Chapter 1

Introduction

Definition 1.1 *An* elliptic curve $\mathcal{E}$ *over a field* k *of characterstic* $\neq 2$ *is defined by an equation*

$$y^2 = x^3 + ax^2 + bx + c,$$

where the cubic on the right has distinct roots.

Remarks:

1. There are several ways of defining elliptic curves. We have chosen the definition above because it is the most concrete, and requires no further explanation.

2. An alternative definition is that an elliptic curve over a field k is *a non-singular cubic curve* over k containing at least one point defined over k.

   By a cubic curve we mean a curve defined by a cubic polynomial

   $$ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + fxy + gy^2 + hx + iy + j = 0.$$

   We will see in Chapter 2 exactly what is meant by *non-singular*; but informally it means that the curve does not cross itself like

   $$y^2 = x^3 + x,$$

   or have a cusp like

   $$y^2 = x^3.$$

   We shall see too that the curve

   $$y^2 = x^3 + ax^2 + bx + c$$

   is non-singular precisely when the cubic on the right is *separable*, ie has distinct roots.

3. The additive group on an elliptic curve is most naturally seen in this context; if P, Q, R are three points defined over k (ie with coordinates in k) on the cubic curve then $P + Q + R = 0$ if and only if P, Q, R are collinear.

   Note one subtle (and important) point about this definition: if P, Q are two points on the curve defined over k then the line PQ meets the curve in a third point *defined over* k. This follows from the fact that if two of the roots α, β of the cubic polynomial

   $$p(x) = Ax^3 + Bx^2 + Cx + D \qquad (A, B, C, D \in k)$$

   lie in k then so does the third root γ, since

   $$\alpha + \beta + \gamma = -B/A.$$

   It follows that the points defined over k form a group. Since a curve defined over k is also defined over any extension field $K \supset k$, there is a group $\mathcal{E}(K)$ defined for each such field.

   In particular, in the rational case $k = \mathbb{Q}$ which specially concerns us we can consider the groups over $\mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$, as well as over the p-adic fields $\mathbb{Q}_p$ which we shall introduce in Chapter **??**. Each of these groups tells us something about the elliptic curve we are studying.

4. We've skated over one difficulty; the line PQ may not meet the curve again. We have to pass from *affine* to *projective* geometry, in effect adding a line at infinity where PQ can meet the curve in this case. All this will be detailed in Chapter 2

5. There is an even more general definition. To every curve there corresponds a non-negative integer g, the *genus* of the curve. An elliptic curve over k is *a curve of genus 1 over* k containing at least one point defined over k.

   (The reason for adding the condition that the curve must contain a point over k is that the set of points defined over k form an abelian group, as we have said; and a group, by definition, must be non-empty.)

   Lines and conics are curves of genus 0. Such curves are said to be *rational*, since the points on the curve can be parametrised by rational functions, at least if k is algebraically closed. For example, the circle $x^2 + y^2 = 1$ can be parametrised by

   $$x = \frac{t^2 - 1}{t^2 + 1}, \; y = \frac{2t}{t^2 + 1}.$$

From this point of view, elliptic curves are the least complicated curves *after* the conics studied by the ancient Greeks.

Our earlier definitions of an elliptic curve were set in the plane; but this definition — an elliptic curve is a curve of genus 1 — extends to curves in any number of dimensions.

6. An elliptic curve defined by an equation

$$y^2 = x^3 + ax^2 + bx + c$$

is said to be in *Weierstrass normal form*, or just *normal form*.

If the characteristic of k is $\neq 2$ or 3, we can simplify this equation by the change of coordinate $x' = x + a/3$, making the coefficient of x^2 zero, ie bringing the equation to the form

$$y^2 = x^3 + bx + c.$$

We shall say that the curve in this case is in *Weierstrass reduced form*, or just *reduced form*.

7. Although we excluded fields k of characteristic 2 in our definition above, we *do* consider elliptic curves over such fields. But in this case we have to allow the equation to take the more general form

$$y^2 + c_1xy + c_3 = x^3 + c_2x^2 + c_4x + c_6.$$

(We shall see in due course the reason for this rather curious numbering of the coefficients. Note that there is no coefficient c_5.)

We shall say that the curve in this case is in *Weierstrass general form*.

Note that if the characteristic of k is not 2 then we can bring the equation above to standard form by 'completing the square' on the left:

$$(y + c_1x/2 + c_3/2)^2 = x^3 + (c_2 + c_1^2/4)x^2 + (c_4 + c_1c_3/2)x + (c_6 + c_3^2/4),$$

ie by the change of coordinate $y' = y + c_1x/2 + c_3/2$.

8. There is another way of looking at elliptic curves, through the theory of *doubly periodic functions $f(z)$ of a complex variable*. Although this does not lend itself to a definition, it was in fact the origin of the theory of elliptic curves, as well as the explanation for the use of the word 'elliptic'.

The familiar trigonometric functions $\cos x$, $\sin x$, $\tan x$, etc, are *singly periodic functions $f(x)$ of a real variable*:

$$f(x + 2\pi) = f(x).$$

By analogy, we say that $f(z)$ is doubly periodic, with periods ω_1, ω_2 (where $\omega_1/\omega_2 \notin \mathbb{R}$), if

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

It turns out (as we shall see in Chapter 8) that all such functions can be expressed in terms of one such function, *Weierstrass' elliptic function*

$$\wp(z) = \wp_{\omega_1,\omega_2}(z).$$

More precisely, if $f(z)$ is even then it is a rational function of $\wp(z)$:

$$f(z) = \frac{P(\wp(z)}{Q(\wp(z)}$$

where $P(w), Q(w)$ are polynomials.

As we shall see, $\wp(z)$ and its derivative $\wp'(z)$ satisfy an equation

$$\wp'(z)^2 = 4\wp(z)^3 + B\wp(z) + C.$$

(This is where the term *elliptic* comes from; because of this relation the function $\wp(z)$ can be used to compute integrals around an ellipse.)

We see from this equation that the points $(\wp(z), \wp'(z)/2)$ parametrise the elliptic curve

$$y^2 = x^3 + bx + c,$$

where $b = B/4,\ c = C/4$ — much as $(\cos t, \sin t)$ parametrises the circle $x^2 + y^2 = 1$. It turns out that every elliptic curve over $\mathbb{C}$ can be parametrized by a Weierstrass elliptic function in this way; and this provides a powerful analytical tool for studying elliptic curves.

1.1 The discriminant

Since our definition requires that the cubic polynomial

$$p(x) = x^3 + ax^2 + bx + c$$

on the right hand side of our equation should be *separable*, ie should have distinct roots, it is useful to establish a criterion for this.

Definition 1.2 *Suppose the polynomial*

$$f(x) = x^n + c_1 x^{n-1} + \cdots + c_n$$

has roots $\alpha_1, \ldots, \alpha_n$. *The* discriminant *of* f *is defined to be*

$$D(f) = \prod_{i<j} (\alpha_i - \alpha_j)^2.$$

Equivalently,

$$D(f) = (-1)^{n(n-1)/2} \prod_{i \neq j} (\alpha_i - \alpha_j),$$

where now each pair occurs twice, once as $\alpha_i - \alpha_j$ and once as $\alpha_j - \alpha_i$.

The following is an immediate consequence of the definition.

Proposition 1.1 *The polynomial* $f(x)$ *is separable (has distinct roots) if and only if*

$$D(f) \neq 0.$$

Since $D(f)$ is a *symmetric* function of the roots (ie any permutation of the roots leaves $D(f)$ unchanged) it is expressible as a polynomial in the coefficients of f:

$$D(f) = D(c_1, \ldots, c_n).$$

To determine this polynomial explicitly we start with the following result.

Proposition 1.2 *The polynomial* $f(x)$ *has a multiple root if and only if* $f(x)$ *and its derivative* $f'(x)$ *have a factor in common:*

$$f(x) \text{ separable} \iff \gcd(f, f') = 1.$$

Proof ► Suppose first that $f(x)$ has a multiple root, say

$$f(x) = (x - \alpha)^r g(x).$$

Then

$$f'(x) = (x - \alpha)^{r-1} \left(g(x) + (x - \alpha) g'(x) \right).$$

Thus if $r > 1$,

$$(x - \alpha) \mid \gcd(f(x), f'(x)).$$

Conversely, suppose this is so. If

$$f(x) = (x - \alpha) g(x)$$

then

$$f'(x) = g(x) + (x - \alpha) g'(x)$$

and so

$$\begin{aligned}(x - \alpha) \mid f'(x) &\implies (x - \alpha) \mid g(x) \\ &\implies (x - \alpha)^2 \mid f(x).\end{aligned}$$

◄

As this suggests, the discriminant of a polynomial is closely related to the *resultant* of two polynomials, which tells us if those polynomials have a root in common.

Definition 1.3 *Suppose the polynomials*

$$f(x) = x^m + a_1 x^{m-1} + \cdots + a_m, \quad g(x) = x^n + b_1 x^{n-1} + \cdots + b_n.$$

have roots

$$\alpha_1, \ldots, \alpha_m \text{ and } \beta_1, \ldots, \beta_n,$$

respectively. Then the resultant $R(f, g)$ *of* f *and* g *is defined to be*

$$R(f, g) = \prod_{1 \le i \le m, 1 \le j \le n} (\beta_j - \alpha_i).$$

The following result is immediate.

Proposition 1.3 *The polynomials* $f(x), g(x)$ have a root in common if and only if $R(f, g) = 0$.

Now

$$f(x) = (x - \alpha_1) \cdots (x - \alpha_m), \quad g(x) = (x - \beta_1) \cdots (x - \beta_n).$$

Thus

$$R(f, g) = g(\alpha_1) g(\alpha_2) \cdots g(\alpha_m).$$

Since the expression on the right is symmetric in $\alpha_1, \ldots, \alpha_m$, it follows that $R(f, g)$ can be expressed as a polynomial in the coefficients of f and g.

Proposition 1.4 *The resultant $R(f,g)$ can be expressed as an $(m+n)\times(m+n)$ determinant:*

$$R(f,g) = \det \begin{pmatrix} 1 & a_1 & a_2 & \dots & a_m & 0 & \dots & 0 \\ 0 & 1 & a_1 & \dots & a_{m-1} & a_m & \dots & 0 \\ & & & \dots & & & & \\ 0 & 0 & 0 & \dots & & \dots & a_{m-1} & a_m \\ 1 & b_1 & b_2 & \dots & b_n & 0 & \dots & 0 \\ 0 & 1 & b_1 & \dots & b_{n-1} & b_n & \dots & 0 \\ & & & \dots & & & & \\ 0 & 0 & 0 & \dots & & \dots & b_{n-1} & b_n \end{pmatrix}$$

Proof ► Let us denote this determinant by $S(f,g)$. Suppose $f(x)$ and $g(x)$ have a root, say t, in common. Consider the $m+n$ equations

$$\begin{aligned} t^{m-1}f(t) &= 0 \\ t^{m-2}f(t) &= 0 \\ &\dots \\ f(t) &= 0 \\ t^{n-1}g(t) &= 0 \\ t^{n-2}g(t) &= 0 \\ &\dots \\ g(t) &= 0 \end{aligned}$$

as linear equations in $t^{m+n-1}, t^{m+n-2}, \dots, 1$. The determinant of these linear equations is precisely $S(f,g)$. Thus $S(f,g) = 0$ if $f(x)$ and $g(x)$ have a root in common.

This will certainly be the case if any of the mn relations

$$\alpha_i - \beta_j = 0 \qquad (1 \le i \le m, 1 \le j \le n)$$

holds. It follows by the Remainder Theorem that each of these is a factor of $S(f,g)$; and so

$$R(f,g) \mid S(f,g).$$

But now if we express the coefficients of $f(x)$ and $g(x)$ in terms of the α's and β's we see that $R(f,g)$ and $S(f,g)$ are of the same degree in $\beta_1, \dots, \beta_n$; and on comparing the coefficients of $\beta_1^m \dots \beta_n^m$ in $R(f,g)$ and $S(f,g)$ we conclude that

$$R(f,g) = S(f,g).$$

◄

Let us apply this argument to the polynomials $f(x), f'(x)$. We have seen that $f(x)$ has a repeated root if $D(f) = 0$; and we have also seen that $f(x)$ has a repeated root if $R(f,f') = 0$. It is not surprising therefore to find that there is a relation between these entities.

Proposition 1.5 *If $f(x)$ is a monic polynomial then*

$$D(f) = (-1)^{n(n-1)/2} R(f,f').$$

Proof ► On differentiating

$$f(x) = \prod (x - \alpha_i)$$

and setting $x = \alpha_j$,

$$f'(\alpha_j) = \prod_{i \ne j} (\alpha_j - \alpha_i).$$

It follows that

$$\begin{aligned} R(f,f') &= \prod_j f'(\alpha_j) \\ &= \prod_{i \ne j} (\alpha_j - \alpha_i) \\ &= (-1)^{n(n-1)/2} \prod_{j<i} (\alpha_j - \alpha_i)^2 \\ &= (-1)^{n(n-1)/2} D(f). \end{aligned}$$

In other words,

$$D(f) = (-1)^{n(n-1)/2} R(f,f').$$

◄

Now we can apply this result to our cubic. First we consider the reduced case.

Proposition 1.6 *The discriminant of the polynomial*

$$f(x) = x^3 + bx + c$$

is

$$D(f) = -(4b^3 + 27c^2).$$

Proof ► We have

$$f'(x) = 3x^2 + b,$$

and so

$$\begin{aligned} D(f) &= -R(f, f') \\ &= -\det \begin{pmatrix} 1 & 0 & b & c & 0 \\ 0 & 1 & 0 & b & c \\ 3 & 0 & b & 0 & 0 \\ 0 & 3 & 0 & b & 0 \\ 0 & 0 & 3 & 0 & b \end{pmatrix} \\ &= -4b^3 - 27c^2. \end{aligned}$$

◄

It is probably a good idea to remember the discriminant in this reduced case, but not the more general case we turn to now.

Proposition 1.7 *The discriminant of the polynomial*

$$f(x) = x^3 + ax^2 + bx + c$$

is

$$D(f) = -4a^3c + 18abc - 4b^3 - 27c^2.$$

Proof ► We could determine this in the same way, by computing the determinant

$$D(f) = -\det \begin{pmatrix} 1 & a & b & c & 0 \\ 0 & 1 & a & b & c \\ 3 & 2a & b & 0 & 0 \\ 0 & 3 & 2a & b & 0 \\ 0 & 0 & 3 & 2a & b \end{pmatrix}.$$

Alternatively, it may be simpler to observe that $D(f)$ is left unaltered by the "change of origin" $x' = x + a/3$, since this leaves each factor $(\alpha_i - \alpha_j)$ unchanged. Thus we can derive the formula for $D(f)$ from the reduced case $a = 0$ by substituting $b - a^2/3$ for b and $c + 2a^3/27 - ab/3$ for c:

$$D(f) = -4\left(b - a^2/3\right)^3 - 27\left(c + 2a^3/27 - ab/3\right)^2.$$

In either case, the details are left to you! ◄

1.2 Weights

The transformation

$$x \mapsto d^2x, \ y \mapsto d^3y$$

leaves our equation in standard form, taking

$$y^2 = x^3 + ax^2 + bx + c$$

into

$$y^2 = x^3 + a'x^2 + b'x + c'$$

where

$$a' = d^2a, b' = d^4b, c' = d^6c.$$

We may say that the terms a, b, c have *weights* $2, 4, 6$ respectively. The various invariants we shall meet — in particular the *discriminant* defined above — are all *homogeneous*, ie consist of terms of the same weight. This offers a valuable check on the sometimes complicated formulae we shall encounter.

In particular, we see that the disciminant is of weight 12. So it could not contain, for example, a term a^2b, since that has weight 8.

Chapter 1

Introduction

A simple geometric construction allows us to add points on an elliptic curve — that is, a non-singular cubic curve. The resulting abelian group is the basis for the application of elliptic curves in cryptography, number theory and elsewhere.

Our aim in this Chapter is to explain informally — so for the moment we are not on oath! — how points are added, and why this operation is associative. Then in Chapter 3, when we have the tools of projective geometry at our disposal, we can set the theory on a rigorous footing.

1.1 The operation $*$

Let Γ be a cubic curve over the field k defined by a polynomial equation

$$f(x,y) = 0,$$

where $f(x,y)$ is a polynomial of degree 3 with coefficients in k, say

$$f(x,y) = a_1x^3 + a_2x^2y + a_3xy^2 + a_4y^3 + a_5x^2 + a_6xy + a_7y^2 + a_8x + a_9y + a_{10}.$$

Let $\Gamma(k)$ denote the set of points $P = (x,y) \in \Gamma$ defined over k, ie with coordinates $x,y \in k$.

Suppose $P, Q \in \Gamma(k)$. Let ℓ be the line PQ if $P \neq Q$, or the tangent at P if $P = Q$. Then ℓ meets Γ in a third point $R \in \Gamma(k)$.

For if ℓ is the line

$$y = mx + d$$

then

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

if $P \neq Q$; while

$$m = \frac{(\partial f/\partial x)_P}{(\partial f/\partial y)_P}$$

if $P = Q$. In either case,

$$m \in k;$$

and so also

$$d = y_1 - mx_1 \in k.$$

But PQ meets Γ where

$$u(x) = f(x, mx + d) = 0.$$

Now $u(x)$ is a cubic polynomial, say

$$u(x) = b_0x^3 + b_1x^2 + b_2x + b_3,$$

with coefficients $b_0, b_1, b_2, b_3 \in k$.

If the roots of this equation are x_1, x_2, x_3 then

$$x_1 + x_2 + x_3 = -\frac{b_1}{b_0} \in k.$$

Thus

$$x_1, x_2 \in k \Longrightarrow x_3 = -(x_1 + x_2 + b_1/b_0) \in k.$$

Since

$$y_3 = mx_3 + d \in k,$$

it follows that

$$R = (x_3, y_3) \in \Gamma(k),$$

as we claimed.

We set

$$R = P * Q.$$

Evidently this binary operation is commutative:

$$Q * P = P * Q. \qquad (*1)$$

Moreover, the relation between P, Q, R is symmetric:

$$R = P * Q \Longrightarrow P = Q * R \Longrightarrow Q = R * P.$$

In other words,

$$P * (P * Q) = Q. \qquad (*2)$$

It follows from this that

$$P * Q = P * R \Longleftrightarrow Q = R. \qquad (*3)$$

We have skated round two problems in the discussion above:

1. The line PQ may not meet the curve Γ again, since the coefficient of x^3 in the polynomial $u(x)$ may vanish, leaving a quadratic with the two solutions x_1, x_2.

   For example, consider the curve
$$x^2 = y^3 + 1.$$
   The points $P = (2, 3)$, $Q = (-2, 3)$ lie on this curve; but the line
$$y = 3$$
   joining them only meets the curve at these two points.

   As we shall see in Chapter 2, we can solve this problem completely by passing to the *projective plane* – in effect adding a 'line at infinity' to the affine plane k^2. Now every line PQ in the projective plane *does* meet the curve in three points, the third point perhaps being on the line at infinity.

2. More seriously, in the case $P = Q$ the tangent at P *may be undefined*. This happens if
$$\partial f/\partial x = \partial f/\partial y = 0$$
   at this point. Such a point is said to be *singular*.

   We have to restrict ourselves to *non-singular* curves, ie those without singular points. That is why we define an 'elliptic curve' as a *non-singular cubic curve*. This again will be dealt with in Chapter 2.

1.2 Addition

The operation $*$ is not associative. For if it were it would follow from $(*1)$ that if $S = P * P$ then
$$S * Q = (P * P) * Q = P * (P * Q) = Q$$
for all Q, which is absurd.

Remarkably though, if we choose *any* point $O \in \Gamma(k)$, and set
$$P + Q = O * (P * Q)$$
for $P, Q \in \Gamma(k)$ then the operation $+$ is not only commutative — that is obvious — but is also associative:
$$P + (Q + R) = (P + Q) + R$$

for all $P, Q, R \in \Gamma$. That is far from obvious.

It is clear however that O is a neutral (or zero) element with respect to this operation:
$$O + P = O * (O * P) = P,$$
by $(*1)$. Moreover, if we set
$$S = O * O$$
then the point
$$P' = S * P$$
is the additive inverse of P. For
$$P' * P = (S * P) * P = S$$
and so
$$P' + P = O * S = O * (O * O) = O.$$
Thus we may write
$$-P = S * P.$$

It follows that if the operation is associative then it defines an abelian group on $\Gamma(k)$.

It might seem surprising that we can choose *any* point $O \in \Gamma$ as the neutral (or zero) point. However, that is not really so. For if we have an abelian group structure on a set A then we take any element $a \in A$ and define a new abelian group structure on A by the operation
$$x \dagger y = x + y - a.$$
It is readily verified that this new operation is associative:
$$(x \dagger y) \dagger z = x + y + z - 2a = x \dagger (y \dagger z).$$
Moreover
$$x \dagger a = x + a - a = x,$$
so the element a is the new zero element; and if we set
$$x' = -x + 2a$$
then
$$x + x' = x + (-x) + 2a - a = a,$$
ie x' is the inverse of x with respect to the new operation.

In effect, all that we have done is to 'move the origin' from 0 to a, through the transformation
$$x \mapsto x - a.$$

1.3 The choice of O

Recall that we can choose *any* point $O \in \Gamma(k)$ as the zero point of our group. What is the best choice?

We saw that

$$-P = S * P,$$

where $S = O * O$ (ie S is the point where the tangent at O meets Γ again).

It turns out that life is much simpler if we can choose O so that $S = O$, ie

$$O * O = O.$$

That is, the tangent at O meets Γ in three coincident points: O, O, O. In other words, O is a *point of inflexion* (or *flex*) on Γ.

For then, as we have seen,

$$-P = O * P.$$

It also follows in this case that

$$P + Q + R = 0 \iff P, Q, R \text{ are collinear.}$$

For if P, Q, R are collinear then

$$\begin{aligned} R = P * Q &\implies O * R = O * (P * Q) \\ &\implies -R = P + Q \\ &\implies P + Q + R = 0. \end{aligned}$$

Conversely, if $P + Q + R = 0$ then

$$\begin{aligned} P + Q + R = 0 &\implies -R = P + Q \\ &\implies O * R = O * (P * Q) \\ &\implies R = P * Q \\ &\implies P, Q, R \text{ collinear} \end{aligned}$$

However, in general a cubic Γ over k *does not contain a point of inflexion over k*. In fact, Γ may contain *no* points defined over k at all — let alone points of inflexion — as for example the curve

$$\Gamma : x^3 + 2y^3 = 4$$

over $\mathbb{Q}$. For if $(x, y) \in \Gamma$, where $x, y \in \mathbb{Q}$, then we can write

$$x = \frac{X}{Z}, \; y = \frac{Y}{Z},$$

where $X, Y, Z \in \mathbb{Z}$ and $\gcd(X, Y, Z) = 1$; and now

$$X^3 + 2Y^3 = 4Z^3.$$

Evidently $2 \mid X$, say $X = 2X'$. Then

$$4X'^3 + Y^3 = 2Z^3.$$

It follows that $2 \mid Y$, say $Y = 2Y'$. But now

$$2X'^3 + 4Y'^3 = Z^3.$$

Hence $2 \mid Z$; and so $2 \mid X, Y, Z$, contradicting our assumption that $\gcd(X, Y, Z) = 1$.

On the other hand, we shall show in Chapter 3 that if the elliptic curve $\mathcal{E}$ *does* contain a point $P \in \mathcal{E}(k)$ then we can find a *birational transformation* over k taking $\mathcal{E}$ into another elliptic curve $\mathcal{E}'$ over k having a point of inflexion $O \in \mathcal{E}'(k)$. Moreover, this birational transformation preserves the group structure; so nothing is lost, from our point of view, in passing from $\mathcal{E}$ to $\mathcal{E}'$.

We may describe an elliptic curve with this property (having a point of inflexion O defined over the base field) as *Weierstrassian*, since in this case — as we shall see in Chapter 3 — the equation of the curve can be taken in a simple form, due to Weierstrass.

In the rest of the course we shall assume that *every elliptic curve $\mathcal{E}$ is Weierstrassian*, unless the contrary is stated.

1.4 Associativity

There are several ways of showing that our addition is associative. But since we defined addition geometrically, it is appropriate to give a geometric proof of associativity. For the moment, we merely sketch the proof; we shall fill in the details in Chapter 3.

We want to show that

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

ie

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

Since

$$O * X = O * Y \iff X = Y,$$

we can 'hive off' the last $O*$; it is sufficient to show that

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

There is an equivalent, more symmetric, form of this identity: for any 4 points $X, Y, Z, T \in \mathcal{E}$,

$$(X * Y) * (Z * T) = (X * Z) * (Y * T). \qquad (*4)$$

To see that this follows from the associative law, note first that it is sufficient to prove the result in any extension of the ground field k; so we may assume that k is algebraically closed. In that case we can certainly find a point of inflexion $O \in \mathcal{E}$; and on taking this as our zero point,

$$X * Y = O * (X + Y) = -(X + Y).$$

Thus

$$(X * Y) * (Z * T) = X + Y + Z + T = (X * Z) * (Y * T).$$

Conversely, suppose this result holds. On taking $X = O,\ Y = P*Q,\ Z = Q * R,\ T = Q$, we derive the required result:

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

It remains to prove the identity $(*4)$.

The general cubic curve Γ, as we saw, is defined by 10 coefficients:

$$\Gamma : a_1x^3 + a_2x^2y + a_3xy^2 + a_4y^3 + a_5x^2 + a_6xy + a_7y^2 + a_8x + a_9y + a_{10} = 0.$$

Suppose we are given 8 points $P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8$ in the plane, no three of which are collinear. Let us also suppose that there is an elliptic curve $\mathcal{E}$, ie a non-singular cubic, passing through these 8 points.

The cubic Γ passes through a given point P if the coefficients $(a_1, \ldots, a_{10})$ satisfy a certain homogeneous linear equation. Thus Γ will pass through the 8 points if the 10 coefficients satisfy 8 homogeneous linear equations.

Now we know from linear algebra that the solutions of m linear homogeneous equations in n unknowns form a vector space of dimension $\geq n - m$. Thus the cubics passing through our 8 points form a vector space of dimension $d \geq 2$.

Suppose first that $d > 2$. We shall show that this leads to a contradiction. For in this case we can impose 2 further homogeneous linear equations; in particular we can find a cubic Γ passing through any further two points Q, R.

Let us choose these two points on the line $\ell = P_1P_2$, say, then this line will meet Γ in 4 points, and so will lie wholly in Γ, which must therefore be degenerate:

$$\Gamma = \ell C,$$

where C is a conic.

But this conic C must pass through the 6 points $P_3, P_4, P_5, P_6, P_7, P_8$. Now a general conic is defined by 6 coefficients:

$$C : b_1x^2 + b_2xy + b_3y^2 + b_4x + b_5y + b_6 = 0.$$

It follows that we can always find a conic passing through 5 points Q_1, Q_2, Q_3, Q_4, Q_5.

In fact, if no 3 of these 5 points are collinear, then there is *exactly one such conic*. For if there were two we would have a pencil

$$C = \mu_1C_1 + \mu_2C_2;$$

and we could find a conic in this pencil passing through any further point R. But now if we choose R on $\ell = Q_1Q_2$, say, then the line ℓ meets C in 3 points, and so lies wholly in C. Thus C is degenerate:

$$C = \ell m,$$

and the line m must pass through Q_3, Q_4, Q_5, contrary to our assumption that these points were not collinear.

Let C be the conic determined by the points P_4, P_5, P_6, P_7, P_8. Then it follows from the argument above that this conic passes through P_3. But there was nothing special about our choice of P_1, P_2 out of the 8 points; we could equally well have chosen P_2, P_3 and P_1, P_3, in which case we would conclude that C passed through P_1 and P_2. It follows that all 8 points must lie on the conic C.

But a conic C and a cubic Γ meet in at most 6 points, unless the cubic is degenerate and contains the conic:

$$\Gamma = \ell C.$$

Thus all the cubics in our pencil must be degenerate. But that is impossible, since we supposed that there was a non-degenerate cubic (the elliptic curve $\mathcal{E}$) passing through the 8 points.

We have shown, therefore, that $d = 2$, ie the pencil of cubics through the 8 points takes the form

$$\Gamma = \lambda_1\Gamma_1 + \lambda_2\Gamma_2.$$

Now Γ_1 and Γ_2 meet in at most 9 points. For on eliminating y say from the equations for Γ_1 and Γ_2 we obtain a polynomial equation of degree 9 in x, to which the x-coefficients of $P_1, \dots, P_8$ provide 8 solutions. It follows that there is a 9th solution, giving a 9th common point P_9 on Γ_1 and Γ_2. (It also follows — although we make no use of this — that if $P_1, \dots, P_8 \in \Gamma(k)$ then $P_9 \in \Gamma(k)$, by the same argument we used to show that if $P, Q \in \Gamma(k)$ then $P * Q \in \Gamma(k)$.)

We have proved (more-or-less) the remarkable result that given any 8 points $P_1, \dots, P_8$ (no 3 of which are collinear) there exists a unique 9th point P_9 with the property that every cubic Γ through $P_1, \dots, P_8$ also passes through P_9.

To prove the associative law, we apply this result to the 8 points

$$X, Y, Z, T, X * Y, X * Z, Y * T, Z * T.$$

These points all lie on the elliptic curve $\mathcal{E}$, of course, and they also lie on 2 sets of 3 lines, as follows

	ℓ	m	n
ℓ'	X	Y	$X * Y$
m'	Z	T	$Z * T$
n'	$X * Z$	$Y * T$	

Now consider the 3 cubics

$$\mathcal{E}, \ \ell m n, \ \ell' m' n'.$$

Each of these passes through the 8 points, and so belongs to the pencil defined by those points. Hence

$$\mathcal{E} = \lambda \ell m n + \lambda' \ell' m' n'$$

for some $\lambda, \lambda' \in k$.

Moreover, $\mathcal{E}$ and $\ell m n$ meet in the further point

$$(X * Y) * (Z * T) \in \mathcal{E} \cap \ell m n;$$

while $\mathcal{E}$ and $\ell' m' n'$ meet in the further point

$$(X * Z) * (Y * T) \in \mathcal{E} \cap \ell' m' n';$$

It therefore follows from our argument above that

$$(X * Y) * (Z * T) = (X * Z) * (Y * T).$$

This establishes the identity $(*4)$, and so the associativity of our addition.

Chapter 2

From Affine to Projective Geometry

2.1 Projective spaces

One of the great discoveries of the Italian school of algebraic geometry around the turn of the century was that life becomes much easier if one "completes" the affine space k^n by adding "points at infinity" to form the projective space $\mathbb{P}^n(k)$.

Suppose V is a vector space over the field k. The associated *projective space* $\mathbb{P}V$ is the set of 1-dimensional subspaces of V. In other words, $\mathbb{P}V$ is the quotient-set

$$\mathbb{P}V = (V - \{0\}) / k^\times,$$

where $k^\times$ denotes the multiplicative group on the set $k - \{0\}$.

Thus each non-zero vector $v \in V$ defines a point of $\mathbb{P}V$; 2 non-zero vectors u, v defining the same point if they are scalar multiples of one another, ie

$$v = \rho u \qquad (\rho \in k^\times).$$

The *dimension* of $\mathbb{P}V$ is defined to be

$$\dim \mathbb{P}V = \dim V - 1.$$

Each r-dimensional vector subspace $U \subset V$ defines an $(r-1)$-dimensional projective subspace of $\mathbb{P}V$.

