k(C) : \alpha^* k(C')]$. The map α is said to be *separable* or *purely inseparable* according as $k(C)$ is a separable of purely inseparable extension of $\alpha^* k(C')$. If the separable degree of $k(C)$ over $\alpha^* k(C')$ is m, then the map $C(k^{\text{al}}) \to C'(k^{\text{al}})$ is $m : 1$, except over the finite set where it is ramified.

Proposition 29.3. *The Frobenius map $\varphi_p : C \to C^{(p)}$ is purely inseparable of degree p, and any purely inseparable map $\varphi : C \to C'$ of degree p (of complete nonsingular curves) factors as*

$$C \xrightarrow{\varphi_p} C^{(p)} \xrightarrow{\approx} C'.$$

Proof. For $C = \mathbb{P}^1$, this is obvious, and the general case follows because $\mathbb{F}(C)$ is a separable extension of $\mathbb{F}(T)$. See [S1, II.2.12] for the details. □

Brief review of the points of order p on elliptic curves. Let E be an elliptic curve over an algebraically closed field k. The map $p : E \to E$ (multiplication by p) is of degree p^2. If k has characteristic zero, then the map is separable, which implies that its kernel has order p^2. If k has characteristic p, the map is never separable: either it is purely inseparable (and so E has no points of order p) or its separable and inseparable degrees are p (and so E has p points of order dividing p). The first case occurs for only finitely many values of j.

[32] These $T(p)$'s are sometimes denoted $U(p)$.

The Eichler-Shimura relation. The curve $X_0(N)$ and the Hecke correspondence $T(p)$ are defined over $\mathbb{Q}$. For almost all primes $p \nmid N$, $X_0(N)$ will reduce to a nonsingular curve $\widetilde{X}_0(N)$.[33] For such a prime p, the correspondence $T(p)$ defines a correspondence $\widetilde{T}(p)$ on $\widetilde{X}_0(N)$.

Theorem 29.4. *For a prime p where $X_0(N)$ has good reduction,*

$$\widetilde{T}(p) = \varphi_p + \varphi_p^{\mathrm{tr}}.$$

(Equality in the ring $\mathcal{A}(\widetilde{X}_0(N))$ of correspondences on $\widetilde{X}_0(N)$ over the algebraic closure $\mathbb{F}$ of $\mathbb{F}_p$.)

Proof. We sketch a proof that they agree as many-valued maps on an open subset of $\widetilde{X}_0(N)$.

Over $\mathbb{Q}_p^{\mathrm{al}}$ we have the following description of $T(p)$ (see above): a point P on $Y_0(N)$ is represented by a homomorphism of elliptic curves $\alpha : E \to E'$ with cyclic kernel of order N; let $S_0, \ldots, S_p$ be the subgroups of order p in E; then $T_p(P) = \{Q_0, \ldots, Q_p\}$ where Q_i is represented by $E/S_i \to E'/\alpha(S_i)$.

Consider a point $\widetilde{P}$ on $\widetilde{X}_0(N)$ with coordinates in $\mathbb{F}$—by Hensel's lemma it will lift to a point on $X_0(N)$ with coordinates in $\mathbb{Q}_p^{\mathrm{al}}$. Ignoring a finite number of points of $\widetilde{X}_0(N)$, we can suppose $\widetilde{P} \in \widetilde{Y}_0(N)$ and hence is represented by a map $\tilde{\alpha} : \widetilde{E} \to \widetilde{E}'$ where $\alpha : E \to E'$ has cyclic kernel of order N. By ignoring a further finite number of points, we may suppose that $\widetilde{E}$ has p points of order dividing p.

Let $\alpha : E \to E'$ be a lifting of $\tilde{\alpha}$ to $\mathbb{Q}_p^{\mathrm{al}}$. The reduction map $E_p(\mathbb{Q}_p^{\mathrm{al}}) \to \widetilde{\mathbb{E}}_p(\mathbb{F}_p^{\mathrm{al}})$ has a kernel of order p. Number the subgroups of order p in E so that S_0 is the kernel of this map. Then each S_i, $i \neq 0$, maps to a subgroup of order p in $\widetilde{E}$.

The map $p : \widetilde{E} \to \widetilde{E}$ has factorizations

$$\widetilde{E} \xrightarrow{\varphi} \widetilde{E}/S_i \xrightarrow{\psi} \widetilde{E}, \quad i = 0, 1, \ldots, p.$$

When $i = 0$, φ is a purely inseparable map of degree p (it is the reduction of the map $E \to E/S_0$—it therefore has degree p and has zero kernel), and so ψ must be separable of degree p (we are assuming $\widetilde{E}$ has p points of order dividing p). Proposition 29.3 shows that there is an isomorphism $\widetilde{E}^{(p)} \to \widetilde{E}/S_0$. Similarly $\widetilde{E}'^{(p)} \approx \widetilde{E}'/S_0$. Therefore Q_0 is represented by $\widetilde{E}^{(p)} \to \widetilde{E}'^{(p)}$, which also represents $\varphi_p(P)$.