We define *n-dimensional projective space* $\mathbb{P}^n(k)$ over k to be

$$\mathbb{P}^n(k) = \mathbb{P}(k^{n+1}) = \left(k^{n+1} - \{0\}\right) / k^\times.$$

Each point of $\mathbb{P}^n(k)$ is represented by a set of $n+1$ *homogeneous coordinates*

$$[X_1, \ldots, X_n, X_{n+1}].$$

not all 0. Proportional coordinates define the same projective point:

$$\rho[X_1, \ldots, X_n, X_{n+1}] = [\rho X_1, \ldots, \rho X_n, \rho X_{n+1}] = [X_1, \ldots, X_n, X_{n+1}].$$

There is a natural embedding of the affine space k^n into the projective space $\mathbb{P}^n(k)$,

$$k^n \subset \mathbb{P}^n(k),$$

defined by the injective map

$$(x_1, \ldots, x_n) \mapsto [x_1, \ldots, x_n, 1].$$

The points of $\mathbb{P}^n(k)$ not in k^n, namely the points of the form

$$[X_1, \ldots, X_n, 0]$$

are called 'points at infinity'. They form an $(n-1)$-dimensional projective subspace of $\mathbb{P}^n(k)$.

2.2 The Projective Plane

We shall be mainly concerned with geometry in the projective plane

$$\mathbb{P}^2(k) = \{[X, Y, Z] : X, Y, Z \in k]\}.$$

We identify the affine plane k^2 with the subset $Z \neq 0$ of $\mathbb{P}^2(k)$, by the map

$$(x, y) \mapsto [x, y, 1] : \ k^2 \to \mathbb{P}^2(k).$$

The points of $\mathbb{P}^2(k)$ not in k^2 form the *line at infinity* $Z = 0$.

Each affine line

$$ax + by + c = 0$$

in the affine plane k^2 extends to the projective line

$$aX + bY + cZ = 0$$

in $\mathbb{P}^2(k)$, with the addition of a point $[-b, a, 0]$ at infinity.

In general each linear homogenous equation

$$aX + bY + cZ = 0$$

defines a line in the projective plane $\mathbb{P}^2(k)$. Each such line except for the line at infinity $Z = 0$ intersects the affine subspace $k^2 \subset \mathbb{P}^2(k)$ in an affine line.

Any 2 distinct projective lines

$$aX + bY + cZ = 0, \quad a'X + b'Y + c'Z = 0$$

intersect in a point; while any 2 distinct points in $\mathbb{P}^2(k)$ define a unique projective line. This perfect duality between points and lines (or in n dimensions, between points and $(n-1)$-dimensional subspaces) is a minor advantage of projective geometry.

Two affine lines are parallel if and only if the corresponding projective lines meet on the line at infinity.

2.3 The Projective Group

An invertible (non-singular) linear map

$$t : V \to V$$

induces a map

$$\bar{t} : \mathbb{P}V \to \mathbb{P}V,$$

where $\mathbb{P}V$ is the corresponding projective space. Such a map is called a *projective transformation.*

Two linear maps $t, \rho t$ $(\rho \in k^\times)$ define the same linear transformation. Thus the projective transformations form the *projective group*

$$PGL(V) = GL(V)/k^\times.$$

In particular

$$PGL(n, k) = GL(n+1, k)/k^\times.$$

If P_1, P_2, P_3, P_4 are 4 points in the projective plane, no 3 of which are collinear, and Q_1, Q_2, Q_3, P_4 is a second similar set, then there is a unique projective transformation sending

$$P_1 \mapsto Q_1, \ P_2 \mapsto Q_2, \ P_3 \mapsto Q_3, \ P_4 \mapsto Q_4.$$

For if we choose coordinates

$$P_i = [X_i, Y_i, Z_i] \quad (i = 1, 2, 3, 4)$$

then

$$[X_4, Y_4, Z_4] = a_1[X_1, Y_1, Z_1] + a_2[X_2, Y_2, Z_2] + a_3[X_3, Y_3, Z_3]$$

for some $a_i \in k$; and the a_i are all non-zero since no 3 of the points are collinear. But now we can take $a_i[X_i, Y_i, Z_i]$ to represent P_i; and then

$$P_4 = P_1 + P_2 + P_3.$$

Similarly we can choose coordinates to represent the second set with

$$Q_4 = Q_1 + Q_2 + Q_3.$$

Each point P can now be written in the form

$$P = \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3.$$

and the required projective transformation is then given by

$$P \mapsto Q = \lambda_1 Q_1 + \lambda_2 Q_2 + \lambda_3 Q_3.$$

In projective geometry, two curves — or other geometric entities — which can be mapped into one another by projective transformations are regarded as 'the same'.

2.4 Affine and Projective Varieties

An *affine variety* in k^n is defined by a set of simultaneous polynomial equations

$$P_1(x_1, \ldots, x_n) = 0, \ldots, P_r(x_1, \ldots, x_n) = 0.$$

(In general one is interested in the solutions of these equations not only in k, but also in its algebraic closure $\bar{k}$.) Algebraic geometry is the study of varieties.

We shall only be concerned with the simplest of varieties, namely curves in 2 dimensions defined by a single polynomial equation

$$F(x, y) = 0.$$

When we pass to projective space $\mathbb{P}^n(k)$ we deal exclusively with *homogeneous* polynomials $P(X_1, \ldots, X_n, X_{n+1})$, ie those with all terms of the same total degree, eg $X^2Y + XZ^2 + 2Y^3$. If $P(X_1, \ldots, X_n, X_{n+1})$ is a homogeneous polynomial of degree d then

$$P(\rho X_1, \ldots, \rho X_n, X_{n+1}) = \rho^d P(X_1, \ldots, X_n, X_{n+1}).$$

Thus it makes sense to speak of the points in projective space $\mathbb{P}^n(k)$ satisfying the equation $P(X_1, \ldots, X_n, X_{n+1})$ if (and only if) P is homogeneous.

If $p(x_1, \ldots, x_n)$ is a polynomial in k^n of degree d then the corresponding homogeneous polynomial is

$$P[X, Y, Z] = Z^d p(X/Z, Y/Z).$$

For example, the homogeneous form of the polynomial

$$p(x, y) = y^2 - x^3 - ax^2 - bx - c$$

of degree 3 is

$$P(X, Y, Z) = Y^2Z - X^3 - aX^2Z - bXZ^2 - cZ^3.$$

If effect we replace x and y by X and Y, and multiply each term by a power of Z to bring it up to degree d.

In this way, every affine variety V in k^n extends to a *projective variety* $\bar{V}$ in $\mathbb{P}^n(k)$, with

$$\bar{V} \cap k^n = V :$$

ie the restriction of the projective variety $\bar{V}$ to affine space is just the affine variety V. In general $\bar{V}$ will contain additional 'points at infinity'.

2.5 Tangents to a projective curve

Suppose γ is an affine curve in k^2 defined by the equation

$$f(x, y) = 0.$$

Let Γ be the corresponding projective curve in $\mathbb{P}^2(k)$, defined by the 'homogenised' equation

$$F(X, Y, Z) = 0,$$

where

$$F(x, y, 1) \equiv f(x, y).$$

We assert that the tangent to Γ at the point $P = [X_0, Y_0, Z_0]$ is the line

$$\frac{\partial F}{\partial X} X + \frac{\partial F}{\partial Y} Y + \frac{\partial F}{\partial Z} Z = 0,$$

where the partial differential coefficients are computed at the point $[X_0, Y_0, Z_0]$.

Let us verify that this is indeed the projective line corresponding to the usual tangent, if P is a point in the affine plane.

First note an important identity satisfied by the partial differential coefficients of a homogeous polynomial $F(x, y, z)$. If F is of degree d then

$$F(\rho X, \rho Y, \rho Z) = \rho^d F(X, Y, Z).$$

Differentiating with respect to ρ and setting $\rho = 1$,

$$\frac{\partial F}{\partial X}X + \frac{\partial F}{\partial Y}Y + \frac{\partial F}{\partial Z}Z = dF(X, Y, Z).$$

The tangent to the affine curve $f(x, y) = 0$ at the point (x_0, y_0) is

$$y - y_0 = \frac{dy}{dx}(x - x_0).$$

Differentiating $f(x, y) = 0$ with respect to x,

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = 0.$$

Thus the tangent can be written

$$\frac{\partial f}{\partial x}(x - x_0) + \frac{\partial f}{\partial y}(y - y_0) = 0,$$

or

$$\frac{\partial f}{\partial x}x + \frac{\partial f}{\partial y}y = \frac{\partial f}{\partial x}x_0 + \frac{\partial f}{\partial y}y_0.$$

Now

$$\left(\frac{\partial F}{\partial x}\right)_{(x_0,y_0,1)} = \left(\frac{\partial f}{\partial x}\right)_{(x_0,y_0)},$$

since $F(x, y, 1) = f(x, y)$. Moreover

$$\frac{\partial F}{\partial X}x_0 + \frac{\partial F}{\partial Y}y_0 + \frac{\partial F}{\partial Z} = 0,$$

since $F(x_0, y_0, 1) = 0$. Thus the affine tangent can be written in the form

$$\frac{\partial f}{\partial x}x + \frac{\partial f}{\partial y}y + \frac{\partial f}{\partial z} = 0;$$

corresponding to the projective line

$$\frac{\partial F}{\partial X}X + \frac{\partial F}{\partial Y}Y + \frac{\partial F}{\partial Z}Z = 0,$$

as we claimed.

This tangent is defined *unless*

$$\frac{\partial F}{\partial X} = \frac{\partial F}{\partial Y} = \frac{\partial F}{\partial Z} = 0,$$

In this case we say that P is a *singular* point on the curve. A curve is said to be *non-singular* if contains no singular points, either in k or in any extension field of k.

We say that the curve $F(X, Y, Z) = 0$ is *degenerate* if the polynomail F factorises:

$$F(X, Y, Z) = G(X, Y, Z)H(X, Y, Z).$$

A degenerate curve is always singular. For the points where the constituents meet,

$$G(X, Y, Z) = H(X, Y, Z) = 0,$$

are necessarily singular.

2.6 The characteristic 2 case

Now that we have defined what we mean by a singular point or a singular curve we can extend our definition of an elliptic curve over k to the case where $\mathrm{char}(k) = 2$.

Definition 2.1 *An elliptic curve over a field k is given by an equation*

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6 \qquad (c_1, c_2, c_3, c_4, c_6 \in k)$$

subject to the condition that the curve must be non-singular.

Note that the new definition is equivalent to our original definition of an elliptic curve when $\mathrm{char}(k) \neq 2$. For in that case we can complete the square on the left, and bring the equation to standard form; and we have seen that the curve is non-singular in this case precisely when the condition in our original definition — that the cubic on the right should be separable — is satisfied.

First we verify that there is no singularity at infinity.

Proposition 2.1 *The curve*

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

meets the line at infinity in just one point, $[0, 1, 0]$*. This is a point of inflection on the curve, and is non-singular.*

Proof ► The homogeneous form of the curve in this case is

$$F(X,Y,Z) = Y^2Z + c_1XYZ + c_3YZ^2 - X^3 - c_2X^2Z - c_4XZ^2 - c_6Z^3 = 0.$$

This meets the line at infinity $Z = 0$ where

$$X^3 = 0,$$

ie thrice at the point $[0,1,0]$, which is thus a point of inflection. To see that this point is non-singular, note that

$$\begin{aligned}\frac{\partial F}{\partial Z} &= Y^2 + c_1XY + 2c_3YZ - c_2X^2 - 2c_4XZ - 3c_3Z^2\\ &= 1\end{aligned}$$

at $[0,1,0]$, since all the terms except the first vanish. ◄

Now suppose $\mathrm{char}(k) = 2$. We shall establish a condition on the coefficients c_i for non-singularity.

We have seen that the point $[0,1,0]$ on the line at infinity is non-singular. So any singular point is in the affine plane.

In characteristic 2, $-1 = 1,\ 2 = 0,\ 3 = 1$, etc; so we have

$$\begin{aligned}\frac{\partial F}{\partial X} &= c_1YZ + X^2 + c_4Z^2,\\ \frac{\partial F}{\partial Y} &= c_1XZ + c_3Z^2 = Z(c_1X + c_3Z),\\ \frac{\partial F}{\partial Z} &= Y^2 + c_1XY + c_2X^2 + c_6Z^2.\end{aligned}$$

Thus if the point $(x,y) = [x,y,1]$ is singular then

$$\begin{aligned}&c_1y + x^2 + c_4 = 0,\\ &c_1x + c_3 = 0,\\ &y^2 + c_1xy + c_2x^2 + c_6 = 0.\end{aligned}$$

From the second equation,

$$c_1x = c_3.$$

If $c_1 = 0$ this implies that $c_3 = 0$, so that $\partial F/\partial Y = 0$ identically. In that case the point $(x,y) = [x,y,1]$ is singular if

$$\begin{aligned}x^2 + c_4 &= 0,\\ y^2 + c_2x^2 + c_6 &= 0.\end{aligned}$$

We may not be able to solve these equations in k, but we can always solve them in an extension of k, for example in its algebraic closure $\bar{k}$. Thus we have established what we said earlier; *the curve*

$$y^2 = x^3 + ax^2 + bx + c$$

is always singular in characteristic 2.

Now suppose $c_1 \neq 0$. In that case

$$x = c_3/c_1.$$

So from the first equation,

$$y = c_3^2/c_1^3 + c_4/c_1;$$

and then from the third equation,

$$c_3^4/c_1^6 + c_4^2/c_1^2 + c_3^3/c_1^3 + c_3c_4/c_1 + c_2c_3^2/c_1^2 + c_6 = 0.$$

(Note that $(a+b)^2 = a^2 + b^2$ in characteristic 2.) Multiplying by c_1^6 and re-ordering the terms,

$$c_1^6c_6 + c_1^4c_2c_3^2 + c_1^4c_4^2 + c_1^3c_3^3 + c_1^5c_3c_4 + c_3^4 = 0.$$

Conversely, if this is so then either $c_1 = c_3 = 0$, or else $c_1 \neq 0$, in which case on taking

$$x = c_3/c_1,\ y = c_3^2/c_1^3 + c_4/c_1$$

we see that

$$\frac{\partial F}{\partial X} = \frac{\partial F}{\partial Y} = \frac{\partial F}{\partial Z} = 0$$

at the point $(x,y) = [x,y,1]$.

Finally we observe that this point is necessarily on the curve, since

$$F(X,Y,Z) = X\frac{\partial F}{\partial X} + Y\frac{\partial F}{\partial Y} + Z\frac{\partial F}{\partial Z}.$$

We have established

Proposition 2.2 *The equation*

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

defines an elliptic curve in characteristic 2 if and only if

$$c_1^6c_6 + c_1^5c_3c_4 + c_1^4c_2c_3^2 + c_1^4c_4^2 + c_1^3c_3^3 + c_3^4 \neq 0.$$

2.7 The discriminant of an elliptic curve

We have established two conditions for non-singularity: the condition above when $\mathrm{char}(k) = 2$, and the condition that if $\mathrm{char}(k) \neq 2$ then the curve

$$y^2 = x^3 + ax^2 + bx + c$$

is non-singular if $D(p) \neq 0$, where $p(x)$ is the cubic polynomial on the right.

It is natural to ask if we can find a polynomial $D(c_1, c_2, c_3, c_4, c_6)$ such that the general Weierstrass equation is non-singular — and so defines an elliptic curve — if and only if $D \neq 0$, *in all characteristics*. We shall show that this is indeed the case, though the polynomial we get is so complicated that we shall never write it out explicitly.

Suppose for the moment that $\mathrm{char}(k) \neq 2$. Then we can bring the curve

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

to standard form by completing the square on the left, giving

$$y^2 = x^3 + ax^2 + bx + c$$

with

$$a = c_2 + c_1^2/4,\ b = c_4 + c_1c_3/2,\ c = c_6 + c_3^2/4.$$

We know that the curve is non-singular in this case if

$$D(p) = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2 \neq 0.$$

Substituting for a, b, c gives us a horrendous polynomial, say

$$\Delta(c_1, c_2, c_3, c_4, c_6).$$

It is clear that the coefficients of this polynomial will have denominators of the form $1/2^r$. We claim that the highest power of 2 appearing in these denominators is $2^4 = 16$. In other words, *the polynomial* 16Δ *has integer*

coefficients. To see that this is so, consider

$$\begin{aligned}
2^6\Delta = &- 2^8(c_2 + c_1^2/4)^3(c_6 + c_3^2/4)\\
&+ 2^6(c_2 + c_1^2/4)^2(c_4 + c_1c_3/2)^2\\
&+ 2^73^2(c_2 + c_1^2/4)(c_4 + c_1c_3/2)(c_6 + c_3^2/4)\\
&- 2^8(c_4 + c_1c_3/2)^3\\
&- 2^63^3(c_6 + c_3^2/4)^2\\
= &- (4c_2 + c_1^2)^3(4c_6 + c_3^2)\\
&+ (4c_2 + c_1^2)^2(2c_4 + c_1c_3)^2\\
&+ 2^23^2(4c_2 + c_1^2)(2c_4 + c_1c_3)(4c_6 + c_3^2)\\
&- 2^5(2c_4 + c_1c_3)^3\\
&- 2^23^3(4c_6 + c_3^2)^2.
\end{aligned}$$

Working modulo 4,

$$2^6\Delta \equiv -c_1^6c_3^2 + c_1^6c_3^2 \bmod 4.$$

Thus $2^4\Delta$ is a polynomial with integral coefficients.

Definition 2.2 *We define the discriminant of the curve*

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

to be

$$D(\mathcal{E}) = 2^4\Delta.$$

Proposition 2.3 *The equation*

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

defines an elliptic curve if and only if

$$D(\mathcal{E}) \neq 0.$$

Proof ▶ There is nothing to prove if $\mathrm{char}(k) \neq 2$, since the factor 2^4 then makes no difference; $D(\mathcal{E}) = 0$ if and only if the discriminant of the cubic $x^3 + ax^2 + bx + c$ is 0, which we know is the condition for the curve to be singular.

If $\mathrm{char}(k) = 2$ it is sufficient to show that $D(\mathcal{E})$ reduces to the polynomial in Proposition 2.2. In effect, we have to determine $2^6\Delta \bmod 8$. From the formulae in the calculation mod4 above,

$$\begin{aligned}
2^6\Delta &\equiv -12c_1^4c_2c_3^2 + 4c_1^6c_6 + 4c_1^4c_4^2 + 4c_1^5c_3c_4 + 2^23^2c_1^3c_3^3 - 2^33^3c_3^4 \bmod 8\\
&\equiv 4\left(c_1^4c_2c_3^2 + c_1^6c_6 + c_1^4c_4^2 + c_1^5c_3c_4 + c_1^3c_3^3 + c_3^4\right) \bmod 8
\end{aligned}$$

Thus in characteristic 2

$$D(\mathcal{E}) = 2^4\Delta = c_1^6c_6 + c_1^5c_3c_4 + c_1^4c_2c_3^2 + c_1^4c_4^2 + c_1^3c_3^3 + c_3^4,$$

which is exactly the polynomial which we showed vanished if and only if the curve is singular. ◄

2.8 On the Intersection of Curves

Suppose Γ_1, Γ_2 are 2 non-degenerate curves in $\mathbb{P}^2$ defined by homogeneous equations

$$F_1(X,Y,Z) = 0, \ F_2(X,Y,Z) = 0,$$

of degrees n_1, n_2; and suppose

$$P \in \Gamma_1 \cap \Gamma_2.$$

Then one can define an integer $I(P;\Gamma_1,\Gamma_2) \geq 1$, the *intersection number* of Γ_1 and Γ_2 at P.

In the 'generic' case, where Γ_1 and Γ_2 are non-singular at P, and the tangents to the 2 curves at P are distinct, the intersection number $I(P;\Gamma_1,\Gamma_2) = 1$.

We shall not define the intersection number in the general case — although the definition is not particularly complicated — but only in the special case which we need, where one (or both) of the curves is a line.

Let Λ then be the line

$$aX + bY + cZ = 0;$$

and let Γ be the curve

$$F(X,Y,Z) = 0,$$

where $F(X,Y,Z)$ is homogeneous of degree d.

If

$$P_1 = [X_1,Y_1,Z_1], \ P_2 = [X_2,Y_2,Z_2]$$

are 2 points of Λ then the general point $P \in \Lambda$ can be written

$$P = uP_1 + vP_2 = [uX_1 + vX_2, uY_1 + vY_2, uZ_1 + vZ_2].$$

We may regard u, v as homogeneous coordinates for the line Λ.

This line meets the curve where

$$H(u,v) \equiv F(uP_1 + vP_2) = 0,$$

which is a homogenous equation of degree d in u, v.

If now $P = (u_0, v_0) \in \Lambda \cap \Gamma$ then $uv_0 - vu_0$ is a factor of $H(u,v)$. We define the intersection number $I(P;\Lambda,\Gamma)$ to be the multiplicity of this factor in $H(u,v)$.

It is readily verified that this number is independent of the choice of points $P_1, P_2 \in \Lambda$.

If the ground field k is algebraically closed then $H(u,v)$ factorises completely into linear factors; and it follows that the sum of the intersection numbers is equal to the degree:

$$\sum_{P \in \Lambda\cap\Gamma} I(P;\Lambda,\Gamma) = \deg \Gamma.$$

In the general case — where k is not algebraically closed — this reduces to an inequality:

$$\sum_{P \in \Lambda\cap\Gamma} I(P;\Lambda,\Gamma) \leq \deg \Gamma.$$

These results break down if Λ is a factor of Γ, ie

$$F(X,Y,Z) = (aX + bY + cZ)G(X,Y,Z),$$

where G is of degree $d-1$. In this case the intersection numbers are undefined.

Proposition 2.4 *Suppose P is a point on the non-singular curve Γ of degree ≥ 2. Let Λ denote the tangent to Γ at P. Then*

$$I(P;\Lambda,\Gamma) \geq 2.$$

Proof ► Let us take $P = [X_0,Y_0,Z_0]$ and a second point $Q = [X_1,Y_1,Z_1]$ on Λ to define the homogeneous coordinates (u,v) on Λ. By the 2-dimensional version of Taylor's Theorem,

$$\begin{aligned} H(u,v) =& F(uP + vQ) \\ =& u^d F(P) + \\ & u^{d-1}v\left[\left(\frac{\partial F}{\partial X}\right)_P (X_1 - X_0) + \left(\frac{\partial F}{\partial Y}\right)_P (Y_1 - Y_0) + \left(\frac{\partial F}{\partial Z}\right)_P (Z_1 - Z_0)\right] + \cdots. \end{aligned}$$

Since P and Q both satisfy

$$\frac{\partial F}{\partial X}X + \frac{\partial F}{\partial Y}Y + \frac{\partial F}{\partial Z}Z = 0,$$

the coefficient of $u^{d-1}v$ is 0; while the coefficient of u^d is 0 since $F(P) = 0$. Thus H has a double zero at $u = 0$, ie

$$I(P; \Lambda, \Gamma) \geq 2.$$

◄

Remarks:

1. The result still holds if Γ is singular, provided the tangent Λ is not a factor of Γ.

2. We can use the intersection number to define the 'badness' or multiplicity of a singularity. For suppose P is a singular point on the curve Γ. It follows from our equation for $H(u, v)$ above that the coefficient of $u^{d-1}v$ vanishes identically, for any line Λ through P. Thus

   $$\min_{\Lambda \ni P} I(P; \Lambda, \Gamma) \geq 2.$$

   We define this minimum to be the *multiplicity* of the singularity at P.

2.8.1 Bezout's Theorem

Proposition 2.5 *Two curves Γ_1, Γ_2 in $\mathbb{P}^2$ of degrees n_1, n_2 cannot meet in more than $n_1 n_2$ points, unless they have a factor in common.*

Proof ► We may assume that the field k we are working over is infinite; for otherwise we can pass to an infinite extension of k (for example, the algebraic closure $\bar{k}$ of k, or the field $k(t)$ of rational functions over k).

Let the curves be given by the homogeneous equations

$$F_1(X, Y, Z) = 0, \; F_2(X, Y, Z) = 0,$$

of degrees n_1, n_2.

Suppose the curves have $n_1 n_2 + 1$ points in common, say

$$P_0, P_1, \ldots, P_{n_1 n_2}.$$

We can find a line $ax + by + cz$ not passing through any of these points; and we can take this line as the line at infinity. Thus we may assume that the $n_1 n_2 + 1$ points are all in the affine plane k^2. In this way we can reduce the problem to the affine case, in which the curves are given by affine equations

$$f_1(x, y) = 0, \; f_2(x, y) = 0,$$

where f_1, f_2 are non-homogeneous polynomials of degrees $\leq n_1, n_2$.

By making a further change of coordinates, if necessary, we may assume that the $n_1 n_2 + 1$ points

$$P_i = (x_i, y_i)$$

have distinct x-coordinates and distinct y-coordinates.

Now let us regard f_1, f_2 as polynomials in y, and let us compute their resultant $R(f_1, f_2)$. This is a polynomial of degree $\leq n_1 n_2$ in x.

For each x_i the polynomials $f_1(x_i, y), f_2(x_i, y)$ have a factor $y - y_i$ in common. It follows that the resultant $R(x)$ must vanish for these values of x. Thus $R(x)$ has more roots than its degree, and so must vanish identically.

But that implies that the polynomials $f_1(x, y), f_2(x, y)$ have a factor in common, say

$$f_1(x, y) = m(x, y) g_1(x, y), \; f_2(x, y) = m(x, y) g_1(x, y).$$

It follows that the original homogeneous polynomials have a factor in common:

$$F_1(X, Y, Z) = M(X, Y, Z) G_1(X, Y, Z), \; F_2(X, Y, Z) = M(X, Y, Z) G_1(X, Y, Z).$$

◄

Remarks:

1. If the curves have a factor in common, and if the field we are working over is infinite, then of course the curves have an infinity of points in common.

2. The Proposition above is a very feeble form of Bezout's Theorem, which states in its fullness that *if Γ_1, Γ_2 are curves in $\mathbb{P}^2(k)$, where k is an algebraically closed field, and Γ_1, Γ_2 have no factor in common, then*

   $$\sum_{P \in \Gamma_1 \cap \Gamma_2} I(P; \Gamma_1, \Gamma_2) = \deg \Gamma_1 \, \deg \Gamma_2.$$

   In other words, the number of points of intersection, if each is counted with due multiplicity, is equal to the product of the degrees.

There is a small addendum to Bezout's Theorem which we shall find very useful.

Proposition 2.6 *Suppose the curves Γ_1, Γ_2 of degrees n_1, n_2 over k have $(n_1 n_2 - 1)$ points over k in common, but have no factor in common. Then they have a further point over k in common.*

Proof ▶ When we eliminate Z say as above (in the proof of Bezout's Theorem) we are left with a homogeneous polynomial over k of degree $n_1 n_2$ in X, Y. We know that this polynomial has $(n_1 n_2 - 1)$ roots in k. It follows that the last root is also in k, by the homogeneous analogue of the fact that the sum of the roots of the polynomial

$$t^d + a_1 t^{d-1} + \cdots + a_d = 0$$

is equal to $-a_1$. ◀

In effect we have used a particular case of this result (with $n_1 = 1,\ n_2 = 3$) in our assertion that if $P, Q \in \mathcal{E}$ then $P * Q \in \mathcal{E}$; the line PQ meets $\mathcal{E}$ in two points over k, so it meets $\mathcal{E}$ in a third point over k.

2.9 Points of Inflection

Consider the curve

$$y^2 + c_1 xy + c_3 y = x^3 + c_2 x^2 + c_4 x + c_6,$$

or in homogeneous form,

$$F(X, Y, Z) = Y^2 Z + c_1 XYZ + c_3 YZ^2 - X^3 - c_2 X^2 Z - c_4 XZ^2 - c_6 Z^3 = 0.$$

This meets the line at infinity $Z = 0$ where

$$X^3 = 0,$$

ie thrice at the point $[0, 1, 0]$. Thus the line at infinity is the tangent to the curve at $[0, 1, 0]$ — but it is more than that, it is a point of inflection.

Definition 2.3 *A non-singular point P on the curve*

$$\Gamma : F(X, Y, Z) = 0$$

is said to be a point of inflection *(or* flex*) if the tangent Λ at P intersects Γ with multiplicity at least 3:*

$$I(P; \Lambda, \Gamma) \geq 3.$$

Proposition 2.7 *Suppose P is a non-singular point on*

$$\Gamma : F(X, Y, Z) = 0,$$

where $F(X, Y, Z)$ is a homogeneous polynomial of degree ≥ 2. Then P is a point of inflection on Γ if and only if it satisfies the hessian equation

$$H(X, Y, Z) \equiv \det \begin{pmatrix} \frac{\partial^2 F}{\partial X^2} & \frac{\partial^2 F}{\partial X \partial Y} & \frac{\partial^2 F}{\partial X \partial Z} \\ \frac{\partial^2 F}{\partial X \partial Y} & \frac{\partial^2 F}{\partial Y^2} & \frac{\partial^2 F}{\partial Y \partial Z} \\ \frac{\partial^2 F}{\partial X \partial Z} & \frac{\partial^2 F}{\partial Y \partial Z} & \frac{\partial^2 F}{\partial Z^2} \end{pmatrix} = 0.$$

Proof ▶ Let $P = [X, Y, Z]$; and suppose $Q = [X', Y', Z']$. Each point of the line PQ can be written in the form

$$uP + vQ = [uX + vX', uY + vY', uZ + vZ'].$$

We can regard (u, v) as *homogeneous coordinates* on the line PQ. This line meets Γ where

$$F(uP + vQ) = 0.$$

If $\deg F = d$, this expands to

$$u^d F(P) + u^{d-1} v \left[\frac{\partial F}{\partial X} X' + \frac{\partial F}{\partial Y} Y' + \frac{\partial F}{\partial Z} Z' \right] +$$
$$\frac{1}{2} u^{d-2} v^2 \left[\frac{\partial^2 F}{\partial X^2} X'^2 + \frac{\partial^2 F}{\partial Y^2} Y'^2 + \frac{\partial^2 F}{\partial Z^2} Z'^2 + 2 \frac{\partial^2 F}{\partial X \partial Y} X'Y' + 2 \frac{\partial^2 F}{\partial X \partial Z} X'Z' + 2 \frac{\partial^2 F}{\partial X \partial Z} Y'Z' \right] + \cdots$$

Thus the line PQ will intersect Γ at P with multiplicity ≥ 3 if and only if

$$L(X', Y', Z') \equiv \frac{\partial F}{\partial X} X' + \frac{\partial F}{\partial Y} Y' + \frac{\partial F}{\partial Z} Z' = 0$$

and

$$M(X', Y', Z') \equiv \frac{\partial^2 F}{\partial X^2} X'^2 + \frac{\partial^2 F}{\partial Y^2} Y'^2 + \frac{\partial^2 F}{\partial Z^2} Z'^2 + 2 \frac{\partial^2 F}{\partial X \partial Y} X'Y' + 2 \frac{\partial^2 F}{\partial X \partial Z} X'Z' + 2 \frac{\partial^2 F}{\partial X \partial Z} Y'Z' = 0.$$

The first condition simply states that PQ is the tangent to Γ at P.

On setting $Q = P$,

$$F(uP + vP) \equiv 0.$$

Hence

$$M(P) = 0.$$

Thus the equation

$$Q(X, Y, Z) = 0$$

represents a conic through P.

Lemma 1 *The tangent to the conic $M(X, Y, Z) = 0$ at P coincides with the tangent to Γ at P.*

Remark: It would be surprising if this were not so; for in that case we would have defined in an intrinsic way a second line passing through any point P of a curve. One might think of the normal to the curve at P. But angle is not a projective invariant, so this would not make sense.