When $i \neq 0$, φ is separable (its kernel is the reduction of S_i), and so ψ is purely inseparable. Therefore $\widetilde{E} \approx \widetilde{E}_i^{(p)}$, and similarly $\widetilde{E}' \approx \widetilde{E}_i'^{(p)}$, where $\widetilde{E}_i/\widetilde{E}/S_i$ and $\widetilde{E}_i' = \widetilde{E}'/S_i$. It follows that $\{Q_1, \ldots, Q_p\} = \varphi_p^{-1}(P) = \varphi_p^{\mathrm{tr}}(P)$. $\square$

The zeta function of an elliptic curve revisited. We begin with an elementary result from linear algebra.

Proposition 29.5. *Let Λ be a free $\mathbb{Z}$-module of finite rank, and let $\alpha : \Lambda \to \Lambda$ be a $\mathbb{Z}$-linear map with nonzero determinant. Then the kernel of the map*

$$\tilde{\alpha} : (\Lambda \otimes \mathbb{Q})/\Lambda \to (\Lambda \otimes \mathbb{Q})/\Lambda$$

defined by α has order $|\det(\alpha)|$.

[33] In fact, it is known that $X_0(N)$ has good reduction for all primes $p \nmid N$, but this is hard to prove. It is easy to see that $X_0(N)$ does not have good reduction at primes dividing N.

Proof. Consider the commutative diagram:

$$\begin{tikzcd}
0 \arrow[r] & \Lambda \arrow[r] \arrow[d, "\alpha"] & \Lambda\otimes\mathbb{Q} \arrow[r] \arrow[d, "\alpha\otimes 1"] & (\Lambda\otimes\mathbb{Q})/\Lambda \arrow[r] \arrow[d, "\widetilde{\alpha}"] & 0 \\
0 \arrow[r] & \Lambda \arrow[r] & \Lambda\otimes\mathbb{Q} \arrow[r] & (\Lambda\otimes\mathbb{Q})/\Lambda \arrow[r] & 0.
\end{tikzcd}$$

Because $\det(\alpha)\neq 0$, the middle vertical map is an isomorphism. Therefore the snake lemma gives an isomorphism

$$\operatorname{Ker}(\widetilde{\alpha})\to\operatorname{Coker}(\alpha),$$

and it is easy to see that $\operatorname{Coker}(\alpha)$ is finite with order equal to $|\det(\alpha)|$. $\square$

We apply this to an elliptic curve E over $\mathbb{C}$. Then $E(\mathbb{C})=\mathbb{C}/\Lambda$ for some lattice Λ, and $E(\mathbb{C})_{\text{tors}}=\mathbb{Q}\Lambda/\Lambda$ where

$$\mathbb{Q}\Lambda=\{r\lambda\mid r\in\mathbb{Q},\lambda\in\Lambda\}=\{z\in\mathbb{C}\mid mz\in\Lambda\text{ some }m\in\mathbb{Z}\}=\mathbb{Q}\otimes\Lambda.$$

The degree of an endomorphism α of E is the order its kernel in $E(\mathbb{C})_{\text{tors}}$, and so we find that $\deg(\alpha)$ is the determinant of α acting on Λ. We shall need a generalization of this to other fields.

Let E be an elliptic curve over an algebraically closed field k, and let ℓ be a prime not equal to the characteristic of k. Then $E(k)_{\ell^n}\approx(\mathbb{Z}/\ell^n\mathbb{Z})^2$. The *Tate module* $T_\ell E$ of E is defined to be

$$T_\ell E=\varprojlim E(k)_{\ell^n}.$$

Thus, it is a free $\mathbb{Z}_\ell$-module of rank 2 such that $T_\ell E/\ell^n T_\ell E=E(k)_{\ell^n}$ for all n. For example, if $k=\mathbb{C}$ and $E(\mathbb{C})=\mathbb{C}/\Lambda$, then

$$E(\mathbb{C})_{\ell^n}=\frac{1}{\ell^n}\Lambda/\Lambda=\Lambda/\ell^n\Lambda=\Lambda\otimes(\mathbb{Z}/\ell^n\mathbb{Z}),$$

and so

$$T_\ell E=\Lambda\otimes\mathbb{Z}_\ell.$$

More canonically,

$$T_\ell E=H_1(E,\mathbb{Z})\otimes\mathbb{Z}_\ell.$$

Proposition 29.6. *Let E and ℓ be as above. For an endomorphism α of E,*

$$\det(\alpha|T_\ell E)=\deg\alpha.$$

Proof. When $k=\mathbb{C}$, then the statement follows from the above discussion. For k of characteristic zero, it follows from the case $k=\mathbb{C}$. For k of characteristic $p\neq 0$, see [S1]. $\square$

When Λ is a free module over some ring R and $\alpha:\Lambda\to\Lambda$ is R-linear, $\operatorname{Tr}(\alpha|\Lambda)$ denotes the trace (sum of diagonal terms) of the matrix of α relative to some basis for Λ—it is independent of the choice of basis.