Proof of Lemma ▷ To avoid confusion, let us for a moment set $P = [X_0, Y_0, Z_0]$. Then the tangent to $M(X, Y, Z) = 0$ at P is

$$\left(\frac{\partial^2 F}{\partial X^2}X_0 + \frac{\partial^2 F}{\partial X \partial Y}Y_0 + \frac{\partial^2 F}{\partial X \partial Z}Z_0\right) X + \left(\frac{\partial^2 F}{\partial X \partial Y}X_0 + \frac{\partial^2 F}{\partial Y^2}Y_0 + \frac{\partial^2 F}{\partial Y \partial Z}Z_0\right) Y + \left(\frac{\partial^2 F}{\partial X \partial Z}X_0 + \frac{\partial^2 F}{\partial Y \partial Z}Y_0 + \frac{\partial^2 F}{\partial Z^2}Z_0\right) Z = 0.$$

Now $\partial F/\partial X, \partial F/\partial Y, \partial F/\partial Z$ are all homogeneous polynomials of degree $d-1$. But recall that if $F(X, Y, Z)$ is a homogeneous polynomial of degree d then

$$\frac{\partial F}{\partial X}X + \frac{\partial F}{\partial Y}Y + \frac{\partial F}{\partial Z}Z = dF(X, Y, Z).$$

Applying this with $\partial F/\partial X$ in place of F,

$$\frac{\partial^2 F}{\partial X^2}X + \frac{\partial^2 F}{\partial X \partial Y}Y + \frac{\partial^2 F}{\partial X \partial Z}Z = (d-1)\frac{\partial F}{\partial X}.$$

Similarly

$$\frac{\partial^2 F}{\partial X \partial Y}X + \frac{\partial^2 F}{\partial Y^2}Y + \frac{\partial^2 F}{\partial Y \partial Z}Z = (d-1)\frac{\partial F}{\partial Y}$$
$$\frac{\partial^2 F}{\partial X \partial Z}X + \frac{\partial^2 F}{\partial Y \partial Z}Y + \frac{\partial^2 F}{\partial Z^2}Z = (d-1)\frac{\partial F}{\partial Z}.$$

Thus the tangent to the conic $M(X, Y, Z)$ at P is just

$$\frac{\partial F}{\partial X}X + \frac{\partial F}{\partial Y}Y + \frac{\partial F}{\partial Z}Z = 0,$$

which is the tangent to Γ at P ◁

Now suppose P is a point of inflection. Then

$$L(X, Y, Z) = 0 \Longrightarrow M(X, Y, Z) = 0.$$

It follows that L is a factor of M, say

$$M(X, Y, Z) = L(X, Y, Z)L'(X, Y, Z),$$

where L' is a second line. In particular the conic $M(X, Y, Z)$ is degenerate.

Lemma 2 *The conic*

$$C(X, Y, Z) \equiv aX^2 + bY^2 + cZ^2 + 2fYZ + 2gXZ + 2hYZ = 0$$

degenerates into 2 lines if and only if

$$\det A = 0,$$

where

$$A = \begin{pmatrix} a & h & g \\ h & b & f \\ g & f & c \end{pmatrix} = 0.$$

Proof of Lemma ▷ Suppose

$$C(X, Y, Z) \equiv L_1(X, Y, Z)L_2(X, Y, Z).$$

Let the lines $L_1 = 0,\ L_2 = 0$ meet in the point (X_0, Y_0, Z_0). Then the tangent at (X_0, Y_0, Z_0) is undefined. Thus

$$Av_0 = 0,$$

where

$$v_0 = \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix}.$$

Hence A is singular, ie $\det A = 0$.

Conversely, suppose $\det A = 0$. Then we can find X_0, Y_0, Z_0 satisfying the equation $Av_0 = 0$. It follows that the tangent to Γ at *any* point P passes through $P_0 = [X_0, Y_0, Z_0]$. But now take any point P. The tangent at P cuts the conic $C(X, Y, Z) = 0$ twice at P and at P_0. But a line can only cut a conic twice. It follows that the line P_0P lies wholly in the conic, which must thus degenerate into 2 lines. ◁

Putting this together, if P is a flex, then the conic $M(X, Y, Z) = 0$ is degenerate and so $H(X, Y, Z) = 0$.

Conversely, if $H(X, Y, Z) = 0$ then $M(X, Y, Z) = 0$ is degenerate. Since the tangent to this conic at P is $L(X, Y, Z) = 0$, this line must be one of the lines making up the conic:

$$M(X, Y, Z) = L(X, Y, Z)L_1(X, Y, Z).$$

Thus L is a factor of M, and so P is a flex. ◀

As we saw, the point $[0, 1, 0]$ is a flex on an elliptic curve given by Weierstrass' equation. We shall always take this point as the zero element O of the group on the curve. The other flexes are just the points of order 3 in the group. Thus flexes play an important rôle in the theory.

The hessian curve of a cubic is itself a cubic. But 2 cubics meet in at most 9 points — as may be seen by considering the resultant of the 2 polynomials, which is a homogeneous polynomial of degree 9 in 2 variables. It follows that an elliptic curve has at most 9 flexes.

We shall see that an elliptic curve over the reals $\mathbb{R}$ has at most 3 flexes; and the same is therefore true of an elliptic curve over the rationals $\mathbb{Q}$ (which is our main focus of interest).

2.10 Milestones on the Road to Modern Geometry

Euclid (c325BC–c265BC) Whether the work of one man or a school, the introduction of axiomatic methods in *Euclid's Elements* surely marks the greatest leap in the history of mathematics.

René Descartes (1596–1650) By representing a point P in the plane by its coordinates (x, y), Descartes converted geometric into algebraic problems — the start of algebraic geometry.

Bernard Riemann (1826–1866) Although not explicitly geometrical, Riemann's study of what are today known as Riemann surfaces had a profound influence on the theory of curves — in particular his definition of the *genus*, the most important characteristic of a curve.

Felix Klein (1849–1925) In his *Erlangen program* Klein distinguished between different geometries according to their transformation groups — as for example, Euclidean geometry, affine geometry and projective geometry.

David Hilbert (1862–1943) The polynomials satisfied by the points on a variety form an *ideal* in the ring $k[x_1, \ldots, x_n]$. Hilbert showed in his *Finite Basis Theorem* that every such ideal is generated by a finite number of polynomials.

Severi (1879–1961) and the Italian School studied general *algebraic varieties*, that is, the points satisfying a set of polynomial equations.

André Weil (1906–1998) In his seminal work, *The Foundations of Algebraic Geometry*, Weil provided a secure foundation for the work of the Italian school, and extended it to varieties over finite and other fields, not just $\mathbb{C}$.

Alexandre Grothendieck (1928–) In what is perhaps the greatest mathematical work of the 20th century, Grothendieck merged algebraic geometry with commutative algebra, by extending the notion of variety to include the "scheme" of a commutative ring. For example, to the integers $\mathbb{Z}$ there corresponds a 'scheme' — a generalized variety — over the space

$$\operatorname{spec}(\mathbb{Z}) = \{0, 2, 3, 5, \ldots\},$$

whose points correspond to primes p (more precisely, to prime *ideals*, hence the inclusion of 0).

Chapter 3

The Group on an Elliptic Curve

Every elliptic curve $\mathcal{E}(k)$ has a natural structure as an abelian group. We will always represent this group additively, denoting the *sum* of two points $P, Q \in \mathcal{E}(k)$ by $P + Q$.

The basic idea is that

$$P + Q + R = 0 \Longleftrightarrow P, Q, R \text{ are collinear.}$$

But as we shall see, this is not quite sufficient to define the group structure. Also, since this is the basis for the entire theory of elliptic curves we need to ensure that we are on a firm foundation.

Proposition 3.1 *Suppose P, Q are points on the elliptic curve $\mathcal{E}(k)$. Then the line PQ (or the tangent at P if $P = Q$) meets $\mathcal{E}(k)$ again at a unique point R.*

Proof ► Let the line PQ (or the tangent at P) be

$$lX + mY + nZ = 0.$$

If $n \neq 0$ then we can eliminate Z by substituting

$$Z = -\frac{lX + mY}{n}$$

in the original cubic equation, giving a homogeneous cubic in X, Y:

$$a_0 X^3 + a_1 X^2 Y + a_2 X Y^2 + a_3 Y^3 = 0.$$

(If $n = 0$ then we eliminate X or Y in the same way instead.)

Two of the roots of this cubic are given by P, Q, leaving the third root (which must be in the field k) to determine the point R. ◄

Remark: When we speak of a root of a homogeneous polynomial in X, Y we mean of course the ratio $X_0 : Y_0$; and when we say that the root is in k we mean that we can find $X_0, Y_0 \in k$ in this ratio.

The proposition that if $n - 1$ roots of a polynomial $p(x) \in k[x]$ lie in k then so does the nth root carries over unchanged to the homogeneous case.

3.1 Choice of zero point

Recall that the elliptic curve

$$\mathcal{E}(k) : y^2 + c_1 xy + c_3 y = x^3 + c_2 x^2 + c_4 x + c_6$$

has just one point on the line at infinity, namely $[0, 1, 0]$. *We will always choose this as the zero point of our abelian group:*

$$O = [0, 1, 0].$$

Accordingly, the inverse $-P$ of any point P is the point where OP meets $\mathcal{E}(k)$ again:

$$-P = O * P.$$

This gives us the definition of $P + Q$.

Definition 3.1 *Let $\mathcal{E}(k)$ be the elliptic curve*

$$y^2 + c_1 xy + c_3 y = x^3 + c_2 x^2 + c_4 x + c_6.$$

The sum of two points $P, Q \in \mathcal{E}(k)$ is defined to be

$$P + Q = O * (P * Q).$$

It is evident that this operation is commutative:

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

It is clear too that the point O serves as neutral element:

$$O + P = O * (O * P) = P.$$

Also each point P has negation $-P = O * P$, since

$$\begin{aligned} P + (O * P) &= O * (P * (O * P)) \\ &= O * (P * (P * O)) \\ &= O * O \\ &= O, \end{aligned}$$

since the tangent at O meets $\mathcal{E}$ again at O, as O is a point of inflection.

However, it is far from evident that the operation is associative:

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

We shall prove this important result in the next Chapter. But for the moment we shall assume that it is true, and look at some concrete examples of the group on an elliptic curve.

First though, let us get an explicit expression for $-P$ when

$$P = (x_0, y_0) = [x_0, y_0, 1].$$

The line OP is

$$X - x_0 Z = 0,$$

since this certainly goes through P and $O = [0, 1, 0]$.

In affine terms this is the line

$$x = x_0,$$

ie the line through P parallel to the y-axis.

Suppose the elliptic curve is in standard form

$$y^2 = x^3 + ax^2 + bx + c = 0.$$

In this case the line $x = x_0$ meets the curve again at the point $(x_0, -y_0)$. Thus *if the elliptic curve is given in standard form then*

$$-(x, y) = (x, -y).$$

In the more general case

$$y^2 + c_1 xy + c_3 y = x^3 + c_2 x^2 + c_4 x + c_6$$

the line $x = x_0$ meets the curve where

$$y^2 + (c_1 x_0 + c_3)y - (x_0^3 + c_2 x_0^2 + c_4 x + c_6) = 0.$$

One root of this equation for y is y_0. If the other root is y_1 then

$$y_0 + y_1 = -(c_1 x_0 + c_3),$$

ie

$$y_1 = -y_0 - c_1 x_0 - c_3.$$

Thus

$$-(x, y) = (x, -y - c_1 x - c_3).$$

3.2 Examples

1. Consider the curve

   $$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + 1$$

   over the rationals $\mathbb{Q}$. There are 5 obvious points on this curve:

   $$P = (-1, 0),\ Q = (0, 1),\ -Q = (0, -1),\ R = (2, 3),\ -R = (2, -3).$$

   (These all have integer coordinates; but it is important to bear in mind that we are interested in any *rational* solutions.)

   Let us determine $P + Q$. Suppose the line PQ is

   $$y = mx + c.$$

   The slope m is

   $$m = \frac{1-0}{0-(-1)} = 1.$$

   Thus PQ is the line

   $$y = x + 1.$$

   This meets the curve where

   $$(x-1)^2 = x^3 + 1.$$

   We know two of the roots: $-1, 0$ from P, Q. It follows (by looking at the coefficient of x^2 that if the third root is x_2 then

   $$-1 + 0 + x_2 = 1,$$

   ie

   $$x_2 = 2.$$

   Thus $y_2 = x_2 + 1 = 3$, ie

   $$P * Q = (2, 3) = R.$$

   It follows that

   $$P + Q = -R = (2, -3).$$

   Next let us determine $2R$. To determine the slope at R note that

   $$2y\frac{dy}{dx} = 3x^2,$$

ie

$$\frac{dy}{dx} = \frac{3x^2}{2y}.$$

In particular the slope at R is

$$m = \frac{12}{6} = 2;$$

and so the tangent at R is

$$y = 2x - 1.$$

This line meets the curve again where

$$(2x + c)^2 = x^3 + 1.$$

Two of the roots of this are $2, 2$ from R (twice). Thus if the other root is x_2 then (from the coefficient of x^2)

$$2 + 2 + x_2 = 2^2,$$

ie

$$x_2 = 0.$$

Thus

$$R * R = (0, -1) = -Q,$$

and so

$$2R = Q.$$

Note that

$$-P = (-1, 0) = P,$$

ie

$$2P = 0;$$

the point P is of order 2.

In fact it is clear that *the point $P = (x_0, y_0)$ on the curve*

$$y^2 = x^3 + ax^2 + bx + c$$

is of order 2 if and only if $y_0 = 0$, ie *if and only if P lies on the x-axis.*

More generally, suppose the curve is given by

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6.$$

If $P = (x_0, y_0)$ then as we saw

$$-P = (x_0, -y_0 - c_1x_0 - c_3).$$

Thus P is of order 2 if and only if

$$2y_0 + c_1x_0 + c_3 = 0,$$

ie if and only if P lies on the 'line of symmetry'

$$2y + c_1x + c_3 = 0.$$

In either case, the line meets the curve in 0, 1 or 3 points. Thus there are *either 0, 1 or 3 points of order 2 on an elliptic curve.*

Finally, let us determine $2Q$. The slope at Q is

$$m = \frac{0}{2} = 0.$$

Thus the tangent at Q is $y = 1$. If this meets the curve again at (x_2, y_2) then

$$0 + 0 + x_2 = 0^2,$$

ie

$$x_2 = 0.$$

Hence

$$Q * Q = Q,$$

ie

$$2Q = -Q,$$

ie

$$3Q = 0.$$

Thus Q is of order 3, ie Q is a point of inflection on the curve. Since $P + Q = -R$, while P is of order 2, it follows that $\pm R$ are of order 6; and the 6 elements

$$\{0, P, \pm Q, \pm R\}$$

form a cyclic group of order 6.

We shall see later that these are the only rational points on this elliptic curve:

$$\mathcal{E}(\mathbb{Q}) = C_6.$$

In particular there are no integers such that

$$y^2 = x^3 + 1$$

except $(x, y) = (-1, 0), (0, \pm 1), (2, \pm 3)$. However, this will require considerable apparatus to establish.

The group on the elliptic curve in this case is finite. *There is no known algorithm to determine whether the group on a general elliptic curve over $\mathbb{Q}$ is finite or infinite.* There are techniques which are likely to work in any given case, but there is no guarantee that they will work.

One important property of the group *is* known: *Mordell's Theorem* states that *the group on an elliptic curve $\mathcal{E}$ over $\mathbb{Q}$ is finitely-generated.* In other words, there are points $P_1, \ldots, P_r \in \mathcal{E}$ such that every rational point $P \in \mathcal{E}$ is expressible in the form

$$P = n_1 P_1 + \cdots + n_r P_r.$$

Our main aim in the first part of the course is to prove Mordell's Theorem.

2. Let us look at the same equation

$$\mathcal{E}(\mathcal{F}_5) : y^2 = x^3 + 1$$

but now over the finite field $\mathcal{F}_5$. The curve is still non-singular, since

$$D = -4 = 1$$

in $\mathcal{F}_5$.

We can easily find all the points on the curve. We have to find all (x, y) with $0 \le x, y \le 4$, or if we prefer $x, y \in \{0, \pm 1, \pm 2\}$, for which

$$y^2 \equiv x^3 + 1 \bmod 5.$$

In other words, we have to determine for each x whether or not $x^3 + 1$ is a quadratic residue mod5.

The quadratic residues mod5 are $0, 1, 4$. The results are given in the following table:

x	x^3+1	y
0	1	± 1
1	2	
2	4	± 2
-2	3	
-1	0	0

We see that there are 6 points in the group, including the zero point $O = [0, 1, 0]$:

$$O, (0, \pm 1), (2, \pm 2), (-1, 0).$$

There is only one abelian group of order 6, namely the cyclic group $C_6 = \mathbb{Z}/(6)$. Thus

$$\mathcal{E}(\mathcal{F}_5) = C_6.$$

There is just one element of order 2, namely $P = (-1, 0)$, since this is the only point of the curve on the x-axis $y = 0$.

Let us determine the order of $Q = (0, 1)$. The method is exactly the same as in the rational case. As there, the slope of the curve is given by

$$\frac{dy}{dx} = \frac{3x^2}{2y}.$$

In particular, the slope at Q is $m = 0$, so that the tangent at Q is

$$y = 1.$$

Since this is the only point with $y = 1$ it follows that

$$Q * Q = Q,$$

and so

$$Q + Q = -Q,$$

ie

$$3Q = 0.$$

Thus Q is of order 3, as also is $-Q$. The remaining 2 points must be of order 6, since C_6 has 1 element each of orders 1 and 2, and 2 elements each of orders 3 and 6.

(You may feel a little queasy about using the differential calculus over a finite field, or even the rationals. But in fact we are only using the derivative in a formal or algebraic sense, as for example if $f(x)$ is a polynomial over k then

$$f(x) - f(a) \equiv (x-a)f'(a) \bmod (x-a)^2,$$

ie

$$f(x) - f(a) = (x-a)f'(a) + (x-a)^2 g(x)$$

for some polynomial $g(x)$.)

What is $P + Q$? We leave that to the reader.

Elliptic curves over finite fields are used in *cryptography*, both in creating codes and in trying to crack them.

More generally, such curves provide on of the most powerful tools for trying to *factorise* large numbers.

Determining the number of points on an elliptic curve over a finite field has been an important topic in the development of the theory of elliptic curves, and many questions in this area remain open. If we take an elliptic curve over the field $\mathcal{F}_p$ (where $p \neq 2$) in the form

$$\mathcal{E}(\mathcal{F}_p) : y^2 = x^3 + ax^2 + bx + c$$

then we may expect the cubic $p(x) = x^3 + ax^2 + bx + c$ to be a quadratic residue for about half the values $x \in \{0, 1, \dots, p-1\}$. Each of these will give two solutions $\pm y$ unless $y = 0$, in which case it gives one. To these we must add the point $O = [0, 1, 0]$. Thus the 'expected' number of solutions is about $p+1$. *Hasse's Theorem* tells us that if the number of points is actually $p+1+a_p$, then the 'discrepancy' a_p is bounded by

$$|a_p| < 2\sqrt{p}.$$

The values of a_p for the same equation but different primes p have remarkable and mysterious properties, related to modular forms and Fermat's Last Theorem, which have still not been elucidated.

That is well beyond the scope of this course (although we shall have something to say about modular forms), but there is one related topic that we shall deal with.

It turns out that any elliptic curve $\mathcal{E}(\mathbb{Q})$ over the rationals can be 'reduced mod$p$' to give a curve $\mathcal{E}(\mathcal{F}_p)$ over the finite field $\mathcal{F}_p$. This curve may be singular for a finite set of so-called 'bad' primes (for that particular curve), but it will remain an elliptic curve for the remaining primes. Furthermore it will emerge that there is a natural homomorphism

$$\mathcal{E}(\mathbb{Q}) \to \mathcal{E}(\mathcal{F}_p)$$

for each of these 'good' primes p; and the study of these homomorphisms is one of the many tools we shall have to hand for studying the curve $\mathcal{E}(\mathbb{Q})$.

3. Let us look now at the elliptic curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 - 2x.$$

We see that this contains the points

$$P = (0,0),\ Q = (2,2),\ -Q = (2,-2).$$

We know that P has order 2.

Let us determine $2Q$. The slope is given by

$$2y\frac{dy}{dx} = 3x^2 - 2,$$

ie

$$\frac{dy}{dx} = \frac{3x^2-2}{2y}$$

At P,

$$m = \frac{10}{4} = \frac{5}{2}.$$

Thus the tangent at P is

$$(y-2) = \frac{5}{2}(x-2),$$

ie

$$5x - 2y - 6 = 0.$$

If this tangent meets the curve again at (x_2, y_2) then

$$2 + 2 + x_2 = m^2 = \frac{25}{4},$$

ie

$$x_2 = \frac{9}{4}.$$

Thus

$$P * P = \left(\frac{9}{4}, \frac{21}{8}\right),$$

and so

$$2P = \left(\frac{9}{4}, -\frac{21}{8}\right).$$

We shall show later that *a point (x, y) of finite order on the elliptic curve*

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

necessarily has integer coordinates $x, y \in \mathbb{Z}$. (This is quite difficult to prove — though not as difficult as Mordell's Theorem! Essentially we have to show that as we successively double the point, $2Q, 4Q, 8Q, \ldots$, the denominator of the slope m gets larger and larger.)

It will follow from this that the point Q is of infinite order. In particular *the group $\mathcal{E}(\mathbb{Q})$ in this case is infinite.*

4. Next, let us look at a curve in general Weierstrass format:

$$\mathcal{E}(\mathbb{Q}) : y^2 - y = x^3 - x.$$

We could bring this to standard form, as follows. Completing the square on the left,

$$(y - 1/2)^2 = x^3 - x + 1/4,$$

ie

$$y_1^2 = x^3 - x + 1/4$$

after the change of coordinate $y_1 = y - 1/2$.

Note that an equation in standard form remains in standard form under any change of coordinates of the form

$$x_2 = a^2x,\ y_2 = a^3y_1,$$

since the coefficients of y^2 and x^3 will still be the same after such a change. In the present case, if we take $a = 2$ the equation becomes

$$y_2^2 = x_2^3 - 16x_2 + 16,$$

under the change of coordinates

$$x_2 = 4x,\ y_2 = 8y - 4.$$

This device can be used to bring any equation

$$\mathcal{E}(\mathbb{Q}) : y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

with rational coefficients to an equation

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

with integer coefficients a, b.

However, this is not necessarily the best policy, since the coefficients a, b one finishes up with will in general be much larger than the original coefficients.

In the present case, we shall stick with the original equation

$$\mathcal{E}(\mathbb{Q}) : y^2 - y = x^3 - x.$$

This curve contains a number of obvious points:

$$P = (0, 0),\ Q = (1, 0),\ R = (-1, 0),\ S = (0, 1),\ T = (1, 1),\ U = (-1, 1).$$

If $P = (x, y) \in \mathcal{E}$ then

$$-P = (x, 1 - y).$$

Thus

$$-P = S,\ -Q = T,\ -R = U.$$

Let us determine $P+Q$. The line PQ has slope

$$m = \frac{0}{1} = 0;$$

so PQ is the line

$$y = 0.$$

This meets the curve again at $(-1,0)$. Thus

$$P+Q = -(-1,0) = (-1,1),$$

ie

$$P+Q = U.$$

Now let us determine $2Q$. The slope is given by

$$(2y-1)\frac{dy}{dx} = 3x^2 - 1,$$

ie

$$\frac{dy}{dx} = \frac{3x^2-1}{2y-1}.$$

In particular, the slope at Q is

$$m = \frac{2}{-1} = -2.$$

Thus the tangent at Q is

$$y = -2x + 2.$$

This meets the curve where

$$(-2x+2)^2 - (-2x+2) = x^3 - x.$$

Thus if the tangent meets $\mathcal{E}$ again at (x_2, y_2) then (looking as usual at the coefficient of x^2)

$$1 + 1 + x_2 = m^2 = 4,$$

and so

$$Q * Q = (2, -2).$$

Thus

$$2Q = -(2,-2) = (2,3) = V.$$

We leave it to the reader to determine $2V$. Is the order of Q finite or infinite?

5. Finally, let us look at the same equation over the field $\mathcal{F}_2$:

$$\mathcal{E}(\mathcal{F}_2) : y^2 - y = x^3 - x.$$

First we must verify that this *is* an elliptic curve, ie that the curve remains non-singular under 'reduction mod2'.

The curve takes the homogeneous form (remember that in characterstic 2, $-x = x$, so that we do not need to worry about sign):

$$F(X,Y,Z) \equiv Y^2Z + YZ^2 + X^3 + XZ^2 = 0.$$

Hence

$$\frac{\partial F}{\partial X} = X^2 + Z^2,$$
$$\frac{\partial F}{\partial Y} = Z^2,$$
$$\frac{\partial F}{\partial Z} = Y^2.$$

Thus at a singular point, $Y = Z = 0$, ie the point would be $[1,0,0]$, which is not on the curve.

The projective plane $\mathbb{P}^2(\mathcal{F}_2)$ contains just 7 points: 4 points in the affine plane $\mathcal{F}_2^2$, and 3 points on the line at infinity. (In general, the projective plane $\mathbb{P}^2(\mathcal{F}_q)$, over a finite field with q elements, contains $q^2 + q + 1$ points.

It is trivial to see that $\mathcal{E}(\mathcal{F}_2)$ contains just 5 points: all 4 affine points $(0,0), (0,1), (1,0), (1,1)$ together with the point $O = [0,1,0]$ at infinity.

The only abelian (or non-abelian) group with 5 elements is the cyclic group of order 5. Thus

$$\mathcal{E}(\mathcal{F}_2) = C_5.$$

As an exercise, verify that if $P = (0,0)$ then $5P = 0$.

3.3 Change of origin

It is perhaps worth noting that we can choose any element in an abelian group A as neutral or zero element. More precisely, if $a \in A$ then we can define a new group operation on A by

$$x \dagger y = x + y - a.$$

This operation is evidently commutative; and it is associative, since

$$(x \dagger y) \dagger z = x + y + z - 2a = x \dagger (y \dagger z).$$

The element a acts as new zero element, since

$$x \dagger a = x + a - a = x;$$

while x has inverse $2a - x$ since

$$x \dagger (2a - x) = a,$$

which is now the neutral element.

Thus we could have taken any point $A \in \mathcal{E}(k)$ on the elliptic curve as zero element. However, unless A is a point of inflection we must lose the geometric property that

$$P + Q + R = 0 \iff P, Q, R \text{ are collinear.}$$

In fact, if P, Q, R are collinear then

$$P \dagger Q \dagger R = P + Q + R - 2A = -2A,$$

so

$$P \dagger Q \dagger R = A \iff 3A = 0 \iff A \text{ is a point of inflection.}$$

As we shall see, an elliptic curve can have up to 9 points of inflection. But in general the curve

$$\mathcal{E}(k) : y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

has just one point of inflection: $O = [0, 1, 0]$.

Chapter 4

The Associative Law

Theorem 4.1 *The addition*

$$P + Q = O * (P * Q)$$

on the elliptic curve

$$\mathcal{E}(k) : y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

is associative:

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

Proof ► We have

$$P + (Q + R) = O * (P * (Q + R)), \; (P + Q) + R = O * ((P + Q) * R).$$

Since

$$O * (O * P) = P,$$

it follows that

$$O * A = O * B \iff A = B.$$

Thus it is sufficient to show that

$$P * (Q + R) = (P + Q) * R,$$

ie

$$P * (O * (Q * R)) = (O * (P * Q) * R.$$

Lemma 3 *The associative law holds if and only if*

$$(P * Q) * (R * S) = (P * R) * (Q * S)$$

for any four points $P, Q, R, S \in \mathcal{E}(k)$.

Proof of Lemma ▷ Suppose the associative law holds, so that $\mathcal{E}(k)$ is an additive group. Recall that

$$P * Q = -(P + Q).$$

Thus

$$\begin{aligned}(P * Q) * (R * S) &= -\left((P * Q) + (R * S)\right)\\ &= -\left(-(P + Q) - (R + S)\right)\\ &= (P + Q) + (R + S).\end{aligned}$$

Similarly,

$$\begin{aligned}(P * R) * (Q * S) &= (P + R) + (Q + S)\\ &= (P + Q) + (R + S)\\ &= (P * Q) * (R * S).\end{aligned}$$

Conversely, if this relation holds for all P, Q, R, S then in particular, on setting $P = O$,

which as we have seen is equivalent to the associative law. ◁

This reduces the theorem to a rather complicated geometric result, involving 10 points on the curve:

$$\begin{gathered}X_1 = P,\ X_2 = Q,\ X_3 = R,\ X_4 = S,\\ X_5 = P * Q,\ X_6 = R * S,\ X_7 = P * R,\ X_8 = Q * S,\\ X_9 = X_5 * X_6,\ X_{10} = X_7 * X_8.\end{gathered}$$

The following are collinear:

$$\ell_1 = X_1X_2X_5,\ \ell_2 = X_3X_4X_6,\ \ell_3 = X_1X_3X_7,\ \ell_4 = X_2X_4X_8,\ \ell_5 = X_5X_6X_9,\ \ell_6 = X_7X_8X_{10}.$$

We have to show that

$$X_9 = X_{10}.$$

We shall establish this identity for any non-singular cubic curve.

The basic idea is to use *pencils of cubics.* Suppose

$$\Gamma_1 : F_1(X, Y, Z) = 0,\ \Gamma_2 : F_2(X, Y, Z) = 0,$$

are two cubic curves. By the *pencil* defined by Γ_1, Γ_2 we mean the family of cubic curves

$$\Gamma_{r,s} : rF_1(X, Y, Z) + sF_2(X, Y, Z) = 0.$$

This is a *one-dimensional* pencil, since each cubic in the family is determined by the ratio $[r, s]$. More generally, we can consider two-dimensional pencils

$$\Gamma_{r,s,t} : rF_1(X, Y, Z) + sF_2(X, Y, Z) + tF_3(X, Y, Z) = 0,$$

etc.

Note that a general cubic Γ (we are not concerned with singularity or non-singularity for the moment) is defined by 10 coefficients:

$$\Gamma : a_1X^3{+}a_2X^2Y{+}a_3X^2Za_4XY^2{+}a_5XYZ{+}a_6XZ^2{+}a_7Y^3{+}a_8Y^2Z{+}a_9YZ^2{+}a_{10}Z^3 = 0.$$

The cubic is unchanged if we multiply all the cubics by the same scalar $\rho \in k^\times$, so we may say that the cubics form *a projective space of dimension 9.*

We can always find a cubic passing through any 9 points, since m simultaneous homogeneous linear equations in $n > m$ unknowns always have a non-zero solution.

In general there will be just one such cubic; but there may well be more than one for some sets of 9 points.

Note that three lines ℓ, m, n define a cubic

$$\Gamma = \ell mn.$$

So our pencil could perfectly well consist of cubics

$$\Gamma_{r,s} = r\ell_1m_1n_1 + s\ell_2m_2n_2,$$

where $\ell_1, m_1, n_1, \ell_2, m_2, n_2$ are 6 lines.

◀

Chapter 5

The p-adic Case

5.1 The p-adic valuation on $\mathbb{Q}$

The absolute value $|x|$ on $\mathbb{Q}$ defines the metric, or distance function,

$$d(x, y) = |x - y|.$$

Surprisingly perhaps, there are other metrics on $\mathbb{Q}$ just as worthy of study.

Definition 5.1 *Let p be a prime. Suppose*

$$x = \frac{m}{n} \in \mathbb{Q},$$

where $m, n \in \mathbb{Z}$ with $\gcd(m, n) = 1$. Then we set

$$\|x\|_p = \begin{cases} 0 & \text{if } x = 0, \\ p^{-e} & \text{if } p^e \parallel m, \\ p^e & \text{if } p^e \parallel n. \end{cases}$$

We call the function $x \mapsto \|x\|_p$ the p-adic valuation *on $\mathbb{Q}$.*

Another way of putting this is: If $x \in \mathbb{Q}$, $x \neq 0$, then we can write

$$x = \frac{m}{n} p^e$$

where $p \nmid m, n$. The p-adic value of x is given by

$$\|x\|_p = p^{-e}.$$

Note that all integers are quite small in the p-adic valuation:

$$x \in \mathbb{Z} \Longrightarrow \|x\|_p \leq 1.$$

High powers of p are very small:

$$p^n \to 0 \text{ as } n \to \infty.$$

The following result is immediate.