Corollary 29.7. *Let E be an elliptic curve over $\mathbb{F}_p$. Then the trace of*

$$\operatorname{Tr}(\varphi_p|T_\ell E)=a_p=_{df}p+1-N_p.$$

Proof. For any 2×2 matrix A, $\det(A - I_2) = \det A - \operatorname{Tr} A + 1$. On applying this to the matrix of φ_p acting on $T_\ell E$, and using the proposition, we find that

$$\deg(\varphi_p - 1) = \deg(\varphi_p) - \operatorname{Tr}(\varphi_p | T_\ell E) + 1.$$

As we noted in Section 19, $\deg(\varphi_p - 1) = N_p$ and $\deg(\varphi_p) = p$. □

As we noted above, a correspondence $T : X \vdash X$ defines a map $J(X) \to J(X)$. When E is an elliptic curve, $E(k) = J(E)$, and so T acts on $E(k)$, and hence also on $T_\ell(E)$.

Corollary 29.8. *Let E be an elliptic curve over $\mathbb{F}_p$. Then*

$$\operatorname{Tr}(\varphi_p^{\mathrm{tr}} | T_\ell E) = \operatorname{Tr}(\varphi_p | T_\ell E).$$

Proof. Because φ_p has degree p, $\varphi_p \circ \varphi_p^{\mathrm{tr}} = p$. Therefore, if α, β are the eigenvalues of φ_p, so that in particular $\alpha\beta = \deg \varphi = p$, then

$$\operatorname{Tr}(\varphi_p^{\mathrm{tr}} | T_\ell E) = p/\alpha + p/\beta = \beta + \alpha.$$

□

The action of the Hecke operators on $H_1(E, \mathbb{Z})$. Again, we first need an elementary result from linear algebra.

Let V be a real vector space and suppose that we are given the structure of a complex vector space on V. This means that we are given an $\mathbb{R}$-linear map $J : V \to V$ such that $J^2 = -1$. The map J extends by linearity to $V \otimes_{\mathbb{R}} \mathbb{C}$, and $V \otimes_{\mathbb{R}} \mathbb{C}$ splits as a direct sum

$$V \otimes_{\mathbb{R}} \mathbb{C} = \mathbb{V}^+ \oplus \mathbb{V}^-,$$

with $V^\pm$ the ± 1 eigenspaces of J.

Proposition 29.9. *(a) The map*

$$V \xrightarrow{v \mapsto v \otimes 1} V \otimes_{\mathbb{R}} \mathbb{C} \xrightarrow{\mathit{project}} \mathbb{V}^+$$

is an isomorphism of complex *vector spaces.*

(b) Denote by $w \mapsto \bar{w}$ the map $v \otimes z \mapsto v \otimes \bar{z} : V \otimes_{\mathbb{R}} \mathbb{C} \to V \otimes_{\mathbb{R}} \mathbb{C}$; this is an $\mathbb{R}$-linear involution of $V \otimes_{\mathbb{R}} \mathbb{C}$ interchanging V^+ and V^-.

Proof. Easy exercise. □

Corollary 29.10. *Let α be an endomorphism of V which is $\mathbb{C}$-linear. Write A for the matrix of α regarded as an $\mathbb{R}$-linear endomorphism of V, and A_1 for the matrix of α as a $\mathbb{C}$-linear endomorphism of V. Then*

$$A \sim A_1 \oplus \bar{A}_1.$$

(By this I mean that the matrix A is equivalent to the matrix $\begin{pmatrix} A_1 & 0 \\ 0 & \bar{A}_1 \end{pmatrix}$.)

Proof. Follows immediately from the above Proposition. [In the case that V has dimension 2, we can identify V (as a real or complex vector space) with $\mathbb{C}$. For the map "multiplication by $\alpha = a + ib$" the statement becomes,

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \sim \begin{pmatrix} a + ib & 0 \\ 0 & a - ib \end{pmatrix},$$

which is obviously true because the two matrices are semisimple and have the same trace and determinant.] □

Corollary 29.11. *For any $p \nmid N$,*

$$\mathrm{Tr}(T(p) \mid H_1(X_0(N), \mathbb{Z})) = \mathrm{Tr}(T(p) \mid \Omega^1(X_0(N))) + \overline{\mathrm{Tr}(T(p) \mid \Omega^1(X_0(N)))}.$$

Proof. To say that $H_1(X_0(N), \mathbb{Z})$ is a lattice in $\Omega^1(X_0(N))^\vee$ means that

$$H_1(X_0(N), \mathbb{Z}) \otimes_{\mathbb{Z}} \mathbb{R} = \Omega^1(X_0(N))^\vee$$

(as real vector spaces). Clearly

$$\mathrm{Tr}(T(p) \mid H_1(X_0(N), \mathbb{Z})) = \mathrm{Tr}(T(p) \mid H_1(X_0(N), \mathbb{Z}) \otimes_{\mathbb{Z}} \mathbb{R}),$$

and so we can apply the preceding corollary. □

The proof that $c(p) = a_p$.

Theorem 29.12. *Consider an $f = \sum c(n)q^n$ and a map $X_0(N) \to E$, as in (28.3). For all $p \nmid N$,*

$$c(p) = a_p =_{df} p + 1 - N_p(E).$$

Proof. We assume first that $X_0(N)$ has genus 1, and so we may take the map to be an isomorphism: $E = X_0(N)$. Let p be a prime not dividing N. Then E has good reduction at p, and for any $\ell \neq p$, the reduction map $T_\ell E \to T_\ell \widetilde{E}$ is an isomorphism. The Eichler-Shimura relation states that

$$\widetilde{T}(p) = \varphi_p + \varphi_p^{\mathrm{tr}}.$$

On taking traces on $T_\ell \widetilde{E}$, we find (using 29.7, 29.8, 29.11) that

$$2c(p) = a_p + a_p.$$

The proof of the general case is very similar except that, at various places in the argument, an elliptic curve has to be replace either by a curve or the Jacobian variety of a curve. Ultimately, one uses that $T_\ell E$ is the largest quotient of $T_\ell Jac(X_0(N))$ on which $T(p)$ acts as multiplication by $c(p)$ for all $p \nmid N$ (perhaps after tensoring with $\mathbb{Q}_\ell$). □

Aside 29.13. Let X be a Riemann surface. The map $[P] - [P_0] \mapsto \int_{P_0}^{P} \omega$ extends by linearity to map $\mathrm{Div}^0(X) \to Jac(X)$. The famous theorem of Abel-Jacobi says that this induces an isomorphism $J(X) \to \mathrm{Jac}(X)$. The Jacobian variety $\mathrm{Jac}(X)$ of a curve X over a field k (constructed in general by Weil) has the property that $\mathrm{Jac}(X)(k) = J(X)$, at least when $J(k) \neq \emptyset$. For more on Jacobian and Abelian varieties, see my articles in "Arithmetic Geometry" (Eds. Cornell, G., and Silverman, J.).

Reference: The best reference for the material in Sections 23–29 is Knapp's book.

30. Wiles's Proof

> *Somebody with an average or even good mathematical background might feel that all he ends up with after reading [...]'s paper is what he suspected before anyway: The proof of Fermat's Last Theorem is indeed very complicated. (M. Flach)*

In this section, I explain the strategy of Wiles's proof of the Taniyama conjecture for semistable elliptic curves over $\mathbb{Q}$ (i.e., curves with at worst nodal reduction).

Recall, that if S denotes the sphere, then $\pi =_{df} \pi_1(S \setminus \{P_1, \ldots, P_s\}, O)$ is generated by loops $\gamma_1, \ldots, \gamma_s$ around each of the points $P_1, \ldots, P_s$, and that π classifies the coverings of S unramified except over $P_1, \ldots, P_s$.

Something similar is true for $\mathbb{Q}$. Let K be a finite extension of $\mathbb{Q}$, and let $\mathcal{O}_K$ be the ring of integers in K. In $\mathcal{O}_K$, the ideal $p\mathcal{O}_K$ factors into a product of powers of prime ideals: $p\mathcal{O}_K = \prod \mathfrak{p}^{e_\mathfrak{p}}$. The prime p is said to be *unramified* in K if no $e_\mathfrak{p} > 1$.