Proposition 5.1 *1. $\|x\|_p \geq 0$; and $\|x\|_p = 0 \Longleftrightarrow x = 0$;*

2. $\|xy\|_p = \|x\|_p \|y\|_p$;

3. $\|x + y\|_p \leq \max(\|x\|_p, \|y\|_p)$.

From (3) we at once deduce

Corollary 1 *The p-adic valuation satisfies the triangle inequality:*

3' $\|x + y\|_p \leq \|x\|_p + \|y\|_p$.

A *valuation* on a field k is a map

$$x \mapsto \|x\| : k \to \mathbb{R}$$

satisfying (1), (2) and (3'). A valuation defines a *metric*

$$d(x, y) = \|x - y\|$$

on k; and this in turn defines a *topology* on k.

Corollary 2 *If $\|x\|_p \neq \|y\|_p$ then*

$$\|x + y\|_p = \max(\|x\|_p, \|y\|_p).$$

Corollary 3 *In a p-adic equation*

$$x_1 + \cdots + x_n = 0 \qquad (x_1, \ldots, x_n \in \mathbb{Q}_p)$$

no term can dominate, ie at least two of the x_i must attain $\max \|x_i\|_p$.

To emphasize the analogy between the p-adic valuation and the familiar valuation $|x|$ we sometimes write

$$\|x\|_\infty = |x|.$$

5.2 p-adic numbers

The reals $\mathbb{R}$ can be constructed from the rationals $\mathbb{Q}$ by *completing* the latter with respect to the valuation $|x|$. In this construction each Cauchy sequence

$$\{x_i \in \mathbb{Q} : |x_i - x_j| \to 0 \text{ as } i, j \to \infty\}$$

defines a real number, with 2 sequences defining the same number if $|x_i - y_i| \to 0$.

(There are 2 very different ways of constructing $\mathbb{R}$ from $\mathbb{Q}$: by completing $\mathbb{Q}$, as above; or alternatively, by the use of *Dedekind sections.* In this each real number corresponds to a partition of $\mathbb{Q}$ into 2 subsets L, R where

$$l \in L,\ r \in R \Longrightarrow l < r.$$

The construction by completion is much more general, since it applies to any metric space; while the alternative construction uses the fact that $\mathbb{Q}$ is an *ordered* field. John Conway, in *On Numbers and Games,* has generalized Dedekind sections to give an extraordinary construction of rationals, reals and infinite and infinitesimal numbers, starting 'from nothing'. Knuth has given a popular account of Conway numbers in *Surreal Numbers.*)

We can complete $\mathbb{Q}$ with respect to the p-adic valuation in just the same way. The resulting field is called *the field of p-adic numbers*, and is denoted by $\mathbb{Q}_p$. We can identify $x \in \mathbb{Q}$ with the Cauchy sequence $(x, x, x, \ldots)$. Thus

$$\mathbb{Q} \subset \mathbb{Q}_p.$$

To bring out the parallel with the reals, we sometimes write

$$\mathbb{R} = \mathbb{Q}_\infty.$$

The numbers $x \in \mathbb{Q}_p$ with $\|x\|_p \leq 1$ are called *p-adic integers.* The p-adic integers form a ring, denoted by $\mathbb{Z}_p$. For if $x, y \in \mathbb{Z}_p$ then by property (3) above,

$$\|x + y\|_p \leq \max(\|x\|_p, \|y\|_p) \leq 1,$$

and so $x + y \in \mathbb{Z}_p$. Similarly, by property (1),

$$\|xy\|_p = \|x\|_p\|y\|_p \leq 1,$$

and so $xy \in \mathbb{Z}_p$.

Evidently

$$\mathbb{Z} \subset \mathbb{Z}_p.$$

More generally,

$$x = \frac{m}{n} \in \mathbb{Z}_p$$

if $p \not| n$. (We sometimes say that a rational number x of this form is *p-integral.*) In other words,

$$\mathbb{Q} \cap \mathbb{Z}_p = \{\frac{m}{n} : p \not| n\}.$$

Evidently the p-integral numbers form a sub-ring of $\mathbb{Q}$.

Concretely, each element $x \in \mathbb{Z}_p$ is uniquely expressible in the form

$$x = c_0 + c_1 p + c_2 p^2 + \cdots \quad (0 \leq c_i < p).$$

More generally, each element $x \in \mathbb{Q}_p$ is uniquely expressible in the form

$$x = c_{-i}p^{-i} + c_{-i+1}p^{-i+1} + \cdots + c_0 + c_1 p + \cdots \quad (0 \leq c_i < p).$$

We can think of this as the p-adic analogue of the decimal expansion of a real number $x \in \mathbb{R}$.

Suppose for example $p = 3$. Let us express $1/2 \in \mathbb{Q}_3$ in standard form. The first step is to determine if

$$\frac{1}{2} \equiv 0, 1 \text{ or } 2 \bmod 3.$$

In fact $2^2 \equiv 1 \bmod 3$; and so

$$\frac{1}{2} \equiv 2 \bmod 3.$$

Next

$$\frac{1}{3}\left(\frac{1}{2} - 2\right) = -\frac{1}{2} \equiv 1 \bmod 3$$

ie

$$\frac{1}{2} - 2 \equiv 1 \cdot 3 \bmod 3^2.$$

Thus

$$\frac{1}{2} \equiv 2 + 1 \cdot 3 \bmod 3^2$$

For the next step,

$$\frac{1}{3}\left(-\frac{1}{2}-1\right) = -\frac{1}{2} \equiv 1 \bmod 3$$

giving

$$\frac{1}{2} \equiv 2 + 1\cdot 3 + 1\cdot 3^2 \bmod 3^3$$

It is clear that this pattern will be repeated indefinitely. Thus

$$\frac{1}{2} = 2 + 3 + 3^2 + 3^3 + \cdots .$$

To check this,

$$\begin{aligned} 2 + 3 + 3^2 + \cdots &= 1 + (1 + 3 + 3^2 + \cdots) \\ &= 1 + \frac{1}{1-3} \\ &= 1 - \frac{1}{2} \\ &= \frac{1}{2}. \end{aligned}$$

As another illustration, let us expand $3/5 \in \mathbb{Q}_7$. We have

$$\begin{aligned} \frac{3}{5} &\equiv 2 \bmod 7 \\ \frac{1}{7}\left(\frac{3}{5}-2\right) = -\frac{1}{5} &\equiv 4 \bmod 7 \\ \frac{1}{7}\left(-\frac{1}{5}-4\right) = -\frac{3}{5} &\equiv 5 \bmod 7 \\ \frac{1}{7}\left(-\frac{3}{5}-5\right) = -\frac{4}{5} &\equiv 2 \bmod 7 \\ \frac{1}{7}\left(-\frac{4}{5}-2\right) = -\frac{2}{5} &\equiv 1 \bmod 7 \\ \frac{1}{7}\left(-\frac{2}{5}-1\right) = -\frac{1}{5} &\equiv 4 \bmod 7 \end{aligned}$$

We have entered a loop; and so (in $\mathbb{Q}_7$)

$$\frac{3}{5} = 2 + 4\cdot 7 + 5\cdot 7^2 + 2\cdot 7^3 + 1\cdot 7^4 + 4\cdot 7^5 + 5\cdot 7^6 + \cdots$$

Checking,

$$\begin{aligned} 1 + \left(1 + 4\cdot 7 + 5\cdot 7^2 + 2\cdot 7\right)\frac{1}{1-7^4} &= 1 - \frac{960}{2400} \\ &= 1 - \frac{2}{5} \\ &= \frac{3}{5}. \end{aligned}$$

It is not difficult to see that a number $x \in \mathbb{Q}_p$ has a recurring p-adic expansion if and only if it is rational (as is true of decimals).

Let $x \in \mathbb{Z}_p$. Suppose $\|x\|_p = 1$. Then

$$x = c + yp,$$

where $0 < c < p$ and $y \in \mathbb{Z}_p$. Suppose first that $c = 1$, ie

$$x = 1 + yp.$$

Then x is invertible in $\mathbb{Z}_p$, with

$$x^{-1} = 1 - yp + y^2p^2 - y^3p^3 + \cdots .$$

Even if $c \neq 1$ we can find d such that

$$dc \equiv 1 \bmod p.$$

Then

$$dx \equiv dc \equiv 1 \bmod p,$$

say

$$dx = 1 + py,$$

and so x is again invertible in $\mathbb{Z}_p$, with

$$x^{-1} = d\left(1 - yp + y^2p^2 - \cdots\right).$$

Thus the elements $x \in \mathbb{Z}_p$ with $\|x\|_p = 1$ are all *units* in $\mathbb{Z}_p$, ie they have inverses in $\mathbb{Z}_p$; and all such units are of this form. These units form the multiplicative group

$$\mathbb{Z}_p^\times = \{x \in \mathbb{Z}_p : \|x\|_p = 1\}.$$

5.3 In the p-adic neighbourhood of 0

Recall that an elliptic curve $\mathcal{E}(k)$ can be brought to Weierstrassian form

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

if and only if it has a flex defined over k. This is not in general true for elliptic curves over $\mathbb{Q}_p$. For example, the curve

$$X^3 + pY^3 + p^2Z^3 = 0$$

has no points at all (let alone flexes) defined over $\mathbb{Q}_p$. For if $[X, Y, Z]$ were a point on this curve then

$$\|X^3\|_p = p^{3e}, \ \|pY^3\|_p = p^{3f-1}, \ \|p^2Z^3\|_p = p^{3g-2}$$

for some integers e, f, g. But if $a, b, c \in \mathbb{Q}_p$ and

$$a + b + c = 0$$

then two (at least) of a, b, c must have the same p-adic value, by Corollary 3 to Proposition 5.1.

On the other hand, $\mathbb{Q}_p$ is of characteristic 0; so if $\mathcal{E}(\mathbb{Q}_p)$ *is* Weierstrassian — as we shall always assume, for reasons given earlier — then it can be brought to standard form

$$y^2 = x^3 + bx + c.$$

In spite of this, there is some advantage in working with the general Weierstrassian equation, since — as we shall see in Chapter 6 — this allows us to apply the results of this Chapter to study the integer points (that is, points with integer coordinates) on elliptic curves over $\mathbb{Q}$ given in general Weierstrassian form. Such an equation over $\mathbb{Q}$ can of course be reduced to standard form; but the reduction may well transform integer to non-integer points.

As in the real case, we study the curve in the neighbourhood of $0 = [0, 1, 0]$ by taking coordinates X, Z, where

$$(X, Z) = [X, 1, Z].$$

In these coordinates the elliptic curve takes the form

$$\mathcal{E}(\mathbb{Q}_p) : Z + c_1XZ + c_3Z^2 = X^3 + c_2X^2Z + c_4XZ^2 + c_6Z^3.$$

As in the real case, if $Z(P)$ is small then so is $X(P)$.

Proposition 5.2 *If $P \in \mathcal{E}(\mathbb{Q}_p)$ then*

$$\|Z\|_p < 1 \Longrightarrow \|X\|_p < 1;$$

and if this is so then

$$\|Z\|_p = \|X\|_p^3.$$

Proof ► Suppose $\|Z\|_p < 1$. Let

$$\|X\|_p = p^e.$$

If $e \geq 0$ then X^3 will dominate; no other term can be as large, p-adically speaking.

Thus $e < 0$, ie $\|X\|_p < 1$; and now each term

$$\|c_1XZ\|_p, \|c_3Z^2\|_p, \|c_2X^2Z\|_p, \|c_4XZ^2\|_p, \|c_6XZ\|_p < \|Z\|_p.$$

Only X^3 is left to balance Z. Hence

$$\|Z\|_p = \|X^3\|_p = \|X\|_p^3.$$

◄

Definition 5.2 *For each $e > 0$ we set*

$$\mathcal{E}_{(p^e)} = \{(X, Z) \in \mathcal{E} : \|X\|_p \leq p^{-e}, \ \|Z\|_p \leq p^{-3e}\}.$$

Recall that in the real case, we showed that Z could be expressed as a power-series in X,

$$Z = X^3 - c_1X^4 + (c_1^2 + c_2)X^5 + \cdots .$$

valid in a neighbourhood of $O = [0, 1, 0]$. It follows that

$$F(X, Z(X)) = 0$$

identically, where

$$F(X, Z) = Z + c_1XZ + c_3Z^2 - (X^3 + c_2X^2Z + c_4XZ^2 + c_6Z^3).$$

This identity must hold in any field, in particular in $\mathbb{Q}_p$.

Note that in the p-adic case, convergence is much simpler than in the real case. A series in $\mathbb{Q}_p$ converges if and only if its terms tend to 0:

$$\sum a_r \text{ convergent } \iff a_r \to 0.$$

Remember too that in the p-adic valuation integers are *small*,

$$x \in \mathbb{Z} \Longrightarrow \|x\|_p \leq 1.$$

Thus a power-series

$$a_0 + a_1 x + a_2 x^2 + \cdots$$

where $a_i \in \mathbb{Z}$—or more generally, $a_i \in \mathbb{Z}_p$—will converge for all x with $\|x\|_p < 1$.

Proposition 5.3 *Suppose $\|Z\|_p < 1$. Then we can express Z as a power-series in X,*

$$Z = X^3 + a_1 X^4 + a_2 X^5 + \cdots$$

where

1. *$a_1 = -c_1,\ a_2 = c_1^2 + c_2,\ c_3 = -(c_1^3 + 2c_1c_3 + c_3)$;*
2. *each coefficient a_i is a polynomial in c_1, c_2, c_3, c_4, c_6 with integer coefficients;*
3. *the coefficient a_i has weight i, given that c_i is ascribed weight i for $(i = 1 - 4, 6$.*

Proof ▶ By repeatedly substituting for Z on the right-hand side of the equation

$$Z = X^3 + c_2X^2Z + c_4XZ^2 + c_6Z^3 - (c_1XZ + c_3Z^2)$$

we can successively determine more and more terms in the power series. Thus suppose we have shown that

$$Z = X^3\left(1 + a_1X + \cdots + a_{n-1}X^{n-1}\right).$$

On substituting for Z on the right-hand side of the equation and comparing coefficients of X^{n+3},

$$a_n = c_2a_{n-2} + c_4 \sum_{i+j=n-4} a_ia_j + c_6 \sum_{i+j+k=n-6} a_ia_ja_k - c_1a_{n-1} - c_3 \sum_{i+j=n-3} a_ia_j,$$

from which the result follows. ◀

Corollary *If the elliptic curve is given in standard form*

$$y^2 = x^3 + ax^2 + bx + c$$

then

$$Z = x^3 + d_2X^5 + d_4X^7 + \cdots,$$

where

1. *only odd powers of X appear, ie $d_i = 0$ for i odd;*
2. *$d_2 = a,\ d_4 = a^2 + b,\ d_6 = a^3 + 3ab + c$;*
3. *each coefficient d_{2i} is a polynomial in a, b, c with integer coefficients;*
4. *the coefficient d_{2i} has weight i, given that a, b, c are ascribed weights 2,4,6 respectively;*

Proof ▶ We note that in the standard case the (X, Z)-equation

$$Z = X^3 + aX^2Z + bXZ^2 + cZ^3$$

is invariant under the reflection $(X, Z) \mapsto (-X, -Z)$ (corresponding to $P \mapsto -P$). Thus

$$Z(-X) = -Z(X),$$

from which the absence of terms of even degree X^{2i} follows. ◀

As in the real case, the sum of 2 points near O is defined by a function $S(X_1, X_2)$, where

$$X(P_1 + P_2) = S(X(P_1), X(P_2)).$$

Proposition 5.4 *Suppose $\|X_1\|_p, \|X_2\|_p < 1$. Then we can express $S(X_1, X_2)$ as a double power-series in X_1, X_2,*

$$\begin{aligned} S(X_1, X_2) &= X_1 + X_2 + c_1X_1X_2 + \cdots \\ &= \sum_i S_i(X_1, X_2) \\ &= \sum_{i,j} s_{ij} X_1^i X_2^j \end{aligned}$$

where

1. *$S_i(X_1, X_2)$ is a symmetric polynomial in X_1, X_2 of degree i;*
2. *$S_1(X_1, X_2) = X_1 + X_2,\ S_2(X_1, X_2) = c_1X_1X_2$;*
3. *the coefficient s_{jk} of X^jX^k is a polynomial in c_1, c_2, c_3, c_4, c_6 with integral coefficients.*
4. *all the coefficients in $S_i(X_1, X_2)$ have weight i.*

Proof ► As in the real case, let the line

$$P_1P_2: \; Z = MX + D$$

meet $\mathcal{E}$ again in $P_3 = (X_3, Z_3)$, ie

$$P_3 = P_1 * P_2.$$

Then X_1, X_2, X_3 are the roots of the equation

$$\begin{aligned} X^3 + c_2X^2(MX+D) + c_4X(MX+D)^2 + c_6(MX+D)^3 \\ - (MX+D) - c_1X(MX+D) - c_3(MX+D)^2 = 0. \end{aligned}$$

Hence

$$\begin{aligned} X_1 + X_2 + X_3 &= -\frac{\text{coeff of } X^2}{\text{coeff of } X^3} \\ &= \frac{c_1M + 2c_3M^2 - (c_2 + c_4M + c_6M^2)D}{1 + c_2M + c_4M^2 + c_6M^3} \end{aligned}$$

Now

$$\begin{aligned} M &= \frac{Z_2 - Z_1}{X_2 - X_1} \\ &= \frac{X_2^3 - X_1^3}{X_2 - X_1} - c_1\frac{X_2^4 - X_1^4}{X_2 - X_1} + \cdots \\ &= X_1^2 + X_1X_2 + X_2^2 - c_1(X_1^3 + X_1^2X_2 + X_1X_2^2 + X_2^3) + \cdots, \\ D &= \frac{X_2Z_1 - X_1Z_2}{X_2 - X_1} \\ &= X_1X_2\left(\frac{X_2^2 - X_1^2}{X_2 - X_1} - c_1\frac{X_2^3 - X_1^3}{X_2 - X_1} + \cdots\right) \\ &= X_1X_2\left(X_1 + X_2 - c_1(X_2^2 + X_1X_2 + X_2^2) + \cdots\right). \end{aligned}$$

Thus M, D are both expressible as symmetric power-series in X_1, X_2; and

$$\|M\|_p \le p^{-2}, \; \|D\|_p \le p^{-3},$$

or more precisely,

$$\begin{aligned} M &\equiv X_1^2 + X_1X_2 + X_2^2 \bmod p^3 \\ D &\equiv X_1X_2(X_1 + X_2) \bmod p^4. \end{aligned}$$

Hence

$$X_1 + X_2 + X_3 \equiv 0 \bmod p^2.$$

More precisely,

$$X_1 + X_2 + X_3 \equiv c_1(X_1^2 + X_1X_2 + X_2^2) \bmod p^3,$$

ie

$$X_3 \equiv -(X_1 + X_2) + c_1(X_1^2 + X_1X_2 + X_2^2) \bmod p^3.$$

In particular,

$$\|X_3\|_p \le p^{-1},$$

and so

$$\|Z_3\|_p = \|MX_3 + D\| \le p^{-3},$$

ie

$$P_1, P_2 \in \mathcal{E}_{(p)} \Longrightarrow P_3 \in \mathcal{E}_{(p)}.$$

Recall that

$$P_1 + P_2 = O * (P_1 * P_2) = O * P_3.$$

By our formulae above, with O, X_3 in place of X_1, X_2,

$$X(O * P_3) \equiv = -X_3 \bmod p^2,$$

or more precisely

$$X(O * P_3) \equiv = -X_3 + c_1X_3^2 \bmod p^3,$$

Hence

$$X(P_1 + P_2) = X_1 + X_2 \bmod p^2,$$

or more precisely

$$\begin{aligned} X(P_1 + P_2) &= X_1 + X_2 - c_1(X_1^2 + X_1X_2 + X_2^2) + c_1(X_1 + X_2)^2 \bmod p^3 \\ &= X_1 + X_2 + c_1X_1X_2 \bmod p^3 \end{aligned}$$

◀

Finally, we turn to the normal coordinate function $\theta(X)$, defined as in the real case by

$$\begin{aligned}\frac{d\theta}{dX} &= \frac{1}{\partial F/\partial Z}\\ &= \frac{1}{1 + c_1X + 2c_3Z - c_2X^2 - 2c_4XZ - 3c_6Z^2}\end{aligned}$$

Proposition 5.5 *Suppose $\|X\|_p < 1$. Then we can express θ as a power-series in X,*

$$\begin{aligned}\theta &= X + \frac{c}{2}X^2 + \cdots\\ &= \sum t_n X^{n+1}\end{aligned}$$

where

1. *$t_1 = 1$, $t_2 = -c_1/2$;*
2. *for each i, it_i is a polynomial in c_1, c_2, c_3, c_4, c_6 with integral coefficients;*
3. *t_i is of weight i.*

Proof ► Since

$$\begin{aligned}\frac{d\theta}{dX} &= \frac{1}{1 + c_1X + 2c_3Z - c_2X^2 - 2c_4XZ - 3c_6Z^2}\\ &= 1 - (c_1X + 2c_3Z - c_2X^2 - 2c_4XZ - 3c_6Z^2)\\ &\quad + (c_1X + 2c_3Z - c_2X^2 - 2c_4XZ - 3c_6Z^2)^2 + \cdots\end{aligned}$$

the coefficients in the power-series for $d\theta/dX$ are integral polynomials in the c_i. It follows on integration that the coefficients t_i in the power-series for $\theta(X)$ have at worst denominator i.

It remains to show that this power series converges for $\|X\|_p < 1$.

Lemma 4 *For all i,*

$$\|1/i\|_p \le i.$$

Proof of Lemma ▷ Suppose

$$\|i\|_p = p^{-e}.$$

Then

$$\begin{aligned}p^e \mid i &\Longrightarrow p^e \le i\\ &\Longrightarrow \|1/i\| \le i.\end{aligned}$$

◁

If now $\|X\|_p < 1$ then

$$\|X\|_p \le \frac{1}{p};$$

and so

$$\|t_i X^i\|_p \le \frac{i}{p^i},$$

which tends to 0 as $i \to \infty$. The power-series is therefore convergent. ◀

Note that

$$p^i \ge 2^i = (1+1)^i > i^2/2$$

if $i \ge 2$, while if p is odd, $\|1/2\|_p = 1$. Thus

$$\begin{aligned}\|X\|_p \le p^{-1} &\Longrightarrow \|X^i/i\|_p \le p^{-2} \text{ for } i \ge 2 \qquad (p \text{ odd})\\ \|X\|_2 \le 2^{-2} &\Longrightarrow \|X^i/i\|_2 \le 2^{-3} \text{ for } i \ge 2 \qquad (p = 2).\end{aligned}$$

So if p is odd,

$$\theta(X) = X + O(p^2) \text{ if } \|X\|_p \le p^{-1};$$

while if $p = 2$,

$$\theta(X) = X + O(2^3) \text{ if } \|X\|_2 \le 2^{-2}.$$

That is why in our discussion below the argument often applies to $P \in \mathcal{E}_{(p)}$ if p is odd, while if $p = 2$ we have to restrict P to $\mathcal{E}_{2^2}$.

Theorem 5.1 *For each power p^e, where $e \ge 1$,*

$$\mathcal{E}_{(p^e)}(\mathbb{Q}_p)$$

is a subgroup of $\mathcal{E}(\mathbb{Q}_p)$. Moreover the map

$$\theta : \mathcal{E}_{(p^e)}(\mathbb{Q}_p) \to p^e\mathbb{Z}_p$$

is an isomorphism (of topological abelian groups), provided $e \ge 2$ if $p = 2$.

Proof ► The identity

$$\theta(S(X_1, X_2) = \theta(X_1) + \theta(X_2),$$

which we established in the real case, must still hold; and we conclude from it, as before, that

$$\theta(P_1 + P_2) = \theta(P_1) + \theta(P_2)$$

whenever

$$P_1, P_2 \in \mathcal{E}_{(p^e)}(\mathbb{Q}_p).$$

It follows from this that $\mathcal{E}_{(p^e)}$ is a subgroup; and that

$$\theta : \mathcal{E}_{(p^e)} \to p^e\mathbb{Z}_p$$

is a homomorphism, provided $e \geq 2$ if $p = 2$.

Since

$$\theta(X) = X - c_1X^2/2 + \cdots,$$

we have

$$\|\theta(X)\|_p = \|X\|_p$$

for all $\|X\|_p \leq p^{-e}$. In particular

$$\theta(X) = 0 \Longleftrightarrow X = 0.$$

Hence θ is injective.

It is also surjective, as the following Lemma will show.

Lemma 5 *The only closed subgroups of $\mathbb{Z}_p$ are the subgroups*

$$p^n\mathbb{Z}_p \quad (n = 0, 1, 2, \dots),$$

together with $\{0\}$. In particular, every closed subgroup of $\mathbb{Z}_p$, apart from $\{0\}$, is in fact open.

Proof of Lemma ▷ $\mathbb{Z}$ is a dense subset of $\mathbb{Z}_p$:

$$\overline{\mathbb{Z}} = \mathbb{Z}_p.$$

For the p-adic integer

$$x = c_0 + c_1p + c_2p^2 + \cdots \quad (c_i \in \{0, 1, \dots, p-1\})$$

is approached arbitrarily closely by the (rational) integers

$$x_r = c_0 + c_1p + \cdots + c_rp^r.$$

Now suppose S is a closed subgroup of $\mathbb{Z}_p$. Let $s \in S$ be an element of maximal p-adic valuation, say

$$\|s\| = p^{-e}.$$

Then

$$s = p^eu$$

where u is a unit in $\mathbb{Z}_p$, with inverse v, say. Given any $\epsilon > 0$, we can find $n \in \mathbb{Z}$ such that

$$\|v - n\| < \epsilon.$$

Then

$$\begin{aligned} ns - p^e &= p^e(nu - 1) \\ &= p^eu(n - v); \end{aligned}$$

and so

$$\|ns - p^e\| < \epsilon.$$

Since $ns \in S$ and S is closed, it follows that

$$p^e \in S.$$

Hence

$$p^e\overline{\mathbb{Z}} = p^e\mathbb{Z}_p \subset S.$$

Since s was a maximal element in S, it follows that

$$S = p^e\mathbb{Z}_p.$$

◁

It follows from this Lemma that $\operatorname{im}\theta$ is one of the subgroups $p^m\mathbb{Z}_p$. But since

$$\|X\| = p^{-e} \Longrightarrow \|\theta(X)\| = p^{-e},$$

$\operatorname{im}\theta$ must in fact be $p^e\mathbb{Z}_p$, ie θ is surjective.

A continuous bijective map from a compact space to a hausdorff space is necessarily a homeomorphism. (This follows from the fact that the image of every closed, and therefore compact, subset is compact, and therefore closed.) In particular, θ establishes an isomorphism

$$\mathcal{E}_{(p^e)} \cong p^e\mathbb{Z}_p \cong \mathbb{Z}_p.$$

◄

It follows from this Theorem that $\mathcal{E}_{(p^e)}$ is torsion-free, since $\mathbb{Z}_p$ is torsion-free. Thus *there are no points of finite order on $\mathcal{E}$ close to O*, a result which we shall exploit in the next Chapter.

5.4 The Structure of $\mathcal{E}(\mathbb{Q}_p)$

We shall not use the following result, but include it for the sake of completeness.

Theorem 5.2 *Let $\mathcal{F} \subset \mathcal{E}(\mathbb{Q}_p)$ be the torsion subgroup of the elliptic curve $\mathcal{E}(\mathbb{Q}_p)$. Then*

$$\mathcal{E}(\mathbb{Q}_p) \cong \mathcal{F} \oplus \mathbb{Z}_p.$$

Proof ► The torsion subgroup $\mathcal{F}$ splits (uniquely) into its p-component $\mathcal{F}_p$ and the sum $\mathcal{F}_{p'}$ of all components $\mathcal{F}_q$ with $q \neq p$:

$$\mathcal{F} = \mathcal{F}_p \oplus \mathcal{F}_{p'}.$$

(See Appendix A for details.) Explicitly,

$$\mathcal{F}_p = \{P \in \mathcal{E} : p^n P = 0 \text{ for some } n\},$$
$$\mathcal{F}_{p'} = \{P \in \mathcal{E} : mP = 0 \text{ for some } d \text{ with } \gcd(m,p) = 1\}.$$

(We write $\mathcal{E}$ for $\mathcal{E}(\mathbb{Q}_p)$).

We also set

$$\mathcal{E}_p = \{P \in \mathcal{E} : p^n P \to O \text{ as } n \to \infty\}.$$

Evidently

$$\mathcal{E}_p \supset \mathcal{E}_{(p)}.$$

Since $E_{(p)}$ is an open (and therefore closed) subgroup of $\mathcal{E}$, it follows that the same is true of $\mathcal{E}_p$.

Lemma 6 *$p^n \mathcal{E}_p = \mathcal{E}_{(p^e)}$ for some $n, e > 0$.*

Proof of Lemma ▷ For each $P \in \mathcal{E}_p$,

$$p^n P \in \mathcal{E}_{(p)}$$

for some $n > 0$ since $p^n P \to O$ and $\mathcal{E}_{(p)}$ is an open neighbourhood of O. Hence the open subgroups $p^{-n}\mathcal{E}_{(p)}$ cover $\mathcal{E}_p$. Since $\mathcal{E}_p$ is compact, it follows that $p^{-n}\mathcal{E}_{(p)} \supset \mathcal{E}_p$ for some n, ie

$$p^n \mathcal{E}_p \subset \mathcal{E}_{(p)} \cong \mathbb{Z}_p.$$

But by Lemma 5 to Theorem 5.1, the only closed subgroups of $\mathbb{Z}_p$ are the $p^e Z_p$, which correspond under this isomorphism to the subgroups $\mathcal{E}_{(p^e)}$ of $\mathcal{E}_{(p)}$.

We conclude that

$$p^n \mathcal{E}_p = \mathcal{E}_{(p^e)}$$

for some e. ◁

Lemma 7 *Suppose A is a finite p-group; and suppose $\gcd(m,p) = 1$. Then the map $\psi : A \to A$ under which*

$$a \mapsto ma$$

is an isomorphism.

Proof of Lemma ▷ Suppose $a \in \ker A$, ie

$$ma = 0.$$

Then $\operatorname{order}(a) \mid m$. But by Lagrange's Theorem, $\operatorname{order}(a) = p^e$ for some e. Hence $order(a) = 1$, ie $a = 0$.

Thus ψ is injective; and it is therefore surjective, by the Pigeon-Hole Principle. Hence ψ is an isomorphism. ◁

It is not difficult to extend this result to $\mathcal{E}_p$, which is in effect a kind of topological p-group.

Lemma 8 *Suppose $\gcd(m,p) = 1$. Then the map $\psi : \mathcal{E}_p \to \mathcal{E}_p$ under which*

$$a \mapsto ma$$

is an isomorphism.

Proof of Lemma ▷ Suppose $P \in \ker \psi$, ie

$$mP = 0.$$

By Lemma 1,

$$p^n \mathcal{E}_p \subset \mathcal{E}_{(p^2)} \cong \mathbb{Z}_p$$

for some n.

But $\mathbb{Z}_p$ is torsion-free. Thus

$$mP = 0 \Longrightarrow m(p^n P = 0) \Longrightarrow p^n P = 0.$$

Hence

$$m, p^n \mid \operatorname{order}(P) \Longrightarrow \operatorname{order}(P) = 1 \Longrightarrow P = 0$$

since $\gcd(m, p^n) = 1$. Thus

$$\ker \psi = 0,$$

ie ψ is injective.