Now assume $K/\mathbb{Q}$ is Galois with Galois group G. Let p be unramified in K and choose a prime ideal $\mathfrak{p}$ dividing $p\mathcal{O}_K$ (so that $\mathfrak{p} \cap \mathbb{Z} = (p)$). Let $G(\mathfrak{p})$ be the subgroup of G of σ such that $\sigma\mathfrak{p} = \mathfrak{p}$. One shows that the action of $G(\mathfrak{p})$ on $\mathcal{O}_K/\mathfrak{p} = k(\mathfrak{p})$ defines an isomorphism $G(\mathfrak{p}) \to \mathrm{Gal}(k(\mathfrak{p})/\mathbb{F}_p)$. The element $F_\mathfrak{p} \in G(\mathfrak{p}) \subset G$ mapping to the Frobenius element $x \mapsto x^p$ in $\mathrm{Gal}(k(\mathfrak{p})/\mathbb{F}_p)$ is called the *Frobenius element* at $\mathfrak{p}$. Thus $F_\mathfrak{p} \in G$ is characterised by the conditions:

$$\begin{cases} F_\mathfrak{p}\mathfrak{p} &= \mathfrak{p}, \\ F_\mathfrak{p} x &\equiv x^p \mod \mathfrak{p}, \text{ for all } x \in \mathcal{O}_K. \end{cases}$$

If $\mathfrak{p}'$ also divides $p\mathcal{O}_K$, then there exists a $\sigma \in G$ such that $\sigma\mathfrak{p} = \mathfrak{p}'$, and so $F_{\mathfrak{p}'} = \sigma F_\mathfrak{p} \sigma^{-1}$. Therefore, the conjugacy class of $F_\mathfrak{p}$ depends on p—I'll often write F_p for any one of the $F_\mathfrak{p}$. It is known that the $F_\mathfrak{p}$ (varying p) generate G.

The above discussion extends to infinite extensions. Fix a finite nonempty set S of prime numbers, and let K_S be the union of all $K \subset \mathbb{C}$ that are of finite degree over $\mathbb{Q}$ and unramified outside S—it is an infinite Galois extension of $\mathbb{Q}$. For each $p \in S$, there is an element $F_p \in \mathrm{Gal}(K_S/\mathbb{Q})$, well-defined up to conjugation, called the *Frobenius element* at p.

Proposition 30.1. *Let E be an elliptic curve over $\mathbb{Q}$. Let ℓ be a prime, and let*

$$S = \{p \mid E \text{ has bad reduction at } p\} \cup \{\ell\}.$$

Then all points of order ℓ^n on E have coordinates in K_S, i.e., $E(K_S)_{\ell^n} = E(\mathbb{Q}^{al})_{\ell^n}$ for all n.

Proof. See [S1, VII.4.1]. □

Example 30.2. The smallest field containing the coordinates of the points of order 2 on the curve $E : Y^2Z = X^3 + aXZ^2 + bZ^3$ is the splitting field of $X^3 + aX + b$. Those who know a little algebraic number theory will recognize that this field is unramified at the primes not dividing the discriminant Δ of $X^3 + aX + b$, i.e., at the primes where E has good reduction (ignoring 2)

The Galois group G_S acts on $E(K_S)_{\ell^n}$ for all n. Recall from p156 that $T_\ell E$ is the free $\mathbb{Z}_\ell$-module of rank 2 such that

$$T_\ell E/\ell^n T_\ell E = E(K_S)_{\ell^n} = E(\mathbb{Q}^{\mathrm{al}})_{\ell^n}$$

for all n. The action of G_S on the quotients defines a continuous action of G_S on $T_\ell E$, i.e., a continuous homomorphism (also referred to as a representation)

$$\rho_\ell : G_S \to \mathrm{Aut}(T_\ell E) \approx \mathrm{GL}_2(\mathbb{Z}_\ell).$$

Proposition 30.3. *Let E, ℓ, S be as in the previous proposition. For all $p \notin S$,*

$$\mathrm{Tr}(\rho_\ell(F_p) \mid T_\ell E) = a_p =_{df} p + 1 - N_p(E).$$

Proof. Because $p \notin S$, E has good reduction to an elliptic curve E_p over $\mathbb{F}_p$, and the reduction map $P \mapsto \bar{P}$ induces an isomorphism $T_\ell E \to T_\ell E_p$. [For an elliptic curve E over a nonalgebraically closed field k, $T_\ell E = \varprojlim E(k^{\mathrm{al}})_{\ell^n}$.] By definition F_p maps to the Frobenius element in $\mathrm{Gal}(\mathbb{F}/\mathbb{F}_p)$, and the two have the same action on $T_\ell E$. Therefore the proposition follows from (29.7). □

Definition 30.4. A continuous homomorphism $\rho : G_S \to \mathrm{GL}_2(\mathbb{Z}_\ell)$ is said to be *modular* if $\mathrm{Tr}(\rho(F_p)) \in \mathbb{Z}$ for all $p \notin S$ and there exists a cusp form $f = \sum c(n)q^n$ in $\mathcal{S}_{2k}(\Gamma_0(N))$ for some k and N such that

$$\mathrm{Tr}(\rho(F_p)) = c(p)$$

for all $p \notin S$.

Thus, in order to prove that E is modular one must prove that $\rho_\ell : G_S \to \mathrm{Aut}(T_\ell E)$ is modular for some ℓ. Note that then ρ_ℓ will be modular for all ℓ.

Similarly, one says that a continuous homomorphism $\rho : G_S \to \mathrm{GL}_2(\mathbb{F}_\ell)$ is modular if there exists a cusp form $f = \sum c(n)q^n$ in $\mathcal{S}_{2k}(\Gamma_0(N))$ for some k and N such that

$$\mathrm{Tr}(\rho(F_p)) \equiv c(p) \mod \ell$$

for all $p \notin S$. There is the following remarkable conjecture.

Conjecture 30.5 (Serre). *Every odd irreducible representation $\rho : G_S \to \mathrm{GL}_2(\mathbb{F}_\ell)$ is modular.*

"Odd" means that $\det \rho(c) = -1$, where c is complex conjugation. "Irreducible" means that there is no one-dimensional subspace of $\mathbb{F}_\ell^2$ stable under the action of G_S. The Weil pairing [S1,III.8] shows that $\bigwedge^2 E_\ell = \mu_\ell$ (the group of ℓ-roots of 1 in $\mathbb{Q}^{\mathrm{al}}$). Since $c\zeta = \zeta^{-1}$, this shows that the representation of G_S on E_ℓ is odd. It need not be irreducible, for example, if E has a point of order ℓ with coordinates in $\mathbb{Q}$.

As we shall discuss in the next section, Serre in fact gave a recipe for defining the level N and weight $2k$ of modular form.

There is much numerical evidence supporting Serre's conjecture, but few theorems. The most important of these is the following.

Theorem 30.6 (Langlands, Tunnell). *If $\rho : G_S \to \mathrm{GL}_2(\mathbb{F}_3)$ is odd and irreducible, then it is modular.*

Note that $\mathrm{GL}_2(\mathbb{F}_3)$ has order $8 \cdot 6 = 48$. The action of $\mathrm{PGL}_2(\mathbb{F}_3)$ on the projective plane over $\mathbb{F}_3$ identifies it with S_4, and so $\mathrm{GL}_2(\mathbb{F}_3)$ is a double cover $\widetilde{S}_4$ of S_4.

The theorem of Langlands and Tunnell in fact concerned representations $G_S \to \mathrm{GL}_2(\mathbb{C})$. In the last century, Klein classified the finite subgroups of $\mathrm{GL}_2(\mathbb{C})$: their images in $\mathrm{PGL}_2(\mathbb{C})$ are cyclic, dihedral, A_4, S_4, or A_5. Langlands constructed candidates for the modular forms, and verified they had the correct property in the A_4 case. Tunnell verified this in the S_4 case, and, since $\mathrm{GL}_2(\mathbb{F}_3)$ embeds into $\mathrm{GL}_2(\mathbb{C})$, this verifies Serre's conjecture for $\mathbb{F}_3$.

Fix a representation $\rho_0 : G_S \to \mathrm{GL}_2(\mathbb{F}_\ell)$. In future, R will always denote a complete local Noetherian ring with residue field $\mathbb{F}_\ell$, for example, $\mathbb{F}_\ell$, $\mathbb{Z}_\ell$, or $\mathbb{Z}_\ell[[X]]$. Two homomorphism $\rho_1, \rho_2 : G_S \to \mathrm{GL}_2(R)$ will be said to be *strictly equivalent* if

$$\rho_1 = M\rho_2 M^{-1}, \quad M \in \mathrm{Ker}(\mathrm{GL}_2(R) \to \mathrm{GL}_2(k)).$$

A *deformation* of ρ_0 is a strict equivalence class of homomorphisms $\rho : G_S \to \mathrm{GL}_2(R)$ whose composite with $\mathrm{GL}_2(R) \to \mathrm{GL}_2(\mathbb{F}_p)$ is ρ_0.

Let $*$ be a set of conditions on representations $\rho : G_S \to \mathrm{GL}(R)$. Mazur showed, for certain $*$, that there is a universal $*$-deformation of ρ_0, i.e., a ring $\widetilde{R}$ and a deformation $\widetilde{\rho} : G_S \to \mathrm{GL}_2(\widetilde{R})$ satisfying $*$ such that for any other deformation $\rho : G_S \to \mathrm{GL}_2(R)$, there is a unique homomorphism $\widetilde{R} \to R$ such that the composite $G_S \xrightarrow{\widetilde{\rho}} \mathrm{GL}_2(\widetilde{R}) \to \mathrm{GL}_2(R)$ is ρ.

Now assume ρ_0 is modular. Work of Hida and others show that, for certain $*$, there exists a deformation $\rho_{\mathbb{T}} : G_S \to \mathrm{GL}_2(\mathbb{T})$ that is universal for *modular* deformations satisfying $*$.