Now suppose $P \in \mathcal{E}_p$. We have to show that $P = mQ$ for some $Q \in \mathcal{E}_p$.

Since $\mathcal{E}_p/p^n\mathcal{E}_p$ is a finite p-group we can find $Q \in \mathcal{E}_p$ such that

$$mQ \equiv P \bmod p^n\mathcal{E}_p$$

ie

$$mQ = P + R,$$

where

$$R \in p^n\mathcal{E}_p \cong \mathbb{Z}_p.$$

Now the map

$$P \mapsto mP : \mathbb{Z}_p \to \mathbb{Z}_p$$

is certainly an isomorphism, since m is a unit in $\mathbb{Z}_p$ with inverse $m^{-1} \in \mathbb{Z}_p$. In particular we can find $S \in p^n\mathcal{E}_p$ with

$$mS = R.$$

Putting all this together,

$$P = mQ + R = mQ + mS = m(Q + S).$$

Thus the map ψ is surjective, and so an isomorphism. $\triangleleft$

Lemma 9 $\mathcal{E}(\mathbb{Q}_p) = \mathcal{F}_{p'} \oplus \mathcal{E}_p$.

Proof of Lemma $\triangleright$ Suppose

$$P \in \mathcal{F}_{p'} \cap \mathcal{E}_p,$$

say

$$mP = O,$$

where $\gcd(m, p) = 1$.

On considering $p \bmod m$ as an element of the finite group

$$(\mathbb{Z}/m)^\times = \{r \bmod m : \gcd(r, m) = 1\},$$

it follows by Lagrange's Theorem that

$$p^r \equiv 1 \bmod m$$

for some $n > 0$. But then

$$p^r P = P;$$

and so

$$p^n P \to O \Longrightarrow P = O.$$

Now suppose $P \in \mathcal{E}$. Since $\mathcal{E}$ is compact, and $\mathcal{E}_p$ is open, $\mathcal{E}/\mathcal{E}_p$ is finite (eg since $\mathcal{E}$ must be covered by a finite number of $\mathcal{E}_p$-cosets). Let the order of this finite group be mp^e, where $\gcd(m, p) = 1$.

We can find $u, v \in \mathbb{Z}$ such that

$$um + vp^e = 1;$$

and then

$$P = Q + R,$$

where

$$Q = u(mP),\ R = v(p^e P).$$

Now

$$p^e Q = u(mp^e P) \in \mathcal{E}_p.$$

Hence

$$p^n Q \to 0 \text{ as } n \to \infty$$

ie

$$Q \in \mathcal{E}_p.$$

On the other hand,

$$mR = v(mp^e P) \in \mathcal{E}_p.$$

Hence by Lemma 8, there is a point $S \in \mathcal{E}_p$ such that

$$mR = mS,$$

and so

$$T = R - S \in \mathcal{F}_{p'}.$$

Putting these results together,

$$P = T + (Q + S),$$

with $T \in \mathcal{F}_{p'}$ and $Q + S \in \mathcal{E}_p$. $\triangleleft$

Lemma 10 $\mathcal{F}_p \subset \mathcal{E}_p$.

Proof of Lemma ▷ Suppose

$$P = Q + R \in \mathcal{F}_p,$$

where $Q \in \mathcal{F}_{p'}$, $R \in \mathcal{E}_p$. Then

$$p^n P = 0 \Longrightarrow p^n Q = 0,\ p^n R = 0,$$

since the sum is direct. But

$$p^n Q = 0 \Longrightarrow \operatorname{order}(Q) \mid p^n \Longrightarrow \operatorname{order}(Q) = 1 \Longrightarrow Q = 0,$$

since the order of Q is coprime to p by the definition of $\mathcal{F}_{p'}$. Thus

$$P = R \in \mathcal{E}_p.$$

◁

It remains to split $\mathcal{E}_p$ into $\mathcal{F}_p$ and a subgroup isomorphic to $\mathbb{Z}_p$.

Consider the surjection

$$\psi : \mathcal{E}_p \to \mathcal{E}_{(p^e)} \cong \mathbb{Z}_p.$$

Let us choose a point

$$P_0 \in \mathcal{E}_{p^e} \setminus \mathcal{E}_{(p^{e+1})},$$

eg if we identify $\mathcal{E}_{(p^e)}$ with $\mathbb{Z}_p$ we might take the point corresponding to $1 \in \mathbb{Z}_p$. Now choose a point P_1 such that

$$\psi(P_1) = P_0;$$

and let

$$\mathcal{E}_1 = \overline{\langle P_1 \rangle}$$

be the closure in $\mathcal{E}_p$ of the subgroup generated by P_1. We shall show that the restriction

$$\psi_1 = \psi \mid \mathcal{E}_1 : \mathcal{E}_1 \to \mathcal{E}_{(p^e)}$$

is an isomorphism, so that

$$\mathcal{E}_1 \cong \mathcal{E}_{(p^e)} \cong \mathbb{Z}_p.$$

Certainly ψ_1 is surjective. For $\mathcal{E}_1$ is compact, and so its image is closed; while $\langle P_0 \rangle >$ is dense in $\mathcal{E}_{(p^e)} \cong \mathbb{Z}_p$.

Suppose

$$Q \in \ker \psi_1 = \ker \psi \cap \mathcal{E}_1.$$

By definition, Q is the limit of points in $\langle P_1 \rangle$, say

$$n_i P_1 \to Q,$$

where $n_i \in \mathbb{Z}$. But then, since ψ is continuous,

$$n_i P_0 \to \psi(Q) = 0.$$

Hence

$$n_i \to 0$$

in $\mathbb{Z}_p$. But then it follows that

$$n_i P_1 \to 0$$

in $\mathcal{E}_p$, since

$$\bigcap p^n E_p = 0.$$

Hence $Q = 0$, ie $\ker \psi_1 = 0$.

It remains to show that

$$\mathcal{E}_p = \mathcal{F}_p \oplus \mathcal{E}_1.$$

Suppose $P \in \mathcal{E}_p$. Then

$$\psi(P) = \psi(Q),$$

for some $Q \in \mathcal{E}_1$. In other words,

$$p^n(P - Q) = 0.$$

Thus

$$R = P - Q \in \mathcal{F}_p$$

On the other hand, if

$$F_p \cap \mathcal{E}_1 = 0,$$

since as we have seen,

$$\mathcal{E}_1 \cong \mathcal{E}_{(p^e)} \cong \mathbb{Z}_p,$$

and Z_p is torsion-free.

We have shown therefore that

$$\begin{aligned} \mathcal{E} &= \mathcal{F}_{p'} \oplus \mathcal{E}_p \\ &= \mathcal{F}_{p'} \oplus (\mathcal{F}_p \oplus \mathcal{E}_1) \\ &= (\mathcal{F}_{p'} \oplus \mathcal{F}_p) \oplus \mathcal{E}_1 \\ &= \mathcal{F} \oplus \mathcal{E}_1 \\ &\cong \mathcal{F} \oplus \mathbb{Z}_p. \end{aligned}$$

◀

Remark: We can regard $\mathcal{E}_p$ *as a* $\mathbb{Z}_p$*-module*; for since $p^n P \to O$ we can define xP unambiguously for $x \in \mathbb{Z}_p$:

$$n_i \to x \Longrightarrow n_i P \to xP.$$

Moreover, $\mathcal{E}_p$ is a *finitely-generated* $\mathbb{Z}_p$-module; that follows readily from the fact that $\mathcal{E}_{(p)} \cong \mathbb{Z}_p$ is of finite index in $\mathcal{E}_p$.

The Structure Theorem for finitely-generated abelian groups, ie Z-modules, extends easily to $\mathbb{Z}_p$-modules; such a module is the direct sum of copies of $\mathbb{Z}_p$ and cyclic groups $\mathbb{Z}/(p^e)$. (This can be proved in much the same way as the corresponding result for abelian groups.)

Effectively, therefore, all we proved above was that the factor $\mathbb{Z}_p$ occurred just once, which simply reflects the fact that we are dealing with a 1-dimensional curve.

Chapter 6

Points of Finite Order

6.1 The Torsion Subgroup

The elements of finite order in an abelian group A form a subgroup $F \subset A$, since

$$a, b \in F \Longrightarrow ma = 0, nb = 0 \Longrightarrow mn(a+b) = 0 \Longrightarrow a + b \in F.$$

This subgroup F is commonly called the *torsion* subgroup of A. (See Appendix A for further details.)

It turns out to be much easier to determine the torsion subgroup $F \subset \mathcal{E}(\mathbb{Q})$ of an elliptic curve than it is to determine the rank of the curve — that is, the number of copies of $\mathbb{Z}$ in

$$\mathcal{E}(\mathbb{Q}) = F \oplus \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}.$$

In effect the discussion below provides a simple algorithm for determining F, while there is no known algorithm for determining the rank.

Proposition 6.1 *The torsion subgroup of an elliptic curve $\mathcal{E}(\mathbb{Q})$ is finite, ie $\mathcal{E}$ has only a finite number of points of finite order.*

Proof ▶ Suppose $\mathcal{E}$ has equation

$$y^2 + c_1 xy + c_3 y = x^3 + c_2 x^2 + c_4 x + c_6,$$

where $c_i \in \mathbb{Q}$. Choose an odd prime p not appearing in the denominators of the c_i, and consider the p-adic curve $\mathcal{E}(\mathbb{Q}_p)$. Any point $P \in \mathcal{E}(\mathbb{Q})$ of finite order will still have finite order in $\mathcal{E}(\mathbb{Q}_p)$.

We know that $\mathcal{E}(\mathbb{Q}_p)$ has an open subgroup

$$\mathcal{E}_{(p)}(\mathbb{Q}_p) \cong \mathbb{Z}_p.$$

The only point of finite order in this subgroup is 0 (since $\mathbb{Z}_p$ has no other elements of finite order).

It follows that any coset

$$P + \mathcal{E}_{(p)}(\mathbb{Q}_p)$$

contains at most one element of finite order. For if there were two, say P, Q, then $P - Q$ would be a point of finite order in the subgroup.

But $\mathcal{E}(\mathbb{Q}_p)$ is compact, since it is a closed subspace of the compact space $\mathbb{P}^2(\mathbb{Q}_p)$. Hence it can be covered by a finite number of cosets

$$P_1 + \mathcal{E}_{(p)}(\mathbb{Q}_p),\ \ldots, P_r + \mathcal{E}_{(p)}(\mathbb{Q}_p).$$

Since each coset contains at most 1 point of finite order, the number of such points is finite. ◄

Remark: We shall prove in Chapter 8 the much deeper result that the group $\mathcal{E}(\mathbb{Q})$ of an elliptic curve over $\mathbb{Q}$ is *finitely-generated* (Mordell's Theorem), from which the finiteness of F follows, as shown in Appendix A.

6.2 Lessons from the Real Case

Proposition 6.2 *Suppose F is the torsion subgroup of the elliptic curve $\mathcal{E}(\mathbb{Q})$. Then*

$$F \cong \mathbb{Z}/(n) \text{ or } F \cong \mathbb{Z}(2n) \oplus \mathbb{Z}/(2).$$

Proof ► We know that

$$\mathcal{E}(\mathbb{R}) \cong \mathbb{T} \text{ or } \mathbb{T} \oplus \mathbb{Z}/(2).$$

Since

$$\mathcal{E}(\mathbb{Q}) \subset \mathcal{E}(\mathbb{R}),$$

it follows that

$$F \subset \mathbb{T} \text{ or } \mathbb{T} \oplus \mathbb{Z}/(2).$$

Lemma *Every finite subgroup of $\mathbb{T}$ is cyclic; and there is just one such subgroup of each order n.*

Proof of Lemma ▷ The torsion subgroup of

$$\mathbb{T} = \mathbb{R}/\mathbb{Z}$$

is

$$F = \mathbb{Q}/\mathbb{Z}.$$

For if $\bar{t} \in \mathbb{T}$ is of order n then $nt \in \mathbb{Z}$, say $nt = m$, ie $t = m/n \in \mathbb{Q}$. Conversely, if $t \in \mathbb{Q}$, say $t = m/n$, then $n\bar{t} = 0$, and so $\bar{t} \in F$.

Suppose

$$A \subset \mathbb{Q}/\mathbb{Z}$$

is a finite subgroup $\neq 0$. Since each $\bar{t} \in \mathbb{T}$ has a unique representative $t \in [-1/2, 1/2)$, A has a smallest representative $t = m/n > 0$, where we may assume that $m, n > 0$, $\gcd(m, n) = 1$.

In fact $n = 1$; for we can find $u, v, \in \mathbb{Z}$ such that

$$um + vn = 1,$$

and then

$$\frac{1}{n} = u\frac{m}{n} + v,$$

ie

$$\frac{1}{n} \equiv u\frac{m}{n} \bmod \mathbb{Z}$$

Thus

$$\frac{1}{n} \in A.$$

Since $1/n \leq m/n$, this must be our minimal representative: $n = 1$.

Now every element $\bar{t} \in A$ must be of the form m/n; for otherwise we could find a representative

$$t - m/n \in (0, 1/n),$$

contradicting our choice of $1/n$ as minimal representative of A.

We conclude that

$$A = \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\right\} \cong \mathbb{Z}/(n).$$

Moreover, our argument shows that this is the only subgroup of A of order n. ◁

Since this is the only subgroup of $\mathbb{T}$ of order n we can write

$$\mathbb{Z}/(n) \subset \mathbb{T}$$

without ambiguity, identifying

$$r \bmod n \longleftrightarrow r/n \bmod \mathbb{Z}$$

This establishes the result if $F \subset \mathbb{T}$. It remains to consider the case

$$A \subset \mathbb{T} \oplus \mathbb{Z}/(2).$$

By the Lemma, $A \cap \mathbb{T}$ is cyclic, say

$$A \cap \mathbb{T} = \mathbb{Z}/(n).$$

Thus

$$\mathbb{Z}/(n) \subset A \subset \mathbb{Z}/(n) \oplus \mathbb{Z}/(2).$$

Since $\mathbb{Z}/(n)$ is of index 2 in $\mathbb{Z}/(n) \oplus \mathbb{Z}/(n)$ it follows that

$$A = \mathbb{Z}/(n) \text{ or } A = \mathbb{Z}/(n) \oplus \mathbb{Z}/(2).$$

If n is odd then

$$\mathbb{Z}/(n) \oplus \mathbb{Z}/(2) \cong \mathbb{Z}/(2n)$$

by the Chinese Remainder Theorem. Thus either A is cyclic or else

$$A \cong \mathbb{Z}/(n) \oplus \mathbb{Z}/(2)$$

with n even. ◂

Mazur has shown that in fact the torsion group of an elliptic curve can only be one of a small number of groups, namely

$$\mathbb{Z}/(n)\ (n = 1 - 10, 12) \text{ and } \mathbb{Z}/(2n) \oplus \mathbb{Z}/(2)\ (n = 1 - 5).$$

6.2.1 Elements of order 2

We can distinguish between the two cases in Proposition 6.2 by considering the number of points of order 2. For $Z/(n)$ has no points of order 2 if n is odd, and just one point if n is even, say $n = 2m$, namely $m \bmod n$; while $\mathbb{Z}/(2n) \oplus \mathbb{Z}/(2)$ has three points of order 2, namely $(n \bmod 2n, 0 \bmod 2)$, $(n \bmod 2n, 1 \bmod 2)$, $(0 \bmod 2n, 1 \bmod 2)$.

Proposition 6.3 *The point $P = (x, y)$ on the elliptic curve*

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx + c \qquad (a, b, c \in \mathbb{Q})$$

has order 2 if and only if $y = 0$. There are either 0, 1 or 3 points of order 2.

Proof ▸ If $P = (x, y)$ then $-P = (x, -y)$. Thus $2P = 0$, ie $-P = P$, if and only if $y = 0$.

Thus there are as many elements of order 2 as there are roots of $f(x) = x^3 + ax^2 + bx + c$ in $\mathbb{Q}$. But if 2 roots $\alpha, \beta \in \mathbb{Q}$ then the third root $\gamma \in \mathbb{Q}$, since

$$\alpha + \beta + \gamma = -a.$$

◂

In determining whether

$$p(x) = x^3 + ax^2 + bx + c$$

has 0, 1 or 3 rational roots, one idea is very important: *if $a, b, c \in \mathbb{Z}$ then every rational root r of $p(x)$ is in fact integral, and $r \mid n$.* (For on substituting $r = m/n$ and multiplying by n^3, each term is divisible by n except the first.) This usually reduces the search for rational roots to a number of simple cases.

We may also note that if $a, b, c \in \mathbb{Z}$ then a necessary — but not sufficient — condition for $p(x)$ to have 3 rational roots is that the discriminant D should be a perfect square: $D = d^2$. For

$$D = [(\alpha - \beta)(\beta - \gamma)(\gamma - \alpha)]^2 .$$

6.2.2 Elements of order 3

In any abelian group, the elements of order p (where p is a prime), together with 0, form a subgroup; for

$$pa = 0, pb = 0 \Longrightarrow p(a + b) = 0.$$

We can consider this subgroup as a vector space over the finite field $\mathbf{GF}(p)$.

Proposition 6.4 *If p is an odd prime then there are either no points of order p on the elliptic curve $\mathcal{E}(\mathbb{Q})$, or else there are exactly $p - 1$ such elements, forming with 0 the group $\mathbb{Z}/(p)$.*

Proof ▸ An element of $\mathbb{T} \oplus \mathbb{Z}/(2)$ of odd order p is necessarily in $\mathbb{T}$. Thus the result follows from Proposition 6.2 and the Lemma in the proof of that Proposition. ◂

The elements of order 3 have a particularly simple geometric description.

Proposition 6.5 *A point $P \neq 0$ on the elliptic curve $\mathcal{E}(\mathbb{Q})$ has order 3 if and only if it is a point of inflexion. There are either 0 or 2 such points.*

Proof ► Suppose P has order 3, ie

$$P + P + P = 0.$$

From the definition of addition, this means that the tangent at P meets $\mathcal{E}$ in 3 coincident points P, P, P. In other words, P is a point of inflexion.

It follows from the previous Proposition that there are either 0 or 2 such flexes. ◄

Remark: The point 0 is of course a flex (by choice); so there are either 1 or 3 flexes on the elliptic curve $\mathcal{E}(\mathbb{Q})$ given by a general Weierstrass equation.

6.3 Points of Finite Order are Integral

Theorem 6.1 *Suppose $P = (x, y)$ is a point of finite order on the elliptic curve*

$$\mathcal{E}(\mathbb{Q}) : y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6,$$

where $c_1, c_2, c_3, c_4, c_6 \in \mathbb{Z}$. Then either $2P = 0$ or $x, y \in \mathbb{Z}$.

Proof ►

Lemma 1 *For each prime p, if $(x, y) \in \mathcal{E}(\mathbb{Q}_p)$ then*

$$\|x\|_p \leq 1 \Longleftrightarrow \|y\|_p \leq 1.$$

Proof of Lemma ▷ If $\|x\|_p \leq 1$ but $\|y\|_p > 1$ then y^2 will dominate the equation. On the other hand, if $\|x\|_p > 1$ but $\|y\|_p \leq 1$ then x^3 will dominate the equation. ◁

On combining these results for all primes,

$$x \in \mathbb{Z} \Longleftrightarrow y \in \mathbb{Z}.$$

(This last result is easily proved directly; for if $x \in Z$ then the equation for $\mathcal{E}$ can be regarded as a monic quadratic equation for y with integral coefficients; and any rational solution for y is therefore integral; and similarly if $y \in Z$ then the equation for $\mathcal{E}$ can be regarded as a monic cubic equation for x with integral coefficients; and any rational solution for x is therefore integral.)

Lemma 2 *If $P = (x, y) \in \mathcal{E}(\mathbb{Q}_p)$ then either $x, y \in \mathbb{Z}_p$ or else $P \in \mathcal{E}_{(p)}$.*

Proof of Lemma ▷ The equation of the curve in (X, Z)-coordinates is

$$Z + c_1XZ + c_3Z^2 = X^3 + c_2X^2Z + c_4XZ^2 + c_6Z^3.$$

Suppose $P \notin \mathcal{E}_{(p)}$, ie

$$\|X\|_p \geq 1 \text{ or } \|Z\|_p \geq 1.$$

In fact

$$\|X\|_p \geq 1 \Longrightarrow \|Z\|_p \geq 1;$$

for if $\|X\|_p \geq 1$ but $\|Z\|_p < 1$ then X^3 would dominate the equation. Thus

$$\|Z\|_p \geq 1$$

in either case.

Since $y = 1/Z$

$$\|Z\|_p \geq 1 \Longrightarrow \|y\|_p \leq 1.$$

Hence

$$x, y \in \mathbb{Z}_p$$

by Lemma 1. ◁

Lemma 3 *1. If p is odd then $\mathcal{E}_{(p)}$ is torsion-free (ie has no elements of finite order except 0).*

2. $\mathcal{E}_{(2^2)}$ is torsion-free.

Proof of Lemma ▷ This follows at once from the fact that

$$\mathcal{E}_{(p)} \cong \mathbb{Z}_p \ (p \text{ odd}), \qquad \mathcal{E}_{(2^2)} \cong \mathbb{Z}_2,$$

as we saw in Chapter 5. ◁

Lemma 4 *If $P \in \mathcal{E}_{(2)}$ then $2P \in \mathcal{E}_{(2^2)}$.*

Proof of Lemma ▷ Suppose $P = (X, Z)$. Recall that although $\mathcal{E}_{(2)}$ was defined as

$$\mathcal{E}_{(2)} = \left\{(X, Z) \in \mathcal{E} : \|X\|_2, \|Z\|_2 < 2^{-1}\right\},$$

in fact it follows from the equation

$$Z(1 + c_1X + c_2Z) = X^3 + c_2X^2Z + c_4XZ^2 + C_6Z^3$$

that
$$(X,Z) \in \mathcal{E}_{(2)} \Longrightarrow \|Z\|_2 \le 2^{-3}.$$
(More generally, although $\mathcal{E}_{(p^e)}$ is defined as
$$\mathcal{E}_{(p^e)} = \left\{(X,Z) \in \mathcal{E} : \|X\|_p < p^{-e}, \|Z\| < 1\right\},$$
in fact
$$(X,Z) \in \mathcal{E}_{(p^e)} \Longrightarrow \|Z\|_p \le p^{-3e}$$
by induction on e.)

The tangent at P is
$$Z = MX + D$$
where
$$\begin{aligned} M &= \frac{\partial F/\partial X}{\partial F/\partial Z} \\ &= \frac{c_1 Z - (3X^2 + 2c_2 XZ + 3c_4 Z^2)}{1 + c_1 X + 2c_3 Z - (c_2 X^2 + 2c_4 XZ + 3c_6 Z^2)}. \end{aligned}$$
The term $3X^2$ dominates the numerator, while the term 1 dominates the numerator. It follows that
$$\|M\|_2 \le 2^{-2}.$$
Hence
$$\|D\|_2 = \|Z - MX\|_2 \le 2^{-3}.$$
The tangent meets $\mathcal{E}$ where
$$\begin{aligned} &(MX+D)(1 + c_1 X + c_3(MX+D)) \\ &\quad = X^3 + c_2 X^2(MX+D) + c_4 X(MX+D)^2 + c_6(MX+D)^3. \end{aligned}$$
Thus if the tangent meets $\mathcal{E}$ again at (X_1, Z_1) then
$$\begin{aligned} 2X + X_1 &= -\frac{\text{coeff of } X^2}{\text{coeff of } X^3} \\ &= \frac{c_1 M + c_3 M^2 - (c_2 + 2c_4 M + 3c_6 M^2)D}{1 + c_2 M + c_4 M^2 + c_6 M^3}. \end{aligned}$$
Hence
$$\|X_1\|_2 \le 2^{-2}.$$
Since
$$\|Z_1\| = \|MX_1 + D\| \le 2^{-4},$$

it follows that
$$(X_1, Z_1) \in \mathcal{E}_{(2^2)}.$$
We conclude that
$$2P = -(X_1, Z_1) \in \mathcal{E}_{(2^2)},$$
since $\mathcal{E}_{(2^2(}$ is a subgroup of $\mathcal{E}$. ◁

Now suppose $P = (x,y) \in \mathcal{E}(\mathbb{Q})$ is of finite order.

For each odd prime p,
$$P \notin \mathcal{E}_{(p)}$$
by Lemma 3. Thus
$$x, y \in \mathbb{Z}_p$$
by Lemma 2.

Since $2P$ is of finite order,
$$P \in \mathcal{E}_{(2)} \Longrightarrow 2P \in \mathcal{E}_{(2^2)} \Longrightarrow 2P = 0,$$
by Lemmas 4 and 3. Thus if $2P \ne 0$ then
$$x, y \in \mathbb{Z}_2,$$
by Lemma refIntegrality.

Putting these results together, we conclude that either $2P = 0$ or else
$$x, y \in \mathbb{Z}_p \text{ for all } p \Longrightarrow x, y \in \mathbb{Z}.$$
◀

Corollary *If $P = (x,y)$ is a point of finite order on the elliptic curve*
$$y^2 = x^3 + ax^2 + bx + c$$
then $x, y \in \mathbb{Z}$.

Proof ▶ After the Proposition we need only consider the case
$$2P = 0 \Longrightarrow y = 0 \Longrightarrow x^3 + ax^2 + bx + c = 0.$$
Since a rational root of a monic polynomial with integral coefficients is necessarily integral, it follows that $x \in \mathbb{Z}$. ◀

Recall that if $P = (x,y)$ is a point of
$$\mathcal{E}(\mathbb{Q}) : y^2 + c_1 xy + c_3 y = x^3 + c_2 x^2 + c_4 x + c_6$$

then

$$-P = (x, -y - c_1x - c_3).$$

For by definition, $-P$ is the point where the line OP meets the curve again. But the lines through O are just the lines

$$x = c$$

parallel to the y-axis (together with the line $Z = 0$ at infinity). This is clear if we take the line in homogeneous form

$$lX + mY + nZ = 0.$$

This passes through $O = [0, 1, 0]$ if $m = 0$, giving

$$x = X/Z = -n/l.$$

Thus $-P$ is the point with the same x-coordinate as P, say

$$-P = (x, y_1).$$

But y, y_1 are the roots of the quadratic

$$y^2 + y(c_1x + c_3) - (x^3 + c_2x^2 + c_4x + c_6).$$

Hence

$$y + y_1 = -(c_1x + c_3),$$

ie

$$y_1 = -y - c_1x - c_3.$$

It follows that

$$\begin{aligned} 2P = 0 &\iff -P = P \\ &\iff y = -y - c_1x - c_3 \\ &\iff 2y + c_1x + c_3 = 0. \end{aligned}$$

Example: Consider the curve

$$\mathcal{E}(\mathbb{Q}) : y^2 + xy = x^3 + 4x^2 + x.$$

If $P = (x, y)$ is of order 2 then

$$2y + x = 0.$$

This meets the curve where

$$x^2/4 - x^2/2 = x^3 + 4x^2 + x,$$

ie

$$4x^3 + 17x^2 + 4x = 0.$$

This has roots $0, -1/4, -4$. Thus the curve has three points of order 2, namely $(0, 0), (-1/4, 1/8), (4, 2)$.

6.4 Points of Finite Order are Small

Theorem 6.2 (Nagell-Lutz) *Suppose the elliptic curve $\mathcal{E}(\mathbb{Q})$ has equation*

$$y^2 = f(x),$$

where

$$f(x) \equiv x^3 + ax^2 + bx + c \qquad (a, b, c \in \mathbb{Z});$$

and suppose $P = [x, y, 1] \in \mathcal{E}$ is a point of finite order. Then either $y = 0$, or

$$y^2 \mid \Delta(f),$$

where

$$\Delta = 8a^3c - a^2b^2 - 18abc + 4b^3 + 27c^2$$

is the discriminant of $f(x)$.

Proof ► We start by proving the weaker result

$$y \mid \Delta(f),$$

since this brings out the basic idea in a more direct way.

Suppose $P = (x, y)$ is a point of finite order. Then so is $2P = (x_1, y_1)$. Thus by Proposition ,

$$x, y, x_1, y_1 \in \mathbb{Z}.$$

Recall that

$$2x + x_1 = -a + m^2,$$

where

$$m = \frac{f'(y)}{2y}.$$

Since $a \in \mathbb{Z}$, it follows that

$$m^2 \in \mathbb{Z} \Longrightarrow m \in Z \Longrightarrow 2y \mid f'(x).$$

On the other hand

$$y \mid f(x)$$

since $y^2 = f(x)$. Thus

$$y \mid f(x), f'(x).$$

Recall that the resultant $R(f,g)$ of two polynomials

$$f(x) = a_0x^m + a_1x^{m-1} + \cdots + a_m,\ g(x) = b_0x^n + b_1x^{n-1} + \cdots + b_n$$

is the determinant of the $(m+n) \times (m+n)$ matrix

$$\mathbf{R}(f,g) = \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_m & 0 & \dots & 0 \\ 0 & a_0 & a_1 & \dots & a_{m-1} & a_m & \dots & 0 \\ & & & \dots & & & & \\ 0 & 0 & 0 & \dots & & \dots & a_{m-1} & a_m \\ b_0 & b_1 & b_2 & \dots & b_n & 0 & \dots & 0 \\ 0 & b_0 & b_1 & \dots & b_{n-1} & b_n & \dots & 0 \\ & & & \dots & & & & \\ 0 & 0 & 0 & \dots & & \dots & b_{n-1} & b_n \end{pmatrix}$$

We saw earlier that $R(f,g) = 0$ is a necessary and sufficient condition for $f(x), g(x)$ to have a root in common. Our present use of the resultant, though related, is more subtle.

Lemma 1 *Suppose $f(x), g(x) \in \mathbb{Z}[x]$. Then there exist polynomials $u(x), v(x) \in \mathbb{Z}[x]$ such that*

$$u(x)f(x) + v(x)g(x) = R(f,g).$$

Proof of Lemma ▷ Let us associate to the polynomials

$$u(x) = c_0x^{n-1} + c_1x^{n-2} + \cdots + c_{n-1},\ v(x) = d_0x^{m-1} + d_1x^{m-2} + \cdots + d_{m-1}$$

(of degrees $< n$ and $< m$) the $(m+n)$-vector

$$\begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \\ d_0 \\ d_1 \\ \vdots \\ d_{m-1} \end{pmatrix}.$$

It is readily verified that if

$$u(x)f(x) + v(x)g(x) = e_0x^{m_n-1} + \cdots + e_{m+n-1},$$

then the e_k are given by the vector equation

$$\mathbf{R}(f,g) \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \\ d_0 \\ d_1 \\ \vdots \\ d_{m-1} \end{pmatrix} = \begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_{m+n-1} \end{pmatrix}.$$

We are looking for integers c_i, d_j such that

$$\begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_{m+n-1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ R(f,g) \end{pmatrix}$$

The existence of such integers follows at once from the following Sublemma. (For simplicity we prove the result with $\det A$ as first coordinate rather than last; but it is easy to see that this does not matter.)

Sublemma *Suppose A is an $n \times n$-matrix with integer entries. Then we can find a vector v with integer entries such that*

$$A \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \det A \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Proof of Lemma ▷ On expanding $\det A$ by its first column,

$$\det A = a_{11}A_{11} + a_{21}A_{21} + \cdots + a_{n1}A_{n1},$$

where the A_{i1}'s are the corresponding co-factors. On the other hand, if $i \neq n$ then

$$a_{1i}A_{11} + a_{2i}A_{21} + \cdots + a_{ni}A_{n1} = 0,$$

since this is the determinant of a matrix with two identical columns.