Because $\widetilde{\rho}$ is universal for *all* $*$-representations, there exists a unique homomorphism $\delta : \widetilde{R} \to \mathbb{T}$ carrying $\widetilde{\rho}$ into $\rho_{\mathbb{T}}$. It is onto, and it is injective if and only if *every* $*$-representation is modular.

It is now possible to explain Wiles's strategy. First, state conditions $*$ as strong as possible but which are satisfied by the representation of G_S on $T_\ell E$ for E a semistable elliptic curve over $\mathbb{Q}$. Fixing a modular ρ_0 we get a homomorphism $\delta : \widetilde{R} \to \mathbb{T}$.

Theorem 30.7 (Wiles). *The homomorphism $\delta : \widetilde{R} \to \mathbb{T}$ is an isomorphism (and so every $*$-representation lifting ρ_0 is modular).*

Now let E be an elliptic curve over $\mathbb{Q}$, and assume initially that the representation of G_S on E_3 is irreducible. By the Theorem of Langlands and Tunnell, the representation $\rho_0 : G_S \to \mathrm{Aut}(E(K_S)_3)$ is modular, and by Wiles's theorem, *every* $*$-representation is modular. In particular, $\rho_3 : G_S \to \mathrm{Aut}(T_3 E)$ is modular, which implies that E is modular.

What if the representation of G_S on $E(K_S)_3$ is not irreducible, for example, if $E(\mathbb{Q})$ contains a point of order three. It is not hard to show that the representations of G_S on $E(K_S)_3$ and $E(K_S)_5$ can't both be reducible, because otherwise either E or a curve isogenous to E will have rational points of order 3 and 5, hence a point 15, which is impossible. Unfortunately, there is no Langlands-Tunnell theorem for 5. Instead, Wiles uses the following elegant argument.

He shows that there is a semistable elliptic curve E' over $\mathbb{Q}$ such that:

(a) $E'(K_S)_3$ is irreducible;
(b) $E'(K_S)_5 \approx E(K_S)_5$ as G_S-modules.

Because of (a), the preceding argument applies to E' and shows it to be modular. Hence the representation $\rho_5 : G_S \to \mathrm{Aut}(T_5 E')$ is modular, and so also is $\rho_0 : G_S \to \mathrm{Aut}(E'(K_S)_5) \approx \mathrm{Aut}(E(K_S)_5)$. Now, Wiles can apply his original argument with 3 replaced by 5.

31. Fermat, At Last

Fix a prime number ℓ, and let E be an elliptic curve over $\mathbb{Q}$. For a prime p it is possible to decide whether or not E has good reduction at p purely by considering the action of $G = \mathrm{Gal}(\mathbb{Q}^{\mathrm{al}}/\mathbb{Q})$ on the modules $E(\mathbb{Q}^{\mathrm{al}})_{\ell^n}$, for all $n \geq 1$.

Let M be a finite abelian group, and let $\rho : G \to \mathrm{Aut}(M)$ be a continuous homomorphism (discrete topology on $\mathrm{Aut}(M)$). The kernel H of ρ is an open subgroup of G, and therefore its fixed field $\mathbb{Q}^{\mathrm{al}H}$ is a finite extension of $\mathbb{Q}$. We say that ρ is *unramified* at p if p is unramified in $\mathbb{Q}^{\mathrm{al}H}$. With this terminology, we can now state a converse to Proposition 30.1.

Theorem 31.1. *Let ℓ be a prime. The elliptic curve E has good reduction at p if and only if the representation of G on $E(\mathbb{Q}^{al})_{\ell^n}$ is unramified for all n.*

The proof makes use of the theory of Néron models.

There is a similar criterion for $p = \ell$.

Theorem 31.2. *Let ℓ be a prime. The elliptic curve E has good reduction at ℓ if and only if the representation of G on E_{ℓ^n} is flat for all n.*

For the experts, the representation of G on $E(\mathbb{Q}^{\mathrm{al}})_{\ell^n}$ is flat if there is a finite flat group scheme H over $\mathbb{Z}_\ell$ such that $H(\mathbb{Q}_\ell^{\mathrm{al}}) \approx E(\mathbb{Q}_\ell^{\mathrm{al}})_{\ell^n}$ as G-modules. Some authors say “finite” or “crystalline” instead of flat.

These criteria show that it is possible to detect whether E has bad reduction at p, and hence whether p divides the conductor of E, from knowing how G acts on $E(\mathbb{Q}^{\mathrm{al}})_{\ell^n}$ for *all* n—it may not be possible to detect bad reduction simply by looking at $E(\mathbb{Q}^{\mathrm{al}})_\ell$ for example.