Thus the vector

$$v = \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{n1} \end{pmatrix}$$

has the required property. ◁

◁

We apply this Lemma to the polynomials $f(x), f'(x)$, recalling that

$$R(f, f') = -D(f).$$

Thus we can find polynomials $u(x), v(x) \in \mathbb{Z}[x]$ such that

$$u(x)f(x) + v(x)f'(x) = D.$$

Hence

$$y \mid f(x), f'(x) \Longrightarrow y \mid D.$$

Turning now to the full result, suppose again the $P = (x, y)$ is of finite order, and that $2P = (x_1, y_1)$. We know that $x, y, x_1, y_1 \in \mathbb{Z}$.

Lemma 2 *The x-coordinate of $2P$ is*

$$-\frac{g(x)}{4y^2},$$

where

$$g(x) = x^4 - 2bx^2 - 8cx - 4ac + b^2.$$

Proof of Lemma ▷ Recall that

$$x(2P) = 2x + a - m^2,$$

where

$$x = \frac{f'(x)}{2y}.$$

Thus

$$\begin{aligned} x(2P) &= \frac{4y^2(2x + a) - f'(x)^2}{4y^2} \\ &= \frac{4(x^3 + ax^2 + bx + c)(2x + a) - (3x^2 + 2ax + b)^2}{4y^2}, \end{aligned}$$

which yields the given result on simplification. ◁

It follows from the lemma that

$$y^2 \mid g(x);$$

Thus

$$y^2 \mid f(x), g(x)$$

since $y^2 = f(x)$.

Lemma 3 *There exist polynomials $u(x), v(x) \in \mathbb{Z}[x]$ of degrees $3, 2$ such that*

$$u(x)f(x) + v(x)g(x) = D.$$

Proof of Lemma ▷ For simplicity we are going to prove the result in the case $a = 0$. We leave it to the reader to establish the general result.

Let us see if we can find $u(x), v(x) \in \mathbb{Q}[x]$ of the form

$$u(x) = x^3 + Bx + C, \qquad v(x) = x^2 + D$$

such that

$$u(x)f(x) - v(x)g(x) = \text{const.}$$

The coefficients of x^6 and x^5 on the left both vanish. Equating the coefficients of x^4, x^3, x^2, x yield

$$\begin{array}{lll} x^4: & b + B = -2b + D & \Longrightarrow D = B + 3b \\ x^3: & c + C = -8c & \Longrightarrow C = -9c \\ x^2: & Bb = b^2 - 2Db & => 2D + B = b \\ x: & Bc + Cb = -8Dc & \Longrightarrow B - 9b = -8D. \end{array}$$

Substituting for D in the third equation gives

$$B = -5b/3, \qquad D = 4b/3.$$

The final equation then reduces to

$$-5b/3 - 9b = -32b/3,$$

which is an identity.

Accordingly, we take

$$u(x) = 3x^3 - 5bx - 27c, \qquad v(x) = 3x^2 + 4b,$$

and then

$$u(x)f(x) - v(x)g(x) = -27c^2 - 4b^2 = D,$$

as required $\triangleleft$

The result now follows as before; since $x, y \in \mathbb{Z}$,

$$y^2 \mid f(x), g(x) \Longrightarrow y^2 \mid D.$$

◀

Remark: The resultant of $f(x), g(x)$ turns out to be

$$R(f, g) = -D^2,$$

so our earlier Lemma would be insufficient. It is not entirely clear (to me at least) *why* we can find $u(x), v(x)$ — of lower degrees than expected — such that

$$u(x)f(x) + v(x)g(x) = D.$$

6.5 Examples

In these examples we compute the torsion group F of various elliptic curves $\mathcal{E}(\mathbb{Q})$.

1. We look first at the curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + 1.$$

Recall that the discriminant of the polynomial

$$p(x) = x^3 + bx + c$$

is

$$D = -\left(4b^3 + 27c^2\right).$$

Thus in the present case

$$D = -27.$$

It follows from Nagell-Lutz (Theorem 6.2) that

$$y = 0, \pm 1, \pm 3.$$

There is just one point of order 2, ie with $y = 0$, namely $(-1, 0)$.

If $y = \pm 1$ then $x = 0$, giving the two points $(0, \pm 1)$.

If $y = \pm 3$ then $x^3 = 8$, giving the two points $(2, \pm 3)$.

It remains to determine which of these points $(0, \pm 1), (2, \pm 3)$ is of finite order – remembering that the Nagell-Lutz condition $y^2 \mid D$ is *necessary* (if $y \neq 0$) but by no means *sufficient.*

The tangent at $P = (0, 1)$ has slope

$$m = \frac{p'(x)}{2y} = \frac{3x^2}{2y} = 0.$$

Thus the tangent at P is

$$y = 1.$$

This meets $\mathcal{E}$ where

$$x^3 = 0,$$

ie thrice at P. In other words P is a flex, and so of order 3.

Turning to the point $(2, 3)$ we have

$$m = \frac{3x^2}{2y} = 2.$$

and so the tangent at this point is

$$y = 2x - 1,$$

which meets $\mathcal{E}$ again at $(0, -1)$. Thus

$$2(2, 3) = -(0, -1) = (0, 1).$$

We conclude that $(2, 3)$ (and $(2, -3) = -(2, 3)$) are of order 6, and

$$F = \mathbb{Z}/(6).$$

2. Consider the curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 - 1.$$

Again, $D = -27$, and there is one point $(1, 0)$ of order 2.

But now

$$y = \pm 1 \Longrightarrow x^3 = 2,$$
$$y = \pm 3 \Longrightarrow x^3 = 10,$$

neither of which has solutions in $\mathbb{Z}$. We conclude that

$$F = \mathbb{Z}/(2).$$

3. Suppose F is the torsion subgroup of

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + x$$

We have

$$D = -4,$$

and so

$$y = 0, \pm 1, \pm 2.$$

There is just one point of order 2, ie with $y = 0$, namely $(0, 0)$.

If $y = \pm 1$ then

$$x^3 + x - 1 = 0.$$

Note that *a rational root $\alpha \in \mathbb{Q}$ of a monic polynomial*

$$x^n + a_1 x^{n-1} + \cdots + a_n$$

with integral coefficients $a_i \in \mathbb{Z}$ is necessarily integral: $\alpha \in \mathbb{Z}$. And evidently $\alpha \mid a_n$. Thus in the present case the only possible rational roots of the equation are $x = \pm 1$; and neither of these is in fact a root.

If $y = \pm 2$ then

$$x^3 + x - 4 = 0.$$

The only possible solutions to this are $x = \pm 1, \pm 2, \pm 4$; and it is readily verified that none of these is in fact a solution.

We conclude that

$$F = \mathbb{Z}/(2).$$

4. Consider the curve

$$y^2 = x^3 - x^2.$$

This curve is singular, since $p(x) = x^3 - x^2$ has a double root, (and so $D = 0$). Thus it is not an elliptic curve, and so is outside our present study, although we shall say a little about singular cubic curves in the next Chapter.

5. Consider the curve

$$\mathcal{E}(\mathbb{Q}) : y^2 - y = x^3 - x.$$

This has 6 obvious integral points, namely $(0, 0), (0, 1), (1, 0), (1, 1), (-1, 0), (-1, 1)$.

We can bring the curve to standard form by setting $y_1 = y - 1/2$, ie $y = y_1 + 1/2$, to complete the square on the left. The equation becomes

$$y_1^2 = x^3 - x + 1/4.$$

Now we can make the coefficients integral by the transformation

$$y_2 = 2^3 y_1, \ x_2 = 2^2 x,$$

giving

$$y_2^2 = x_2^3 - 2^4 x_2 + 2^6/4,$$

since the coefficient of x has weight 4, while the constant coefficient has weight 6. (In practice it is probably easier to apply this transformation first, and then complete the square; that way our coefficients always remain integral.) Our new equation is

$$y_2^2 = x_2^3 - 16 x_2 + 16,$$

with discriminant

$$\begin{aligned} D &= -(4 \cdot 2^{12} + 27 \cdot 2^8) \\ &= -2^8(64 + 27) \\ &= -2^8 91. \end{aligned}$$

By Nagell-Lutz, if $(x_2, y_2) \in F$ then $x_2, y_2 \in \mathbb{Z}$ and

$$y_2 = 0, \pm 1, \pm 2, \pm 4, \pm 8, \pm 16.$$

Note however that if P is not of order 2, ie $y_2 \neq 0$, then

$$y = \frac{y_2 - 4}{8} \in \mathbb{Z}$$

by Theorem 6.2. Only the cases $y_2 = \pm 4$ satisfy this condition. Thus we only have to consider

$$y_2 = 0, \pm 4.$$

If $y_2 = 0$ then

$$x_2^3 - 16 x_2 + 16 = 0.$$

But

$$\begin{aligned} 16 \mid x_2^3 &\Longrightarrow 4 \mid x_2 \\ &\Longrightarrow 32 \mid x_2^3, 16 x_2 \\ &\Longrightarrow 32 \mid 16, \end{aligned}$$

which is absurd. Thus there are no points of order 2 on $\mathcal{E}$.

Finally, if $y_2 = \pm 4$ then

$$16 = x_2^3 - 16x_2 + 16 \Longrightarrow x_2^3 - 16x_2 = 0 \Longrightarrow x_2 = 0, \pm 4.$$

This gives the 6 'obvious' points we mentioned at the beginning.

It remains to determine which of these points are of finite order.

Reverting to the original equation, suppose $P = (0,0)$. We have

$$(2y-1)\frac{dy}{dx} = 3x^2 - 1,$$

ie

$$\frac{dy}{dx} = \frac{3x^2-1}{2y-1}.$$

Thus the tangent at P has slope $m = 1$, and so is

$$y = x.$$

This meets the curve again at $(1,1)$. Hence

$$2(0,0) = -(1,1) = (1,0).$$

The tangent at $(1,0)$ has slope $m = -2$, and so is

$$y = -2x + 2,$$

which meets $\mathcal{E}$ where

$$(-2x+2)^2 - x(-2x+2) = x^3 - x,$$

ie

$$x^3 - 6x^2 + 9x - 4 = 0.$$

We know this has two roots equal to 1. The third root must satisfy

$$2 + x = 6,$$

ie

$$x = 4.$$

At this point

$$y = -2x + 2 = -6.$$

We know that this point $(4,-6)$ is not of finite order, by Nagell-Lutz. It follows that $(1,0)$ is of infinite order. Hence so is $(0,0)$ since $2(0,0) = (1,0)$; and so too are $(1,1) = -(1,0)$ and $(0,1) = -(0,0)$

It remains to consider the points $(-1,0$ and $(-1,1) = -(-1,0)$. Note that if these are of finite order then they must be of order 3 (since there would be just 3 points in F), ie they would be flexes.

The tangent at $P = (-1,0)$ has slope $m = -2$, and so is

$$y = -2x - 2.$$

This meets $\mathcal{E}$ where

$$(-2x-1)^2 - x(-2x-1) = x^3 - x.$$

We know that this has two roots -1. Hence the third root is given by

$$-2 + x = 6,$$

ie

$$x = 8,$$

as before. At this point

$$y = -2x + 2 = -14.$$

So

$$2(-1,0) = -(8,-14).$$

Again, we know by Nagell-Lutz that this point is of infinite order, and so therefore is $(-1,0)$ and $(-1,1) = -(-1,0)$.

To verify that $P = (4,-6)$, for example, is not of finite order, we may note that the tangent at this point has slope

$$m = -\frac{47}{11}.$$

But the tangent

$$y = mx + d$$

at P meets the curve again where

$$(mx+d)^2 - x(mx+d) = x^3 - x,$$

ie at a point (x_1, y_1) with

$$2 \cdot 4 + x_1 = m^2 - m.$$

By Nagell-Lutz, $x_1 \in \mathbb{Z}$ (since we have seen that there are no points of order 2), and so $m^2 - m \in \mathbb{Z}$, which is manifestly not the case.

We conclude that the torsion-group of this curve is trivial:

$$F = \{0\}.$$

Chapter 7

Reduction modulo p

7.1 The reduction map

One serendipitous consequence of our adoption of projective (rather than affine) geometry is that this allows us to 'reduce' rational points modulo a prime p.

Proposition 7.1 *Suppose p is a prime. For each dimension n we can define a map*

$$\Pi_p : \mathbb{P}^n(\mathbb{Q}) \to \mathbb{P}^n(\mathbf{GF}p)$$

as follows: Any point $P \in \mathbb{P}^n(\mathbb{Q})$ can be expressed in the form

$$P = [X_0, X_1, \ldots, X_n]$$

where $X_0, X_1, \ldots, X_n \in \mathbb{Z}$ and not all X_i are divisible by p. We set

$$\Pi_p(P) = \bar{P} = [X_0 \bmod p, X_1 \bmod p, \ldots, X_n \bmod p].$$

Proof ► We can ensure that the coordinates X_i are all integral, by multiplying by the lcm of the denominators; and then we can ensure that not all the X_i are divisible by p by dividing by the highest power of p dividing all the X_i.

It remains to show that the resulting point $\bar{P} \in \mathbb{P}^n(\mathbf{GF}p)$ is uniquely determined by the point P. Suppose we have two such expressions for P:

$$P = [X_0, X_1, \ldots, X_n] = [X_0', X_1', \ldots, X_n'].$$

Then

$$[X_0', X_1', \ldots, X_n'] = \rho[X_0, X_1, \ldots, X_n]$$

for some $\rho \in \mathbb{Q}^\times$. Let

$$\rho = \frac{r}{s},$$

where $\gcd(r,s)=1$. Then

$$rX'_i = sX_i$$

for all i. Clearly $p \not| r$; for otherwise $p \mid X_i$ for all i. Similarly $p \not| s$. But then

$$\bar{r}\overline{X'_i} = \bar{s}\overline{X_i}$$

ie

$$[\overline{X'_0}, \overline{X'_1}, \dots, \overline{X'_n}] = \bar{\rho}[\overline{X_0}, \overline{X_1}, \dots, \overline{X_n}],$$

where $\bar{\rho} = \bar{r}/\bar{s}$.

Thus the two representations of P give the same point $\bar{P}$. ◄

Definition 7.1 *We call the map*

$$\mathbb{P}^n(\mathbb{Q}) \to \mathbb{P}^n(\mathbf{GF}p) : P \mapsto \tilde{P}$$

reduction modulo p.

It is not necessary to choose *integral* coordinates for reduction; it is sufficient that they be *p-integral*, that is, of the form $c = a/b$, where a, b are integers with $p \not| b$. Note that if b is p-integral then the 'remainder' $\tilde{c} = \tilde{a}/\tilde{b}$ modulo p is well-defined. The following result is readily verified.

Proposition 7.2 *Suppose*

$$P = [X_0, \dots, X_n],$$

where $X_0, \dots, X_n$ are p-integral but $\bar{X}_0, \dots, \bar{X}_n$ do not all vanish. Then

$$\bar{P} = [\bar{X}_0, \dots, \bar{X}_n].$$

Proposition 7.3 *Each line ℓ in $\mathbb{P}^2(k)$ defines a line $\bar{\ell}$ in $\mathbb{P}^2(\mathbf{GF}p)$; and*

$$P \in \ell \Longrightarrow \bar{P} \in \bar{\ell}.$$

More generally, each curve Γ in $\mathbb{P}^2(k)$ defines a line $\bar{\Gamma}$ in $\mathbb{P}^2(\mathbf{GF}p)$; and

$$P \in \Gamma \Longrightarrow \bar{P} \in \bar{\Gamma}.$$

Proof ► Suppose ℓ is the line

$$aX + bY + cZ = 0.$$

We can ensure that a, b, c are integral, by multiplying by the lcm of their denominators, and we can ensure that a, b, c are not all divisible by p, by dividing a, b, c by a suitable power of p; and then we set

$$\bar{\ell} : \bar{a}X + \bar{b}Y + \bar{c}Z = 0.$$

If now $P = [X, Y, Z]$ where X, Y, Z are all integers, but not all are divisible by p, then

$$aX + bY + cZ = 0 \Longrightarrow \bar{a}\bar{X} + \bar{b}\bar{Y} + \bar{c}\bar{Z} = 0.$$

Thus P lies on the line

$$\bar{\ell} : \bar{a}X + \bar{b}Y + \bar{c}Z = 0.$$

Now suppose Γ is a curve in $\mathbb{P}^2(\mathbb{Q})$, given by the homogeneous polynomial equation

$$F(X, Y, Z) = 0.$$

We can ensure that all the coefficients of F are integral, but not all divisible by p; and then we can define he polynomial

$$\bar{F}[X, Y, Z] \in \mathbf{GF}p[X, Y, Z],$$

by taking each coefficient of F mod p.

Suppose $P = [X, Y, Z]$ where $X, Y, Z \in \mathbb{Z}$ but not all are divisible by p. Then

$$P \in \Gamma \Longleftrightarrow F(X, Y, Z) = 0 \Longrightarrow \bar{F}(\bar{X}, \bar{Y}, \bar{Z}) = 0 \Longleftrightarrow \bar{P} \in \bar{\Gamma}.$$

◄

7.1.1 Reduction of Elliptic Curves

Definition 7.2 *We say that the elliptic curve $\mathcal{E}(\mathbb{Q})$ has* good reduction *mod p if $\tilde{\mathcal{E}}$ is elliptic, ie non-singular.*

We often say that $\mathcal{E}$ has good reduction *at* p.

Consider the elliptic curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx + c,$$

where $a, b, c \in \mathbb{Z}$ (or, more generally, a, b, c are p-integral).

Reduction modulo p gives the curve

$$\tilde{\mathcal{E}} : y^2 = x^3 + \tilde{a}x^2 + \tilde{b}x + \tilde{c}$$

over the finite field $\mathbf{GF}p$.

Proposition 7.4 *The reduction $\tilde{\mathcal{E}}$ of $\mathcal{E}$ modulo p is good if and only if $p \neq 2$ and*

$$p \not| D,$$

where

$$D = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2$$

is the discriminant of the polynomial $p(x) = x^3 + ax^2 + bx + c$:

Proof ► If $p = 2$ then $\bar{\mathcal{E}}$ is necessarily singular.

Suppose $p \neq 2$. We know in this case that $\bar{\mathcal{E}}$ is elliptic (non-singular) if and only if $D(\bar{\mathcal{E}}) \neq 0$. The result follows since

$$D(\bar{\mathcal{E}}) = D(\mathcal{E}) \bmod p.$$

◄

Theorem 7.1 *Suppose the elliptic curve $\mathcal{E}(\mathbb{Q})$ has good reduction modulo p. Then the map*

$$\mathcal{E}(\mathbb{Q}) \to \mathcal{E}(\mathbf{GF}p) : P \mapsto \tilde{P}$$

is a homomorphism.

Proof ► The zero point on $\mathcal{E}$ certainly maps into the zero point on $\tilde{\mathcal{E}}$:

$$[0, 1, 0] \mapsto [0, \tilde{1}, 0].$$

Suppose the 3 points $P, Q, R \in \mathcal{E}(\mathbb{Q})$ satisfy

$$P + Q + R = 0.$$

In other words P, Q, R lie on a line

$$l : ax + by + cz = 0.$$

Let $\tilde{l}$ be the reduction of l modulo p. Evidently $\tilde{l}$ is a line in $\mathbb{P}^2(\mathbf{GF}p)$, which contains $\tilde{P}, \tilde{Q}, \tilde{R}$ by Proposition **??**.

We need to be a little careful at this point. If $\tilde{P}, \tilde{Q}, \tilde{R}$ are distinct then it follows that

$$\tilde{P} + \tilde{Q} + \tilde{R} = 0.$$

But can we be certain of this conclusion if 2 or all 3 of these points coincide? It's not difficult to see that we can.

Lemma 4 *Suppose the line l meets the curve $\Gamma \subset \mathbb{P}^2(\mathbb{Q})$ of degree n in the n rational points $P_1, \ldots, P_n$ (each repeated according to multiplicity). Then $\tilde{l}$ meets $\tilde{\Gamma}$ in $\tilde{P}_1, \ldots, \tilde{P}_n$ (each repeated according to multiplicity).*

Proof of Lemma ▷ Choose 2 points

$$Q = [x, y, z],\ R = [x', y', z']$$

on l such that $\tilde{Q} \neq \tilde{R}$. We may suppose that $x, y, z, x', y', z' \in \mathbb{Z}$ and that each triple x, y, z and x', y', z' is coprime. the line l takes the parametric form

$$P(s, t) = sQ + tR = [sx + tx', sy + ty', sz + tz'].$$

This will meet the curve Γ where

$$f(s, t) = F(sQ + tR) = 0.$$

This is a homogeneous equation of degree n in s, t, which by hypothesis has roots $(s_1, t_1), \ldots, (s_n, t_n)$ corresponding to the points $P_1, \ldots, P_n$. We may suppose that $s_1, \ldots, s_n, t_1, \ldots, t_n \in \mathbb{Z}$, and that each pair $(s_1, t_1), \ldots, (s_n, t_n)$ is coprime. Now

$$f(s, t) = c(st_1 - ts_1) \cdots (st_n - ts_n)$$

for some $c \in \mathbb{Q}$.

◁

Thus

$$P + Q + R = 0 \Longrightarrow \bar{P} + \bar{Q} + \bar{R} = 0.$$

Since it is readily verified that

$$\overline{-P} = -\bar{P},$$

it follows that the map is a homomorphism.

◄

Theorem 7.2 *Suppose the elliptic curve*

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx + c$$

has good reduction at the prime p. Let $T \subset \mathcal{E}(\mathbb{Q})$ be the torsion subgroup (formed by the points of finite order). Then the reduction map

$$\rho : \mathcal{E}(\mathbb{Q}) \to \mathcal{E}(\mathbf{GF}p),$$

sends T injectively onto a subgroup of $\mathcal{E}(\mathbf{GF}p)$.

Proof ► We know by the Nagell-Lutz Theorem 6.1 that the non-zero points

$$P = (X, Y) \in T$$

all have integral coordinates: $X, Y \in \mathbb{Z}$. It follows that

$$\tilde{P} = [\tilde{X}, \tilde{Y}, 1] = (\tilde{X}, \tilde{Y})$$

This can never be O. (It is always finite.) Thus

$$\ker \rho = \{0\},$$

and so ρ is injective. ◄

7.2 An example

By Theorem 7.2, the torsion subgroup T of $\mathcal{E}(\mathbb{Q})$ has an isomorphic image in $\mathcal{E}(\mathbf{GF}p)$ for every good prime p. We can often exploit this to determine T.

In general, the Nagell-Lutz Theorem provides a surer method of determining T. But there may be cases where the method below is quicker.

As an illustration, let us look at the curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + x + 1.$$

Since

$$D = -31.$$

$\mathcal{E}$ has good reduction at all odd primes p except 31.

Consider first reduction at $p = 3$ If $(x, y) \in \mathcal{E}(\mathbf{GF}3)$ then $x^3 + x + 1$ must be a quadratic residue modulo 3, ie

$$x^3 + x + 1 = 0 \text{ or } 1 \bmod 3.$$

This does not hold if $x = 2 = -1$; but it does hold in the other 2 cases

$$x = 0 \text{ and } x = 1.$$

When $x = 0$ we have $y = \pm 1$. When $x = 1$ we have $y = 1$.

It follows that

$$\mathcal{E}(\mathbf{GF}3) = \{(0, 1),\ (0, -1),\ (1, 0),\ [0, 1, 0]\}.$$

We know that the point (X, Y) has order 2 if and only if $Y = 0$. In this case there is just 1 such point, namely $(1, 0)$. Thus $\mathcal{E}(\mathbf{GF}3)$ is of order 4, and has 1 element of order 2. Consequently,

$$\mathcal{E}(\mathbf{GF}3) \cong \mathbb{Z}/(4).$$

New consider the curve defined by the same equation over $\mathbf{GF}5$. We have

$$x^3 + x + 1 = 0, 1 \text{ or } 4 \bmod 5.$$

This does not hold if $x = 1 \bmod 5$. The other cases yield the points:

$$(0, \pm 1),\ (2, \pm 1)m\ (3, \pm 1),\ (4 \pm 2).$$

Thus

$$|\mathcal{E}(\mathbf{GF}5)| = 9,$$

and so

$$\mathcal{E}(\mathbf{GF}5) = \mathbb{Z}/(3) \oplus \mathbb{Z}/(3) \text{ or } \mathbb{Z}/(9).$$

We leave it to the reader to determine which is the case.

This does not affect our present purpose, since in either case

$$T \subset \mathcal{E}(\mathbf{GF}3), \quad T \subset \mathcal{E}(\mathbf{GF}5) \Longrightarrow T = \{O\},$$

by Lagrange's Theorem.

7.3 Singular cubic curves

Recall that a curve Γ in $\mathbb{P}^2(k)$ is said to be *degenerate* if its equation factorizes:

$$\Gamma = \ell C,$$

where ℓ is a line and C a conic.

Proposition 7.5 *A non-degenerate cubic curve has at most one singularity.*

Proof ►

Lemma 5 *If P is a singular point on the non-degenerate curve Γ then every line through P meets Γ least twice at P.*

Proof of Lemma ▷ We may assume (after a suitable projective transformation) that the equation has no terms of the first order:

$$\Gamma : ax^2 + 2hxy + by^2 + O(x^3, y^3).$$

But any line $y = mx$ through P meets Γ where

$$(a + 2hm + bm^2)y^2 + O(x^3, y^3),$$

with a double root (at least) at $y - 0$, ie at $(0, 0)$. ◁

Now suppose P, Q are singularities. Then the line PQ meets Γ tiwce at P and twice at Q, by the Lemma. Thus the line meets Γ four times, which is impossible. Hence there is at most one singularity. ◀

Singularities on cubic curves divide into two kinds: *nodes* and *cusps*. These are distinguished as follows: Let us move the singularity to $(0, 0)$. Then

$$F(X, Y, Z) = aX^2 + 2hXY + bY^2 + O(X, Y)^3.$$

Definition 7.3 *A singularity on a cubic curve is said to be a node if the second order terms split into distinct factors:*

$$aX^2 + 2hXY + bY^2 = a(X + \alpha Y)(X + \beta Y),$$

where $\alpha \neq \beta$; it is said to be a cusp if $\alpha = \beta$, ie if the second order terms form a perfect square.

Definition 7.4 *Suppose $\mathcal{E}(\mathbb{Q})$ is an elliptic curve. Then we say that $\mathcal{E}$ is* stable *at p if the reduction* mod$p$ *is good. We say that $\mathcal{E}$ is* semi-stable *at p if the reduction is bad but the singularity in $\bar{\mathcal{E}}$ is a node. We say that $\mathcal{E}$ is* unstable *at p if the reduction is bad and the singularity in $\bar{\mathcal{E}}$ is a cusp.*

7.3.1 Nodes and cusps

Suppose we have a cubic curve

$$\Gamma(k) : y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6;$$

and suppose $\mathrm{char}(k) \neq 2, 3$. Then we can bring the curve to the form

$$y^2 = x^3 + ax^2 + bx + c.$$

Now suppose Γ has a singularity. We know that there is just one singular point, and that it is a point $(\alpha, 0)$ on the line $y = 0$, where α is a double or triple root of

$$p(x) = x^3 + ax^2 + bx + c.$$

This root $\alpha \in k$. For if α is a double root then $\gcd(p(x), p'(x)) = x - \alpha$, and we can compute this gcd by Euclid's algorithm within the ring $k[x]$; while if α is a triple root then $3\alpha = -b$.

Thus we may assume that $\alpha = 0$, after the transformation $x \mapsto x - \alpha$.

Our equation now takes the form

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

Note that the second-order terms are $y^2 - ax^2$. This has distinct factors unless $a = 0$. Thus by the definition above, *the singularity is a cusp if $a = 0$, and a node if $a \neq 0$.* (This accords with the look of the curve if $k = \mathbb{R}$.)

Let us consider the case where the singularity is a cusp first. Our equation is

$$y^2 = x^3.$$

We parametrize $\Gamma \setminus \{(0, 0)\}$ by the map

$$k \to \Gamma : t \mapsto \begin{cases} (t^{-2}, t^{-3}) & \text{if } t \neq 0, \\ [0, 1, 0] & \text{if } t = 0. \end{cases}$$

In other words,

$$P(t) = [t, 1, t^3]$$

for all $t \in k$.

Suppose the points P, Q, R with parameters p, q, r lie on the line

$$aX + bY + cZ = 0.$$

Then p, q, r are the roots of

$$at + b + ct^3 = 0.$$

Since the coefficient of t^2 is 0,

$$p + q + r = 0.$$

Thus from our definition of addition on $\Gamma \setminus \{(0, 0)\}$,

$$P + Q + R = 0 \Longleftrightarrow p + q + r = 0.$$

In addition, tt is readily verified that

$$-P(t) = P(-t).$$

It follows that the map

$$k \to \Gamma \setminus \{(0,0)\} : t \mapsto P(t)$$

is an isomorphism. Thus *the group on* $\Gamma\setminus\{(0,0)\}$ *is isomorphic to the additive group of* k.

Now let us consider the case where the singularity is a node. For simplicity let us take the curve

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

This has a node at $(0,0)$ with 'quasi-tangents' $y = \pm x$.

The line $y = mx$ meets the curve in just one point apart from $(0,0)$, unless $m = \pm 1$. We parametrize the curve by setting

$$t = \frac{y+x}{y-x}.$$

This gives

$$y = \frac{t+1}{t-1}x;$$

and so

$$\frac{(t+1)^2}{(t-1)^2}x^2 = x^3 + x^2,$$

ie

$$x = \frac{4t}{(t-1)^2}$$

and

$$y = \frac{4t(t+1)}{(t-1)^3}.$$

In homogeneous terms

$$(x,y) = [4t(t-1), 4t(t+1), (t-1)^3] = P(t).$$

It is readily verified that the map

$$k \to \Gamma : t \to P(t)$$

is bijective, with $t = 0$ corresponding to the singular point $(0,0)$. Thus we have a one-one correspondence between $t \in k^\times$ and $P \in \Gamma \setminus \{(0,0\}$.

Suppose the points P, Q, R with parameters p, q, r lie on the line

$$aX + bY + cZ = 0.$$

Then p, q, r are the roots of

$$4at(t-1) + 4bt(t+1) + c(1-t)^3 = 0.$$

Since the coefficients of t^3 and 1 are $\pm c$,

$$pqr = 1.$$

Thus

$$P + Q + R = 0 \Longleftrightarrow pqr = 0.$$

In addition, tt is readily verified that

$$-P(t) = P(1/t).$$

It follows that the map

$$k^\times \to \Gamma \setminus \{(0,0)\} : t \mapsto P(t)$$

is an isomorphism. Thus *the group on* $\Gamma \setminus \{(0,0)\}$ *is isomorphic to the multiplicative group* $k^\times$.

Recall that the the elliptic curve $\mathcal{E}(\mathbb{Q})$ is said to be *semi-stable* at p if $\bar{\mathcal{E}}$ has a node singularity, and *unstable* if $\bar{\mathcal{E}}$ has a cusp singularity, Because of the analysis above, the terms 'multiplicative' and 'additive' are sometimes used in these two cases. (Note though that we have not proved that the group is always $k^\times$ or k in these two cases; the story is a little bit more complicated than that.)

7.4 Hasse's Theorem

Consider the elliptic curve

$$\mathcal{E}(\mathbf{GF}p) : y^2 = x^3 + ax^2 + bx + c.$$

If $(x,y) \in \mathcal{E}$ then

$$p(x) = x^3 + ax^2 + bx + c$$

must be a quadratic residue mod$p$. Of the numbers $\{1, 2, \ldots, p-1\}$ just $(p-1)/2$ are quadratic residues, namely

$$(\pm 1)^2, (\pm 2)^2, \ldots, (\pm (p-1)/2)^2.$$

Thus if the values of $p(x)$ mod p are randomly distributed, the expectation would be that $p(x) = 0$ for one x, and that $p(x)$ would be a quadratic residue for $(p-1)/2$ values of x. The former would give one point $(x, 0)$ on the curve; each of the latter would give two points $(x, \pm y)$. Thus the expected number of points is

$$1 + 2\frac{p-1}{2} = p.$$

To this must be added the point $O = [0, 1, 0]$, giving $p + 1$ points in all.

Definition 7.5 *We set*

$$a(p) = \|\mathcal{E}(\mathbf{GF}p)\| - (p+1).$$

Thus $a(p)$ measures the discrepancy from the expected value.

Hasse showed that

$$|a(p)| < 2\sqrt{p}$$

for all elliptic curves over $\mathbf{GF}p$, ie

$$p + 1 - 2\sqrt{p} < a(p) < p + 1 + 2\sqrt{p}.$$

For example, if $\mathcal{E}$ is an elliptic curve over $\mathbf{GF}7$ then

$$5 \le \|\mathcal{E}(\mathbf{GF}7)\| \le 11.$$

Although the proof of Hasse's Theorem is not particularly difficult, it would take us too far afield to give it here.

Suppose $\mathcal{E}(\mathbb{Q})$ is an elliptic curve. Then $a(p)$ is defined for each good prime p. Shimura conjectured that there was a modular form $f(z)$ associated to $\mathcal{E}$ with the property that the $a(p)$ were the coefficients of the corresponding Fourier series $g(q)$. (See Chapter 8.)

Wiles proved Shimura's Conjecture for semi-stable elliptic curves, that is, those for which the bad primes were at worst semi-stable (ie no cusps). This was the main step in his proof of Fermat's Last Theorem.

Late last year, Shimura's Conjecture was proved for all elliptic curves over $\mathbb{Q}$.

Chapter 8

The Complex Case

8.1 Periods and Lattices

We shall be concerned in this Chapter exclusively with *meromorphic* functions on $\mathbb{C}$, the space of complex numbers. Recall that a complex function $f(z)$ is said to be meromorphic on $\mathbb{C}$ if it is defined and regular at all points of $\mathbb{C}$ except for a discrete set of points, at each of which it has a pole of finite order.

Every rational function $P(z)/Q(z)$ (where $P(z), Q(z)$ are polynomials) is meromorphic on $\mathbb{C}$, as are the trigonometric functions $\cos z, \sin z, \tan z$, the exponential function e^z, etc.

Definition 8.1 *The meromorphic function $f(z)$ on $\mathbb{C}$ is said to have period $\omega \in \mathbb{C}$ if*

$$f(z + \omega) = f(z)$$

whenever $f(z)$ is defined.

Proposition 8.1 *The periods of a non-constant meromorphic function $f(z)$ form a discrete subgroup of the abelian group $\mathbb{C}$.*

Proof ► If ω_1, ω_2 are periods of $f(z)$ then so are $\omega_1 \pm \omega_2$. Hence the periods form a subgroup of $\mathbb{C}$.

To prove that the subgroup is discrete, we have to show that there exists a constant $C > 0$ such that $f(z)$ has no period $|\omega| < C$ except for $\omega = 0$. To this end, consider the behaviour of $f(z)$ in the neighbourhood of a regular point z_0. In some neighbourhood of this point, $f(z)$ has an expansion

$$f(z) = c_0 + c_1(z - z_0) + c_2(z - z_0)^2 + \cdots.$$

This power-series will be dominated by its first non-zero term, and it is easy to deduce that

$$0 < |z - z_0| < C \Longrightarrow f(z) \neq c_0$$

for some constant $C > 0$. It follows that there is no non-zero period with $|\omega| < C$. ◄

Note that as an abelian group, $\mathbb{C} \cong \mathbb{R}^2$.

Proposition 8.2 *A discrete subgroup of $\mathbb{R}^n$ is isomorphic to $\mathbb{Z}^m$ for some $m \leq n$.*

Proof ► Suppose S is a discrete subgroup of $\mathbb{R}^n$. Let $V = \langle S \rangle$ be the vector subspace of $\mathbb{R}^n$ spanned by the elements of S. We argue by induction on $m = \dim V$, showing that S has a Z-basis with m elements.

Let $s_1, \ldots, s_m \in S$ be a basis for V; and let

$$U = \langle s_1, \ldots, s_{m-1} \rangle.$$

By our inductive hypothesis,

$$S' = S \cap U$$

has a Z-basis with $m-1$ elements, say $t_1, \ldots, t_{m-1}$.

Suppose $s \in S$. Clearly $t_1, \ldots, t_{m-1}, s_m$ is a basis for V. Let

$$s = \lambda_1 t_1 + \cdots + \lambda_{m-1} t_{m-1} + \lambda_m s_m.$$

We claim that there is an $s \in S$ minimizing $|\lambda_m|$. For we can find $n_1, \ldots, n_{m-1} \in \mathbb{Z}$ such that

$$|\lambda_i - n_i| \leq \frac{1}{2} \qquad (1 \leq i \leq m-1);$$

and then

$$\begin{aligned} s' &= s - (n_1 t_1 + \cdots + n_{m-1} t_{m-1} \\ &= (\lambda_1 - n_1) t_1 + \cdots + (\lambda_{m-1} - n_{m-1}) t_{m-1} + \lambda_m s_m. \end{aligned}$$

We may assume that $\lambda_m \leq 1$, since s_m is a contender for minimal s. Thus s' has the same λ_m as s and

$$|s'| \leq |t_1| + \cdots + |t_{m-1}| + |s_m| = R,$$

say.

But since S is a *discrete* subgroup, it has only a finite number of elements in the compact disk $|v| \leq R$. Thus we need only consider a finite number of elements $s \in S$ when minimizing $|\lambda_m|$; and so the minimum is certainly attained, at t_m say.

Now suppose $s \in S$. Evidently $t_1, \ldots, t_{m-1}, t_m$ is a basis for V. Hence

$$s = \mu_1 t_1 + \cdots + \mu_{m-1} t_{m-1} + \mu_m t_m.$$

But now we can find $n_m \in \mathbb{Z}$ such that

$$|\mu_m - n_m| \leq \frac{1}{2}.$$

and then

$$\begin{aligned} s' &= s - n_m t_m \\ &= \mu_1 t_1 + \cdots + \mu_{m-1} t_{m-1} + (\mu_m - n_m) t_m \end{aligned}$$

has smaller s_m component than t_m, contradicting the minimality of t_m unless $\mu_m = n_m$, ie $\mu_m \in \mathbb{Z}$.

But now $s' \in S' = S \cap U$; and therefore $\mu_1, \ldots, \mu_{m-1} \in \mathbb{Z}$, by our inductive hypothesis.

We conclude that $t_1, \ldots, t_m$ is a Z-basis for S. ◄

Corollary 4 *A non-trivial discrete subgroup of the additive group $\mathbb{C}$ is isomorphic either to $\mathbb{Z}$ or to $\mathbb{Z} \oplus \mathbb{Z}$.*

Definition 8.2 *A* lattice *in C is a discrete subgroup $\Lambda \subset \mathbb{C}$ isomorphic to $\mathbb{Z} \oplus \mathbb{Z}$.*

Every lattice has a basis λ, μ. This basis is not unique. In fact it is easy to see that

$$\lambda' = a\lambda + b\mu, \ \mu' = c\lambda + d\mu \qquad (a, b, c, d \in \mathbb{Z})$$

will form a basis if and only if $ad - bc = \pm 1$, ie

$$\begin{pmatrix} \lambda' \\ \mu' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$$

where

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \pm 1.$$

Proposition 8.3 *If λ, μ is a basis for the lattice Λ then $\lambda/\mu \notin \mathbb{R}$.*

Proof ► Suppose first that $\lambda/\mu \in \mathbb{Q}$, say

$$\lambda/\mu = m/n.$$

Then

$$n\lambda = m\mu,$$

ie λ, μ are not linearly independent.

(Alternatively, we may suppose that $\gcd(m, n) = 1$. Then there exist $a, b \in \mathbb{Z}$ such that

$$am + bn = 1$$

Thus

$$a\lambda + b\mu = \mu/n \in \Lambda,$$

and

$$\lambda = m(\mu/n),\ \mu = n(\mu/n),$$

ie λ and μ are both multiples of a smaller period.)

Now suppose that

$$\lambda/\mu \in \mathbb{R} \setminus \mathbb{Q},$$

ie the ratio is real but irrational.

Lemma 6 *If α is irrational then given any $\epsilon > 0$ we can find $m, n \in \mathbb{Z}$ such that*

$$|m\alpha - n| < \epsilon.$$

Proof of Lemma ▷ Choose N with $1/N < \epsilon$. For $x \in \mathbb{R}$, let $\{x\}$ denote the fractional part of x, ie

$$\{x\} = x - [x].$$

Consider the $N + 1$ fractional parts

$$0, \{\alpha\}, \{2\alpha\}, \ldots, \{N\alpha\} \in [0, 1).$$

Divide the interval $[0, 1)$ into N equal parts,

$$[0, 1/N), [1/N, 2/N), \ldots [(N-1)/N, 1).$$

By the Pigeon-Hole Principle, two of the fractional parts, say $\{r\alpha\}, \{s\alpha\}$, must lie in the same subinterval. But then

$$|\{r\alpha\} - \{s\alpha\}| < 1/N < \epsilon,$$

ie

$$|r\alpha - [r\alpha] - (s\alpha - [s\alpha])|\epsilon,$$

ie

$$|m\alpha - n| < \epsilon,$$

where $m = r - s,\ n = [r\alpha] - [s\alpha]$. ◁

By the Lemma, we can find $m, n \in \mathbb{Z}$ such that

$$|m(\lambda/\mu) - n| < \epsilon.$$

Hence

$$|m\lambda - n\mu| < \epsilon\mu.$$

Thus we can find lattice points $m\lambda - n\mu \in \Lambda$ arbitrarily close to 0, contradiction the condition that Λ be discrete. ◄

Definition 8.3 *Suppose $\Lambda \subset \mathbb{C}$ is a lattice. An* elliptic function *$f(z)$ with respect to Λ is a meromorphic function whose periods include all elements of Λ, ie*

$$\omega \in \Lambda \Longrightarrow f(z + \omega) \equiv f(z).$$

Suppose λ, μ is a basis for the lattice $\Lambda \subset \mathbb{C}$. Then $f(z)$ is elliptic with respect to Λ if and only if

$$f(z + \lambda) = f(z),\ f(z + \mu) = f(z).$$

In other words *an elliptic function is just a doubly-periodic function.*

Definition 8.4 *A* fundamental parallelogram *for the lattice $\Lambda \subset \mathbb{C}$ is a set*

$$\Pi(\lambda, \mu, c) = \{z \in \mathbb{C} : z = c + x\lambda + y\mu : 0 \le x, y < 1\},$$

where λ, μ is a basis for Λ, and $c \in \mathbb{C}$.

Suppose Π is a fundamental parallelogram for the lattice Λ. Then each $z \in \mathbb{C}$ is *congruent modulo* Λ to a unique point $z_0 \in \Pi$:

$$z \equiv z_0 \bmod \Lambda,$$

by which we mean that

$$z - z_0 \in \Lambda.$$

(Notice that we excluded 2 sides of the parallelogram, to ensure uniqueness.)

8.2 Applications of Cauchy's Theorem

Let us recall some fundamental results from complex analysis:

1. Cauchy's Theorem, the fundamental result of complex analysis, states that *if the function $f(z)$ is continuous on and holomorphic within the Jordan curve C then*
$$\int_C f(z)dz = 0.$$

2. Suppose $f(z)$ has a pole of order n at $z = b$, so that it has an expansion
$$f(z) = \frac{c_{-n}}{(z-b)^n} + \cdots + \frac{c_{-1}}{z-b} + c_0 + \cdots$$
in a neighbourhood of b. Then the *residue* of $f(z)$ at b is defined to be c_1. *Suppose $f(z)$ is continuous on and meromorphic within C; and suppose $f(z)$ has poles at $b_1, b_2, \ldots, b_r$ inside C, with residues $c_1, c_2, \ldots, c_r$. Then*
$$\frac{1}{2\pi i}\int_C f(z)dz = c_1 + c_2 + \cdots + c_r.$$

3. *Suppose $f(z)$ is continuous on and regular within C; and suppose a is inside C. Then*
$$f(a) = \frac{1}{2\pi i}\int_C \frac{f(z)}{z-a}dz,$$
and
$$f'(a) = \frac{1}{2\pi i}\int_C \frac{f(z)}{(z-a)^2}dz.$$
Informally, the second result is derived from the first by differentiating with respect to a under the integral sign.

4. *Liouville's Theorem: Suppose $f(z)$ is regular and bounded on $\mathbb{C}$. Then $f(z)$ is a constant.* For let us take C to be a large circle centered on a with radius R; and let us suppose that $|f(z)| \le c$. Then
$$|f'(a)| \le \frac{1}{2\pi}\frac{2\pi R}{R^2} = \frac{c}{R}.$$
Since R is arbitrary it follows that $f'(a) = 0$ for all a, and so $f(z)$ is constant.

5. *Suppose the meromorphic function $f(z)$ has zeros at $a_1, a_2, \ldots, a_r$ and poles at $b_1, b_2, \ldots, b_s$ inside C; and suppose $f(z)$ has no poles or zeros on C. Then*
$$\frac{1}{2\pi i}\int_C \frac{f'(z)}{f(z)}dz = r - s,$$
with the understanding that poles and zeros are counted with appropriate multiplicity, eg a double zero is counted twice. For the function $f'(z)/f(z)$ has a simple pole with residue d at a zero of order d, and a simple pole with residue $-d$ at a pole of order d.

6. *With the same assumptions,*
$$\frac{1}{2\pi i}\int_C z\frac{f'(z)}{f(z)}dz = (a_1 + \cdots + a_r) - (b_1 + \cdots + b_s).$$
For if $f(z)$ has a zero at a of order m then $zf'(z)/f(z)$ has a simple pole at a with residue ma; while if $f(z)$ has a pole at b of order n then $zf'(z)/f(z)$ has a simple pole at b with residue $-nb$.

7. *If each of the functions $u_n(z)$ is holomorphic in the open set $U \subset \mathbb{C}$ and $\sum u_n(z)$ is uniformly convergent in U then*
$$f(z) = \sum u_n(z)$$
is holomorphic in U, with
$$f'(z) = \sum u_n'(z).$$
Notice that this is much simpler to prove than the corresponding result for real functions, using the fact that
$$f(a) = \frac{1}{2\pi i}\int_C \frac{f(z)}{z-a}dz,$$

8. With the same assumptions, *if C is a contour inside U then*
$$\int_C f(z)dz = \sum \int_C u_n(z)dz.$$

In applying these results to elliptic functions, we usually take a fundamental parallelogram Π for C. Note that if $f(z)$ is elliptic then
$$\frac{1}{2\pi i}\int_\Pi f(z)dz = 0,$$
since the contributions of opposite sides will cancel out.

Proposition 8.4 *An elliptic function $f(z)$ with no poles is necessarily constant.*

Proof ► Let Π be a fundamental parallelogram. Then $f(z)$ is bounded on Π, say $|f(z)| \leq C$, since a continuous function is always bounded on a compact set. But then $f(z)$ is bounded on the whole of $\mathbb{C}$, since we can always find $z_0 \in \Pi$ with $z \equiv z_0 \bmod \Lambda$ and then $|f(z)| = |f(z_0)| \leq C$.

It follows by Liouville's Theorem that $f(z)$ is constant. ◄

Proposition 8.5 *Suppose $f(z)$ is an elliptic function; and suppose Π is a fundamental parallelogram, containing no poles or zeros of $f(z)$ on its boundary. Then the number of poles of $f(z)$ inside Π is equal to the number of zeros inside Π, each counted according to its multiplicity.*

Proof ► This follows at once from the fact that

$$\frac{1}{2\pi i}\int_{\Pi} \frac{f'(z)}{f(z)} dz = r - s.$$

For since $f'(z)/f(z)$ is elliptic, the integral is 0, as explained above. ◄

Corollary 5 *An elliptic function cannot have a singe simple pole inside Π.*

Proof ► By the Proposition, the residue c at a single pole must vanish. But a simple pole cannot have zero residue. ◄

Thus an elliptic function has to have at least 2 poles (or a double pole) in each fundamental parallelogram.

Proposition 8.6 *Suppose $f(z)$ is an elliptic function; and suppose Π is a fundamental parallelogram, containing no poles of $f(z)$ on its boundary. Let the residues of the poles inside Π be $c_1, \ldots, c_r$. Then*

$$c_1 + \cdots + c_r = 0.$$

Note that in this case the poles are *not* counted according to their multiplicity.

Proof ► This follows at once from the fact that

$$\frac{1}{2\pi i}\int_{\Pi} f(z) dz = 0.$$

◄

Proposition 8.7 *Suppose $f(z)$ is an elliptic function; and suppose Π is a fundamental parallelogram, containing no poles or zeros of $f(z)$ on its boundary. Let the zeros of $f(z)$ inside Π be $a_1, \ldots, a_r$, and let the poles inside Π be $b_1, \ldots, b_r$ (each repeated according to its multiplicity). Then*

$$a_1 + \cdots + a_r \equiv b_1 + \cdots + c_r \bmod \Lambda.$$

Proof ► From above,

$$\frac{1}{2\pi i}\int_{\Pi} z\frac{f'(z)}{f(z)} dz = (a_1 + \cdots + a_r) - (b_1 + \cdots + b_s).$$

Thus the result will be proved if we can show that

$$\frac{1}{2\pi i}\int_{\Pi} z\frac{f'(z)}{f(z)} dz \in \Lambda$$

The function $g(z) = zf'(z)/f(z)$ is not elliptic; but

$$g(z+\lambda) - g(z) = \lambda\frac{f'(z)}{f(z)},\ g(z+\mu) - g(z) = \mu\frac{f'(z)}{f(z)}.$$

Thus the sides $[c, c+\mu]$ and $[c+\lambda+\mu, c+\lambda]$ together contribute

$$\frac{1}{2\pi i}\int_{c}^{c+\mu} \lambda\frac{f'(z)}{f(z)} = \frac{\lambda}{2\pi i}\left[\log f(z)\right]_{c}^{c+\mu}.$$

Since $f(c+\lambda) = f(c)$, the function $\log f(z)$ differs at c and $c+\mu$ by $2m\pi i$ for some $m \in \mathbb{Z}$. Thus these 2 sides together contribute $\pm m\lambda$. Similarly the other 2 sides contribute $\pm n\mu$ for some $n \in \mathbb{Z}$. Hence

$$\frac{1}{2\pi i}\int_{\Pi} z\frac{f'(z)}{f(z)} dz = \pm m\lambda + \pm n\mu \in \Lambda.$$

◄

8.3 Weierstrass' Elliptic Function

We have established several properties of elliptic functions. But we have yet to establish that any non-constant elliptic functions exist.

Proposition 8.8 *Suppose $\Lambda \subset \mathbb{C}$ is a lattice. The series*

$$\sum_{\omega\in\Lambda, \omega\neq 0} \frac{1}{|\omega|^e}$$

converges if and only if $e > 2$

Proof ▶ Let λ, μ be a basis for the lattice Λ, so that

$$\omega = m\lambda + n\mu \qquad (m, n \in \mathbb{Z}).$$

Lemma *There are constants C_1, C_2 such that*

$$C_1(m^2 + n^2) \le |m\lambda + n\mu|^2 \le C_2(m^2 + n^2).$$

Proof of Lemma ▷ For $x, y \in \mathbb{R}$,

$$Q(x,y) = |x\lambda + y\mu|^2 = (x\bar{\lambda} + y\bar{\mu})(x\lambda + y\mu) = Ax^2 + 2Bxy + cy^2$$

is a positive-definite quadratic form. Hence

$$Q(x,y) - C_1(x^2 + y^2)$$

is still positive-definite for sufficiently small C_1, and so

$$C_1(x^2 + y^2) \le Q(x,y).$$

On the other hand, $|2xy| \le x^2 + y^2$, and so

$$Q(x,y) \le (A + B + C)(x^2 + y^2).$$

◁

Geometrically, this Lemma states that concentric circles can be drawn inside and outside an ellipse.

Lemma *The series*

$$\sum_{(m,n)\neq(0,0)} \frac{1}{(m^2+n^2)^e}$$

is convergent if and only if $e > 1$.

Proof of Lemma ▷ We compare the sum S with the integral

$$I = \int_0^\infty \int_0^\infty \frac{dx\,dy}{(x^2+y^2)^e}.$$

Changing to polar coordinates,

$$\begin{aligned} I &= \int_0^\infty \int_0^{2\pi} \frac{r dr\, d\theta}{r^{2e}} \\ &= 2\pi \int_0^{2\pi} \int_0^\infty r^{1-2e} dr. \end{aligned}$$

This converges if and only if $1 - 2e < -1$, ie $e > 1$.

To see that S and I converge or diverge together, we note that if $m \ge 0$, $n \ge 0$ then

$$\frac{1}{((m+1)^2 + (n+1)^2)^e} \le \frac{1}{(x^2+y^2)^e} \le \frac{1}{(m^2+y^2)^e}$$

for $m \le x \le m+1$, $n \le y \le n+1$. We leave the completion of the argument, dealing with the terms along the axes, as an exercise. ◁ ◀

Definition 8.5 *For $n = 2, 3, 4, \ldots$ we set*

$$g_n = \sum_{\omega\in\Lambda, \omega\neq 0} \frac{1}{\omega^{2n}}.$$

Note that the sums of odd powers all vanish,

$$\sum_{\omega\in\Lambda, \omega\neq 0} \frac{1}{\omega^{2n+1}} = 0$$

for $n = 1, 2, 3, \ldots$, since the terms in ω and $-\omega$ cancel out.

Proposition 8.9 *The series*

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

is absolutely convergent for each $z \notin \Lambda$, and defined a meromorphic function of $\mathbb{C}$ with a double pole at each $\omega \in \Lambda$.

Proof ▶ Suppose $|\omega| \ge 2|z|$, ie $|z| \le \frac{1}{2}\omega$. Now

$$\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} = \frac{z(2\omega - z)}{\omega^2(\omega-z)^2}.$$

But $|\omega - z| \ge \frac{1}{2}|\omega|$, while $|2\omega - z| \le 3|\omega|$. Hence

$$|\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}| \le \frac{2|z|^3}{|\omega|}.$$

Since $\sum 1/|\omega|^3$ is convergent, it follows that the series

$$\sum_{|\omega|\ge 2C} \frac{1}{(z-\omega)^2} - \frac{1}{\omega^2}$$

is uniformly absolutely convergent — and so defines a holomorphic function — in $|z| \le C$; and the result follows. ◀

Definition 8.6 *The Weierstrass elliptic function $\wp(z)$ with respect to the lattice $\Lambda \subset \mathbb{C}$ is defined by*

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

Proposition 8.10 *The function $\wp(z)$ is elliptic with respect to Λ.*

Proof ► We have to show that if $\omega_0 \in \Lambda$ then

$$f(z+\omega_0) = f(z).$$

The result would be obvious if we could separate $\wp(z)$ into a variable part $1/z^2 + \sum 1/(z-\omega)^2$ and a constant part $\sum 1/\omega^2$. Unfortunately these 2 parts do not converge separately, so a more careful approach—which we sketch below—is required.

Given $\epsilon > 0$, choose R so large that

$$\sum_{|\omega| \geq R} \frac{1}{|\omega|^3} < \epsilon \text{ and } \sum_{|\omega| \geq R} \frac{1}{|z-\omega|^3} < \epsilon;$$

and let

$$\wp(z) = F(z) + R(z),$$

where

$$F(z) = \frac{1}{z^2} + \sum_{|\omega| \leq R + |z| + |\omega_0|} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

and

$$R(z) = \sum_{|\omega| > R + |z| + |\omega_0|} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

Then

$$\wp(z+\omega_0) - \wp(z) = F(z+\omega_0) - F(z) + R(z+\omega_0) - R(z).$$

All the terms in $F(z+\omega_0) - F(z)$ cancel out, except some corresponding to ω satisfying $|\omega| > R$. The contribution of these will be $< \epsilon$, as will $|R(z)|$ and $R(z+\omega_0)|$. Hence

$$|\wp(z+\omega_0) - \wp(z)| < 3\epsilon.$$

Since ϵ can be taken arbitrarily small, the result follows. ◄

8.4 The Field of Elliptic Functions

Proposition 8.11 *$\wp(z)$ is even.*

Proof ► This follows at once from the definition of $\wp(z)$ ◄

Corollary 6 *$\wp'(z)$ is odd.*

Proposition 8.12 *The elliptic functions form with respect to Λ form a field over $\mathbb{C}$, of which the even functions form a sub-field.*

Proof ► If $f(z), g(z)$ are elliptic with respect to Λ, then so are $f(z) \pm g(z)$, $f(z)g(z)$ and $f(z)/g(z)$; and the same is true if $f(z), g(z)$ are even. ◄

Definition 8.7 *We say that $\sigma \in \mathbb{C}$ is a* semilattice point *with respect to the lattice Λ if $2\sigma \in \Lambda$ but $\sigma \notin \Lambda$.*

There are evidently three classes of semilattice points $\bmod \Lambda$, represented by $\lambda/2$, $\mu/2$ and $(\lambda+\mu)/2$.

Proposition 8.13 *An odd elliptic funtion $f(z)$ has a pole or zero at every semilattice point σ.*

Proof ► Suppose σ is not a pole of $f(z)$. Since

$$2\sigma = \omega \in \Lambda$$

ie

$$-\sigma = \sigma - \omega,$$

it follows that

$$f(-\sigma) = f(\sigma - \omega) = f(\sigma).$$

On the other hand, since $f(z)$ is odd.

$$f(-\sigma) = -f(\sigma).$$

Hence

$$f(\sigma) = 0,$$

ie σ is a zero of $f(z)$. ◄

Corollary 7 *Suppose $f(z)$ is an even elliptic funtion. If the semilattice point σ is a pole or zero of $f(z)$ then it is a pole or zero of even order.*

Proof ▶ Suppose σ is a zero of $f(z)$. Since $f^1(z) = f'(z)$ is odd, $f^2(z) = f''(z)$ is even, $f^3(z)$ is odd, etc,

$$f^{(1)}(\sigma) = f^{(3)}(\sigma) = f^{(5)}(\sigma) = \cdots = 0.$$

Thus the first n for which $f^{(n)}(\sigma) \neq 0$ is even. Hence the ordero of the zero is even.

If $f(z)$ has a pole at σ then the result follows on considering $1/f(z)$. ◀

Theorem 8.1 *The field k of even elliptic functions with respect to Λ is generated over $\mathbb{C}$ by the Weierstrass elliptic function: $k = \mathbb{C}(\wp(z))$. In other words, every elliptic function $f(z)$ is expressible as a rational function of $\wp(z)$:*

$$f(z) = \frac{P(\wp(z))}{Q(\wp(z))},$$

where P, Q are polynomials.

Proof ▶ If $f(z)$ has a pole or zero at 0, it must have even multiplicity since $f(z)$ is even. Thus we can find $e \in \mathbb{Z}$ such that

$$g(z) = \wp(z)^e f(z)$$

has no pole or zero at 0.

Suppose $g(z)$ has zeros $a_1, \ldots, a_r$ and poles $b_1, \ldots, b_r$ in the fundamental parallelogram Π. If a is a zero of $g(z)$ then so is $-a \bmod \Lambda$. Moreover if $-a \equiv a \bmod \Lambda$ then the zero is of even order, by the Corollary to Proposition 8.13. Thus the zeros can be divided into pairs $\pm a_1, \ldots, \pm a_t$, where $2t = r$. Similarly the poles can be divided into pairs $\pm b_1, \ldots, \pm b_t$.

The function $\wp(z) - \wp(a)$ has just 2 zeros in Π, at $\pm a \bmod \Lambda$. It follows that we can 'eliminate' poles or zeros at $\pm a$ by multiplying or dividing by $\wp(z) - \wp(a)$. Thus

$$g(z)\frac{(\wp(z) - \wp(a_1)) \ldots (\wp(z) - \wp(a_t))}{(\wp(z) - \wp(b_1)) \ldots (\wp(z) - \wp(b_t))}$$

has neither poles nor zeros, and so is constant. Hence

$$f(z) = c\wp(z)^{-e}\frac{(\wp(z) - \wp(b_1)) \ldots (\wp(z) - \wp(b_t))}{(\wp(z) - \wp(a_1)) \ldots (\wp(z) - \wp(a_t))}.$$

◀

Proposition 8.14 *Every elliptic function $f(z)$ is expressible in the form*

$$f(z) = R(\wp(z)) + \phi'(z)S(\wp(z)),$$

where R and S are rational functions.

Proof ▶ We can split $f(z)$ into even and odd parts:

$$\begin{aligned} f(z) &= \frac{f(z) + f(-z)}{2} + \frac{f(z) - f(-z)}{2} \\ &= F(z) + G(z), \end{aligned}$$

where $F(z)$ is even and $G(z)$ is odd. But then

$$H(z) = G(z)/\wp'(z)$$

is even, and so

$$f(z) = F(z) + \wp'(z)H(z),$$

where $F(z)$ and $H(z)$ are both even elliptic functions. The result now follows from the previous Proposition. ◀

Corollary 8 *The field K of elliptic functions with respect to Λ is generated over $\mathbb{C}$ by $\wp(z)$ and $\wp'(z)$:*

$$K = \mathbb{C}\left(\wp(z), \wp'(z)\right).$$

8.5 The Functional Equation

Since $\wp'(z)$ is odd, $\wp'(z)^2$ is even and so can be expressed as a rational function of $\wp(z)$:

$$\wp'(z)^2 = R(\wp(z))$$

by our argument above. In fact we shall see that R is a cubic polynomial.

Proposition 8.15 *The function $\wp(z)$ satisfies the functional equation*

$$\wp'(z)^2 = 4\left(\wp(z) - \wp(\sigma_1)\right)\left(\wp(z) - \wp(\sigma_2)\right)\left(\wp(z) - \wp(\sigma_3)\right),$$

where $\sigma_1, \sigma_2, \sigma_3$ are semilattice points in distinct classes $\bmod \Lambda$ (eg $\sigma_1 = \lambda/2,\ \sigma_2 = \mu/2,\ \sigma_3 = (\lambda + \mu)/2$).

Proof ► The function on the left has a 6-fold pole at $z=0$, and double zeros at each semilattice point. The function on the right also has a 6-fold pole at $z=0$. Consider the function $f(z)=\wp(z)-\wp(e_i)$. This has a zero at e_i; and it is a double zero since $f'(e_i)=\wp'(e_i)=0$.

Thus the function on the right has exactly the same poles and zeros as the function on the left. Hence they differ only by a multiplicative constant (since their ratio has no poles or zeros).