Recall that Serre conjectured that every odd irreducible representation $\rho : G \to \mathrm{GL}_2(\mathbb{F}_\ell)$ is modular, i.e., that there exists an $f = \sum c(n)q^n \in \mathcal{S}_{2k}(\Gamma_0(N))$, some k and N, such that

$$\mathrm{Tr}(\rho(F_p)) = c(n) \mod \ell$$

whenever ρ is unramified at p.

Conjecture 31.3 (Refined Serre conjecture). *Every odd irreducible representation $\rho : G \to \mathrm{GL}_2(\mathbb{F}_\ell)$ is modular for a specific k and N. For example, a prime $p \neq \ell$ divides N if and only if ρ is ramified at p, and ℓ divides N if and only if ρ is not flat.*

Theorem 31.4 (Ribet and others). *If $\rho : G \to \mathrm{GL}_2(\mathbb{F}_\ell)$ is modular, then it is possible to choose the cusp form to have the weight $2k$ and level N predicted by Serre.*

This proof is difficult.

Now let E be the curve defined in (26.22) corresponding to a solution to $X^\ell + Y^\ell = Z^\ell$, $\ell > 3$. It is not hard to verify, using nontrivial facts about elliptic curves, that the representation ρ_0 of G on $E(\mathbb{Q}^{\mathrm{al}})_\ell$ is irreducible. Moreover, that it unramified for $p \neq 2, \ell$, and that it is flat for $p = \ell$. The last statement follows from the facts that E has at worst nodal reduction at p, and if it does have bad reduction at p, then $p^\ell | \Delta$.

Now

$$E \text{ modular } \implies \rho_0 \text{ modular } \overset{\text{Ribet}}{\implies} \rho_0 \text{ modular for a cusp form of weight 2, level 2.}$$

But $X_0(N)$ has genus 0, and so there is no such cusp form. Wiles's theorem proves that E doesn't exist.

Of the growing number of sources attempting to explain Wiles's theorem, I'll cite just three.

Ribet: Bull AMS, 32.4, 375–402. This is reliable, easy to read, and contains a great list of references.

Murty, Kumar (Ed.). Seminar on Fermat's last theorem, Canadian Math. Soc.. This was mostly written before Wiles found the correct proof, but nevertheless gives much of the background required for the proof.

Darmon, Diamond, Taylor. Fermat's Last Theorem. Contains the most thorough introduction to the proof. A preliminary version was published in: Current Developments in Mathematics, 1995.

Bibliography

[**C1**] Cassels, J.W.S., Diophantine equations with special reference to elliptic curves, J. London Math. Soc. 41 (1966), 193–291.

This survey article was the first modern account of the arithmetic theory of elliptic curves.

[**C2**] Cassels, J.W.S., Lectures on Elliptic Curves, LMS, Student Texts 24, 1991.

Gives a concise elementary treatment of the basics.

[**Cr**] Cremona, J.E., Algorithms for Modular Curves, Cambridge, 1992.

How to compute almost everything of interest connected with elliptic curves, together with big tables of results.

[**F**] Fulton, W., Algebraic Curves, Benjamin, 1969.

Contains the background from algebraic geometry needed for elliptic curves.

[**H**] Husemoller, D., Elliptic Curves, Springer, 1987.

The beginning is quite elementary, but it becomes rapidly more advanced (and sketchy).

[**Kn**] Knapp, A.W., Elliptic Curves.

The first five chapters give a very readable and elementary account of the basics on elliptic curves. The remaining chapters study the L-series of a curve, and explain the relation to modular forms—this is more difficult, but is very important, for example, in the work of Wiles.

[**K1**] Koblitz, N., Introduction to Elliptic Curves and Modular Forms, Springer, 1984.

More modular forms than elliptic curves, but it explains the relation to the classical problem of "congruent numbers".

[**K2**] Koblitz, N., A Course in Number Theory and Cryptography, Springer, 2nd edn, 1987.

The last chapter explains how elliptic curves are used to give an algorithm for factorizing integers that has advantages over all others.

[**S1**] Silverman, J.H., The Arithmetic of Elliptic Curves, Springer, 1986.

This well-written book and its sequel(s) are the basic references for the subject.

[**S2**] Silverman, J.H., Advanced Topics in the Arithmetic of Elliptic Curves, Springer, 1994.

[**ST**] Silverman, J.H., and Tate, J., Rational Points on Elliptic Curves, Springer, 1992.

The first half of the book is a slight revision of the notes from Tate's famous 1961 Haverford lectures, which give a very elementary introduction to the subject.

[T] Tate, J., The arithmetic of elliptic curves, Inv. Math. 23 (1974), 179–206.

The notes of Tate's talks at the 1972 summer meeting of the AMS. They are an excellent survey of what was known and conjectured at the time. (Tate's Haverford lectures, his course in fall 1967 at Harvard, and this article have strongly influenced subsequent accounts.)

THE END