The value of this constant follows on considering the coefficients of $1/z^6$ on both sides:

$$\begin{aligned}\wp(z)=\frac{1}{z^2}+h(z) &\Longrightarrow \wp'(z)=-\frac{2}{z^3}+O(z)\\ &\Longrightarrow \wp'(z)^2=\frac{4}{z^6}+O(\frac{1}{z^2}.\end{aligned}$$

◄

Theorem 8.2 *The functional equation satisfied by $\wp(z)$ takes the form*

$$\wp'(z)^2=4\wp(z)^3-60g_2\wp(z)-140g_3,$$

where

$$g_2=\sum_{w\in\Lambda,w\neq 0}\frac{1}{w^4},\ g_3=\sum_{w\in\Lambda,w\neq 0}\frac{1}{w^6}.$$

Proof ► We know that $\wp(z)$ satisfies a functional equation of the form

$$\wp'(z)^2=4\wp(z)^3+a\wp(z)^2+b\wp(z)+c.$$

To determine the coefficients a,b,c we consider the leading terms in the expansions of $\wp(z)$ and $\wp'(z)$ about $z=0$. We have

$$\begin{aligned}\frac{1}{(z-\omega)^2}-\frac{1}{\omega^2} &= \frac{1}{\omega^2(1-z/\omega)^2}-\frac{1}{\omega^2}\\ &= \frac{1}{\omega^2}\left(1+\frac{2z}{\omega}+\frac{3z^2}{\omega^2}+\cdots\right)-\frac{1}{\omega^2}\\ &= \frac{2z}{\omega^3}+\frac{3z^2}{\omega^4}+\cdots.\end{aligned}$$

Thus

$$\begin{aligned}\wp(z) &= \frac{1}{z^2}+2z\sum_{\omega\neq 0}\frac{1}{\omega^3}+3z^2\sum_{\omega\neq 0}\frac{1}{\omega^4}+\cdots\\ &= \frac{1}{z^2}+3g_2z^2+5g_3z^4+O(z^6).\end{aligned}$$

Differentiating,

$$\wp'(z)=-\frac{2}{z^3}+6g_2z+20g_3z^3+O(z^5).$$

Thus

$$\wp'(z)^2=\frac{4}{z^6}-\frac{24g_2}{z^2}-80g_3+O(z^2),$$

while

$$\wp(z)^3=\frac{1}{z^6}+\frac{9g_2}{z^2}+15g_3+O(z^2),$$

and

$$\wp(z)^2=\frac{1}{z^4}+6g_2+O(z^2),$$

Substituting in the functional equation,

$$\frac{4}{z^6}+\frac{24g_2}{z^2}+80g_3=\frac{4}{z^6}+\frac{36g_2}{z^2}+60g_3+\frac{a}{z^4}+6ag_2+\frac{b}{z^2}+c+O(z^2).$$

Comparing coefficients,

$$a=0,\ b=-60g_2,\ c=-140g_3,$$

as stated. ◄

8.6 Geometrical Interpretation

The functional equation can be interpreted as saying that the point $(\wp(z),\wp'(z))$ lies on the elliptic curve

$$y^2=4x^3-60g_2x-140g_3$$

for all $z\in\mathbb{C}\setminus\Lambda$. If $z\in\Lambda$ then $\wp(z)$ and $\wp'(z)$ are both undefined. We assign them to the infinite point $[0,1,0]$ on $\mathcal{E}$.

Proposition 8.16 *The map $\Phi:\mathbb{C}\to\mathcal{E}(\mathbb{C})$ defined by*

$$z\mapsto\begin{cases}[\wp(z),\wp'(z),1] & \text{if } z\notin\Lambda,\\ [0,1,0] & \text{if } z\in\Lambda,\end{cases}$$

is surjective and continuous; and

$$\Phi(z_1)=\Phi(z_2)\Longleftrightarrow z_1\equiv z_2 \bmod \Lambda.$$

Proof ► Suppose $(x, y) = [x, y, 1] \in \mathcal{E}$. Consider the elliptic function

$$f(z) = \wp(z) - x.$$

This has a double pole at the points of Λ, and so has two zeros in any fundamental parallelogram Π. Since $f(z)$ is even, the two zeros are $\pm a$ mod Λ. But there are just two points $(x, \pm y)$ on $\mathcal{E}$ with a given x-coordinate. It follows that each point $(x, y) \in \mathcal{E}$ arises from some $z \in \mathbb{C}$, ie Φ is surjective.

Since $\wp(z)$ and $\wp'(z)$ are both doubly-periodic,

$$z_1 \equiv z_2 \bmod \Lambda \Longrightarrow \Phi(z_1) = \Phi(z_2).$$

Conversely, if $\wp(z_1) = \wp(z_2)$ then the argument above shows that $z_1 \equiv \pm z_2 \bmod \Lambda$. Since $\wp'(-z) = -\wp'(z)$, it follows that

$$\Phi(z_1) = \Phi(z_2) \Longrightarrow z_1 \equiv z_2 \bmod \Lambda.$$

The map Φ is certainly continuous at all points $z \notin \Lambda$, since $\wp(z)$ and $\wp'(z)$ are both differentiable, and so a fortiori continuous. It remains to show that Φ is continuous at 0. In the neighbourhood of $0 \in \mathcal{E}$,

$$(\wp(z), \wp'(z)) = \left(\frac{1}{z^2} + \cdots, \frac{-2}{z^3} + \cdots\right).$$

Changing to X, Z coordinates, where $[x, y, 1] = [X, 1, Z]$, ie

$$X = \frac{x}{z},\ Z = \frac{1}{z},$$

we see that

$$X = z + O(z^3),\ Z = -\frac{1}{2}z^3 + O(z^5).$$

It follows that Φ is continuous at 0, and so at the other points of Λ. ◄

Corollary 9 *The map Φ induces a homeomorphism*

$$\mathcal{E}(\mathbb{C}) \cong \mathbb{C}/\Lambda.$$

Let λ, μ be a basis for Λ. The quotient-group $\mathbb{C}/\Lambda$ is homeomorphic to the torus $\mathbb{T}^2$, under the map

$$(x \bmod 1,\ y \bmod 1) \mapsto x\lambda + y\mu \bmod \Lambda.$$

Since this map preserves addition, it is in fact an isomorphism of topological groups:

$$\mathbb{C}/\Lambda = \mathbb{T}^2.$$

Thus we have a homeomorphism

$$\mathbb{T}^2 \to \mathcal{E}(\mathbb{C}) : (x \bmod 1,\ y \bmod 1) \mapsto (\wp(x\lambda + y\mu),\ \wp'(x\lambda + y\mu)).$$

This leaves the question: is this map a group isomorphism? That is, does the addition on $\mathbb{C}/\Lambda$ correspond to the addition defined geometrically on $\mathcal{E}$?

8.7 The Addition Formula

Suppose $u, v \in \mathbb{C} \setminus \Lambda$, with $u \not\equiv v \bmod \Lambda$. Then we can find $A, B, C \in \mathbb{C}$ such that

$$\begin{aligned} A\wp(u) + B\wp'(u) + C &= 0 \\ A\wp(v) + B\wp'(v) + C &= 0. \end{aligned}$$

Consider the elliptic function

$$f(z) = A\wp(z) + B\wp'(z) + C.$$

This has a triple pole (at most) at each lattice point $z \in \Lambda$. Hence it has 3 zeros a_1, a_2, a_3 in any fundamental parallelogram Π, satisfying

$$a_1 + a_2 + a_3 \equiv 0 \bmod \Lambda,$$

by Proposition /refZeroPoleSum Two of these are equivalent modulo Λ to u and v. It follows that the third is $\equiv -(u + v) \bmod \Lambda$:

$$A\wp(u + v) - B\wp'(u + v) + C = 0.$$

Thus, eliminating A, B, C,

$$\det \begin{bmatrix} \wp(u+v) & -\wp'(u+v) & 1 \\ \wp(u) & -\wp'(u) & 1 \\ \wp(v) & -\wp'(v) & 1 \end{bmatrix} = 0.$$

This expresses $\Phi(u + v) = (\wp(u + v), \wp'(u + v))$ in terms of $\Phi(u)$ and $\Phi(v)$.

Proposition 8.17 *Suppose $u, v, w \in \mathbb{C}/\Lambda$; and suppose*

$$u + v + w = 0.$$

Then the corresponding points $\Phi(u), \Phi(v), \Phi(w) \in \mathcal{E}$ are collinear.

Proof ► Suppose $u, v, w \neq 0$. We have seen that there exists $(A, B, C) \neq (0, 0, 0)$ such that

$$\begin{aligned} A\wp(u) + B\wp'(u) + C &= 0 \\ A\wp(v) + B\wp'(v) + C &= 0 \\ A\wp(w) + B\wp'(w) + C &= 0. \end{aligned}$$

In other words the 3 points $\Phi(u), \Phi(v), \Phi(w)$ lie on the line

$$Ax + By + C = 0.$$

If say $u = 0$ then $v = -w$, and

$$\Phi(u) = [0, 1, 0],\ \Phi(v) = [\varphi(v), \varphi'(v), 1],\ \Phi(w) = [\varphi(v), -\varphi'(v), 1]$$

lie on the line $x = \varphi(v)z$. ◀

Corollary 10 *The map*

$$\Phi : \mathbb{C}/\Lambda \to \mathcal{E}(\mathbb{C})$$

is an isomorphism of topological abelian groups. In particular,

$$\mathcal{E}(\mathbb{C}) \cong \mathbb{T}^2.$$

In one sense this result is of little practical value, since we already know that

$$\mathcal{E}(\mathbb{R}) = \mathbb{T}^1 \text{ or } \mathbb{T}^1 \oplus \mathbb{Z}/(2),$$

and this gives us more information about $\mathcal{E}(\mathbb{Q})$. For example, the result for $\mathcal{E}(\mathbb{R})$ tells us that the torsion subgroup F, formed by the points of $\mathcal{E}(\mathbb{Q})$ of finite order, is either cyclic $\mathbb{Z}/(n)$, or else of the form $\mathbb{Z}/(2) \oplus \mathbb{Z}/(n)$. The result for $\mathcal{E}(\mathbb{C})$ only tells us that F is either cyclic $\mathbb{Z}/(n)$, or else of the form $\mathbb{Z}/(m) \oplus \mathbb{Z}/(n)$.

Perhaps the main interest of the complex case is that it explains in a natural way *why* there is a group structure on $\mathcal{E}$.

8.8 The modular group

As we have seen, each lattice $\Lambda \subset \mathbb{C}$ gives rise to an elliptic curve

$$\mathcal{E}(\mathbb{C}) : y^2 = x^3 - 15g_2x - 35g_3.$$

It is natural to ask: Does every elliptic curve over $\mathbb{C}$ arises in this way from some lattice Λ?

Suppose $s \in \mathbb{C}^\times$. Consider the lattice

$$s\Lambda = \{s\omega : \omega \in \Lambda\}.$$

We say that $\Lambda, s\Lambda$ are *similar*. Evidently

$$g_k(s\Lambda) = \sum{}' \frac{1}{(s\omega)^{2k}} = s^{-2k} g_k(\Lambda).$$

In particular, $s\Lambda$ gives rise to the elliptic curve

$$y^2 = x^3 - 15s^{-4}g_2(\Lambda)x - 35s^{-6}g_3(\Lambda).$$

But this is just the equation we get if we make the transformation

$$x \mapsto s^{-2}x,\ y \mapsto s^{-3}y,$$

since the coefficients of x and 1 in the Weierstrass equation have weights 4 and 6, respectively. Thus *similar lattices give rise to projectively equivalent elliptic curves.*

In effect, therefore, we are only concerned with lattices *up to similarity.* In other words, we are concerned with the ratio

$$\tau = \lambda/\mu$$

rather than with the basis elements λ, μ themselves. (For the lattice $\langle 1, \tau\rangle$ is similar to the lattice $\langle \lambda, \mu\rangle$.)

Recall that $\tau \notin \mathbb{R}$. Thus τ either lies in the upper half-plane

$$\mathcal{H} = \{z \in \mathbb{C} : \Im(z) > 0\}$$

or else in the lower half-plane $-\mathcal{H}$. It is convenient to restrict ourselves to bases λ, μ with $\lambda/\mu \in \mathcal{H}$. Let us say that the basis is *positive* in this case. (Note that just one of λ, μ and $-\lambda, \mu$ is positive; so we can always make a basis positive by replacing λ with $-\lambda$ if necessary.)

Recall that if λ', μ' is another basis then

$$\begin{pmatrix} \lambda' \\ \mu' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \lambda \\ \mu \end{pmatrix},$$

where $a, b, c, d \in \mathbb{Z}$ and $ad - bc = \pm 1$. On setting $\tau' = \lambda'/\mu'$ this becomes

$$\tau' = \frac{a\tau + b}{c\tau + d}.$$

The following result, although apparently rather technical, will prove very useful.

Proposition 8.18 *Suppose*

$$\tau' = \frac{a\tau + b}{c\tau + d},$$

where

$$T = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2, \mathbb{R}).$$

Then

$$\Im(\tau') = \frac{\det T}{|c\tau + d|^2} \Im(\tau).$$

Proof ► We have

$$\begin{aligned}\Im(\tau') &= \frac{1}{2i}\left(\tau' - \overline{\tau'}\right)\\ &= \frac{1}{2i}\left(\frac{a\tau+b}{c\tau+d} - \frac{a\bar\tau+b}{c\bar\tau+d}\right)\\ &= \frac{1}{2i}\frac{(ad-bc)(\tau-\bar\tau)}{(c\tau+d)(c\bar\tau+d)}\\ &= \frac{\det T}{|c\tau+d|^2}\Im(\tau).\end{aligned}$$

◄

Corollary 11 *If $\tau, \tau' \in \mathcal{H}$ then $\det T > 0$.*

Thus if we restrict ourselves to positive bases (those with $\Im(\lambda/\mu) > 0$) then we need only consider transformations

$$T \in \mathrm{SL}(2,\mathbb{Z}) = \left\{\begin{pmatrix} a & b\\ c & d\end{pmatrix} : a,b,c,d \in \mathbb{Z},\ ad-bc = 1\right\}.$$

The matrix $T \in \mathrm{SL}(2,\mathbb{Z})$ acts on $\mathcal{H}$ by

$$z \mapsto Tz = \frac{az+b}{cz+d}.$$

Notice that the matrices $\pm T$ define the same transformation.

Definition 8.8 *The* modular group G *is the quotient-group*

$$G = \mathrm{SL}(2,\mathbb{Z})/\{\pm I\}.$$

Thus the modular group G acts on the upper half-plane $\mathcal{H}$, by

$$gz = \frac{az+b}{cz+d}.$$

Each $g \in G$ arises from a pair of matrices $\pm T \in \mathrm{SL}(2,\mathbb{Z})$. By 'abuse of notation' we use the matrix T to denote g.

Definition 8.9 *We define $S, T \in G$ by*

$$S = \begin{pmatrix} 0 & -1\\ 1 & 0\end{pmatrix}, T = \begin{pmatrix} 1 & 1\\ 0 & 1\end{pmatrix},$$

Thus T corresponds to the translation

$$z \mapsto z+1,$$

while S corresponds to the inversion

$$z \mapsto -1/z.$$

We shall see shortly that S, T generate the modular group:

$$G = \langle S, T\rangle.$$

Proposition 8.19 *$S^2 = I,\ (ST)^3 = I.$*

Proof ► We have

$$\begin{aligned}S^2 &= -I\\ &= I,\end{aligned}$$

since we are working mod $\pm I$. Also

$$ST = \begin{pmatrix} 0 & -1\\ 1 & 1\end{pmatrix}$$

satisfies its characteristic equation

$$t^2 - t + 1 = 0.$$

Hence ST satisfies

$$(t+1)(t^2-t+1) = t^3+1 = 0,$$

ie

$$\begin{aligned}(ST)^3 &= -I\\ &= I,\end{aligned}$$

since we are working mod $\pm I$. ◄

8.9 The fundamental region

Definition 8.10 *We define the* fundamental region *(for the modular group) $\mathcal{F} \subset \mathcal{H}$ by*

$$\mathcal{F} = \{z \in \mathcal{H} : -\frac{1}{2} < \Re(z) \le \frac{1}{2},\ \ |z| > 1 \textit{ or } |z| = 1 \textit{ and } \Re(z) > 0\}.$$

Notice that we have included half the boundary of $\mathcal{F}$, just as we did (and for much the same reason) with the fundamental parallelogram Π for a lattice Λ.

Notice too that $\mathcal{F}$ contains the points $-\omega^2$ and i; these will play a special rôle in what follows.

Theorem 8.3 *Each point $z \in \mathcal{H}$ has a unique transform*

$$z_0 = gz \in \mathcal{F} \qquad (g \in G).$$

Remark: Note that we are not saying $g \in G$ is unique (we shall deal with that question shortly); only that z_0 is unique.

Proof ▶ The idea is to find a transform gz maximimising $\Im(gz)$. By Proposition 8.18, if

$$gz = \frac{az+b}{cz+d}$$

then

$$\Im(gz) = \frac{1}{|cz+d|^2}\Im(z).$$

For a fixed $z \in \mathcal{H}$, the points

$$\{cz + d : c, d \in \mathbb{Z}\}$$

form a lattice (with basis 1,z). There are only a finite number of lattice points inside the disk $|z| \leq 1$, ie there are only a finite number of $c, d \in \mathbb{Z}$ with

$$|cz + d| \leq 1.$$

It follows that $\Im(gz)$ can only take a finite number of values $\geq \Im(z)$. In particular there must be a maximum such value, attained say at $g_0 z$.

Now translation $z \mapsto z + r$ does not affect $\Im(z)$, so the maximal value is also attained at each pount $T^r g_0 z$.

But we can choose r so that $z_0 = T^r(g_0 z)$ lies in the strip

$$\mathcal{S} = \{z \in \mathbb{H} : -\frac{1}{2} < \Re(z) \leq \frac{1}{2}\}.$$

We claim that this transform $z_0 \in \mathcal{F}$, or else $|z_0| = 1$ and $Sz_0 \in \mathcal{F}$.

Lemma 7 *If $|z| < 1$ then*

$$\Im(Sz) > \Im(z).$$

Proof of Lemma ▷ If $z = re^{i\theta}$ then $Sz = -1/z$ and so

$$\Im(Sz) = \frac{1}{r}\sin\theta > r\sin\theta = \Im(z).$$

◁

In particular, $|z_0| \geq 1$; for otherwise $\Im(Sz_0) > \Im(z_0)$, contradicting the maximality of $\Im(z_0)$. If $|z_0| > 1$ then $z_0 \in \mathcal{F}$; while if $|z_0| = 1$ then either $\Re(z_0) \geq 0$, in which case $z_0 \in \mathcal{F}$, or else $\Re(z_0) < 0$ in which case $Sz_0 \in FF$.

Now suppose $z, gz \in \mathcal{F}$. We may assume (swapping z,gz if necessary) that

$$\Im(gz) \geq \Im(z).$$

By Proposition 8.18, this implies that

$$|cz + d| \leq 1.$$

The lowest points of $\mathcal{F}$ is

$$-\omega^2 = \frac{1}{2} + \frac{\sqrt{3}}{2}i.$$

Hence

$$|\Im(cz + d)| \geq \frac{\sqrt{3}}{2}|c|;$$

and so

$$|c| \leq 1.$$

But now cz lies in the strip $\mathcal{S}$, and so

$$|\Re(cz + d)| \geq |d| - 1/2.$$

Hence

$$|d| \leq 1.$$

The problem is reduced to just 4 cases: $(c,d) = (1,0), (0,1), (1,1), (1,-1)$.

If $c = 0$ then g is a translation

$$gz = z + r;$$

and it is clear that z, gz cannot both lie in the strip $\mathcal{S}$.

If $d = 0$ then we can take $b = 1, c = -1$, and so

$$\begin{aligned} gz &= -\frac{az+1}{z} \\ &= Sz - a. \end{aligned}$$

Now
$$z \in \mathcal{F} \Longrightarrow Sz \in \mathcal{S}.$$
Hence $a = 0$, ie $g = S$. But it is clear that
$$z, Sz \in \mathcal{F} \Longrightarrow |z| = 1;$$
while if $|z| = 1$ then
$$\Re(z) < 0 \Longleftrightarrow \Re Sz > 0.$$
So z, Sz cannot both be in $\mathcal{F}$.

It remains to consider the cases $(c, d) = (1, \pm 1)$. The function
$$|cz + d|$$
must attain its minimum on $\mathcal{F}$ at a boundary point. (It is a general principle that if the function $f(z)$ is holomorphic on the open set U then $|f(z)|$ can only attain its minimum at a point of U if this minimum is 0.) But on going round the boundary of $\mathcal{F}$, it is clear that
$$|z \pm 1| \geq 1,$$
with equality only at the corner-points $\omega, -\omega^2$. It follows that if $z, gz \in \mathcal{F}$ then
$$z = gz = -\omega^2.$$
◄

Remark: The Theorem shows that we can identify the quotient-space $\mathcal{H}/G$ with the fundamental region $\mathcal{F}$.

Suppose the group G acts on the set X. Recall that the *stabilizer* $S(x)$ of an element $x \in X$ is the subgroup
$$S(x) = \{g \in G : gx = x\}.$$

During the proof of the Theorem we almost established the following result. We leave completion of the proof to the reader.

Proposition 8.20 *1.* $S(-\omega^2) = \{I, TS, (TS)^2\}$*;*

2. $S(i) = \{I, S\}$*;*

3. If $z \in \mathcal{F}$, $z \neq -\omega^2, i$ *then* $S(z) = \{I\}$.

Theorem 8.4 *The modular group G is generated by S and T:*
$$G = \langle S, T \rangle.$$

Proof ► Let
$$H = \langle S, T \rangle$$
be the subgroup of G generated by S, T.

On examining the proof of Proposition 8.3 it is clear that the argument holds equally well with H replacing G. In particular, if $z \in \mathcal{H}$ then we can find a transform
$$hz \in \mathcal{F} \qquad (h \in H).$$

Now suppose $g \in G$. Choose any $z \in \mathcal{F}$ *except* $-\omega^2$ or i, and consider the transform gz. By Theorem 8.3 we can find $h \in H$ such that
$$h(gz) \in \mathcal{F}.$$
But then, by the same Theorem,
$$hgz = z;$$
and therefore
$$hg \in S(z) = \{I\},$$
ie
$$hg = I \Longrightarrow g = h^{-1} \in H.$$
Thus $G = H$, ie G is generated by S and T. ◄

8.10 Modular functions

Definition 8.11 *The meromorphic function $f(z)$ on $\mathcal{H}$ is said to be weakly modular of weight $2k$ (where $k \in \mathbb{Z}$) if*
$$f(gz) = (cz + d)^{-2k} f(z)$$
for each modular transformation
$$gz = \frac{az + b}{cz + d}.$$

Remark: Note that it would not make sense to speak of a function of *odd* weight, since $cz + d$ is only determined up to ± 1.

Proposition 8.21 *The meromorphic function $f(z)$ on $\mathcal{H}$ is weakly modular of weight $2k$ if and only if*

$$f(Tz) = f(z), \quad f(Sz) = z^{-2k} f(z).$$

Proof ▶ If $f(z)$ is weakly modular then the condition is certainly satisfied by $S, T \in G$.

Conversely, suppose the condition is satisfied S, T and g, where

$$gz = \frac{az+b}{cz+d}.$$

We shall show that it is satisfied by Sg and Tg.

We have

$$S(gz) = -\frac{1}{gz} = -\frac{cz+d}{az+b};$$

while

$$\begin{aligned} f(Sgz) &= (gz)^{-2k} f(gz) \\ &= (az+b)^{-2k}(cz+d)^{2k} f(gz) \\ &= (az+b)^{-2k}(cz+d)^{2k}(cz+d)^{-2k} f(z) \\ &= (az+b)^{-2k} f(z), \end{aligned}$$

so the result holds for Sg.

More simply,

$$T(gz) = gz + 1 = \frac{(a+c)z + (b+d)}{cz+d};$$

while

$$\begin{aligned} f(Tgz) &= f(gz) \\ &= (cz+d)^{-2k} f(z), \end{aligned}$$

so the result also holds for Tg.

It follows that the result holds where g is any word in S, T, ie for any $g \in \langle S, T \rangle$. Therefore, since S, T generate G, the result holds for all $g \in G$. ◀

Suppose $f(z)$ is a weakly modular function. Then in particular $f(z)$ is *periodic* with period 1:

$$f(z+1) = f(z).$$

The map

$$\Theta : z \mapsto q = e^{2\pi i z}$$

maps $\mathcal{H}$ onto the interior of the disk

$$\mathcal{D} = \{z : |z| < 1\}$$

with the point 0 removed. Moreover

$$\Theta(z_1) = \Theta(z_2) \Longleftrightarrow z_2 - z_1 \in \mathbb{Z}.$$

It follows that $f(z)$ defines a meromorphic function $g(q)$ on $\mathcal{D} \setminus \{0\}$:

$$f(z) = g(e^{2\pi z}).$$

Definition 8.12 *The weakly modular function $f(z)$ is said to have a pole (or zero) of order m at ∞ if that is true of $g(q)$ at $q = 0$. It is said to be* regular *at ∞ if it does not have a pole there; and in that case we set $f(\infty) = g(0)$.*

Definition 8.13 *The weakly modular function $f(z)$ is said to be* modular *if it has at worst a pole of finite order at ∞.*

It follows that a modular function has an 'expansion at ∞'

$$g(q) = \sum_{n \in \mathbb{Z}} a_n q^n,$$

where only finite number of the coefficients a_n with $n < 0$ are $\neq 0$.

Definition 8.14 *A modular function is said to be a* modular form *if it has no poles in $\mathcal{H}$, or at ∞.*

To each modular function $f(z)$ of weight $2k$ we can associate the lattice function $F(\Lambda)$ of weight $2k$ given by

$$F(\langle \lambda, \mu \rangle) = \mu^{-2k} f(\lambda/mu).$$

Conversely, we can recover the modular function from the lattice function by

$$f(z) = F(\langle 1, z \rangle).$$

Definition 8.15 *We define the functions $G_k(z)$ for $k \geq 2$ by*

$$G_k(z) = \sum_{(m,n) \neq (0,0)} \frac{1}{(m+nz)^{2k}}.$$

Thus $G_k(z)$ corresponds to the lattice function

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

Recall that Riemann's zeta function $f(s)$ is defined by

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots.$$

In number theory (in particular in the proof of the Prime Number Theorem), $\zeta(s)$ is considered as a function of a complex variable. But our concern is only with $\zeta(n)$ for integers $n \geq 2$.

Proposition 8.22 *$G_k(z)$ is a modular form of weight $2k$, with*

$$G_k(\infty) = 2\zeta(2k).$$

Proof ► The series for $G_k(z)$ is uniformly absolutely convergent in $\Im(z) \geq \delta$ for any $\delta > 0$, by comparison with the corresponding integral, as in the proof of Proposition 8.8. It follows that $G_k(z)$ is holomorphic in $\mathcal{H}$.

On the other hand, $G_k(z)$ is weakly modular of weight $2k$ from the same property of the associated lattice function $g_2(\Lambda)$.

It remains to see how $G_k(z)$ behaves near ∞. As $z \to \infty$,

$$\frac{1}{(m+nz)^{2k}} \to \begin{cases} 0 & \text{if } n \neq 0 \\ m^{-2k} & \text{if } n = 0. \end{cases}$$

Since the series is uniformly convergent, it follows that

$$G_k(z) \to 2\zeta(2k) \text{ as } z \to \infty.$$

It follows from this that $g(q)$ is regular at $q = 0$, with $g(0) = 2\zeta(2k)$. (For the coefficient a_{-n} in the Laurent series is given by

$$a_{-n} = \frac{1}{2\pi i}\int_C q^{n-1} g(q)\,dq$$

round a small circle C with centre 0, and this vanishes as the radius of the circle tends to 0.) ◄

Proposition 8.23 *A modular function has only a finite number of poles and zeros in $\mathcal{F}$.*

Proof ► The function $g(q)$ has an expansion

$$g(q) = q^n(a_n + a_{n+1}z + \cdots) \qquad (a_n \neq 0)$$

in some neighbourhood of 0. It follows that $g(q)$ is regular and has no zeros in some disk

$$0 < |q| \leq r \leq 1.$$

Hence $f(z)$ has no poles or zeros in the half-plane

$$\{z \in \mathcal{H} : \Im(z) > e^r\}.$$

On the other hand, $f(z)$ has only a finite number of poles or zeros in the compact set

$$\{z \in \bar{\mathcal{F}} : \Im(z) \leq e^r\}.$$

It follows that $f(z)$ has only a finite number of poles or zeros in $\mathcal{F}$. ◄

Definition 8.16 *Suppose $f(z)$ is a meromorphic function on U. For each $u \in U$ we set*

$$v_u(f) = \begin{cases} m & \text{if } f(z) \text{ has a zero of order } m \text{ at } u \\ -m & \text{if } f(z) \text{ has a pole of order } m \text{ at } u \\ 0 & \text{otherwise} \end{cases}$$

Remark: If $f(z), g(z)$ are meromorphic functions on U then

1. $v_u(f+g) \leq \max(v_u(f), v_u(g))$,
2. $v_u(fg) = v_u(f) + v_u(g)$.

Thus $v_u(f)$ is a *valuation* on the field of meromorphic functions on U; in fact, it satisfies the same stronger conditions as the p-adic valuation we considered in Chapter 5.

8.11 The Modular Counting Theorem

Theorem 8.5 *Suppose $f(z)$ is a modular function of weight $2k$. Then*

$$\frac{1}{3}v_\omega(f) + \frac{1}{2}v_i(f) + \sum_{z\neq\omega,-\omega^2,i} v_z(f) = \frac{k}{6}.$$

Proof ► Let

$$I = \frac{1}{2\pi i}\int_\Gamma \frac{f'(z)}{f(z)}dz,$$

where Γ runs round the boundary of $\mathcal{F}$, truncated at the top. More precisely,

$$\Gamma = A + B + C + D + E,$$

where A is the line joining $-\omega^2$ to $1/2+Ri$, B is the line joining $1/2+Ri$ to $-1/2+Ri$, C is the line joining $-1/2+Ri$ to ω, D is the circular arc joining ω to i, and E is the circular arc joining $-\omega^2$.

Let us assume for the moment that $f(z)$ has no poles or zeros on Γ, and also that R is so large that all the poles or zeros of $f(z)$ inside $\mathcal{F}$ are inside Γ.

As we know, if $f(z)$ has a pole or zero at $u \in \mathcal{H}$ then $f'(z)/f(z)$ has a simple zero at u with residue $v_u(f)$. It follows that

$$I = \sum_{u\in\mathcal{F}} v_u(f).$$

We consider the contributions to I from the five parts of the contour.

1. Since $f(z+1) = f(z)$, while the integrals are in opposite directions,

$$\int_A + \int_C = 0,$$

where for simplicity we write

$$\int_X \text{ for } \frac{1}{2\pi i}\int_X \frac{f'(z)}{f(z)}dz.$$

2. As z moves from ω to i on D, Sz moves from $-\omega^2$ to i on E. If $f(z)$ were of weight 0, so that $f(Sz) = f(z)$, then the contributions from D and E would cancel out in the same was as those from A and C. However, if $f(z)$ is of weight $2k$,

$$\begin{aligned} f(Sz) = \frac{1}{z^{2k}}f(z) &\Longrightarrow f'(Sz) = -\frac{2k}{z^{2k+1}}f(z) + \frac{1}{z^{2k}}f'(z) \\ &\Longrightarrow \frac{f'(Sz)}{f(Sz)} = -\frac{2k}{z} + \frac{f'(z)}{f(z)}. \end{aligned}$$

(In effect, $f'(z)/f(z) = d/dz(\log f(z))$.) Thus the main parts of the integral cancel out, leaving

$$\begin{aligned} \int_D + \int_E &= \frac{1}{2\pi i}\int_D \frac{2k}{z}dz \\ &= \frac{2k}{2\pi i}\int_{2\pi/3}^{\pi/2} i\theta d\theta \\ &= k\left(\frac{2}{3} - \frac{1}{2}\right) \\ &= \frac{k}{6} \end{aligned}$$

3. Finally, on B we have

$$f(z) = g(e^{2\pi i z}).$$

Changing variable from z to $q = e^{2\pi i z}$,

$$\frac{f'(z)}{f(z)} = 2\pi i q\frac{g'(q)}{g(q)}, \qquad dz = 2\pi i q\, dq,$$

and so

$$\int_B = \frac{1}{2\pi i}\int_\gamma \frac{g'(q)}{g(q)}dq,$$

where q runs round the small circle

$$\gamma : q = e^{-2\pi R}e^{2\pi x}$$

from $x = \pi$ to $x = -\pi$ in a negative (clockwise) direction.

Now if $g(q)$ has a pole or zero at $q = 0$ then $g'(q)/g(q)$ has a simple pole there with residue

$$v_0(g) = v_\infty(f);$$

while $g'(q)/g(q)$ is regular at $q = 0$ if $g(q)$ has neither pole nor zero there. It follows in all cases that

$$\int_B = -v_\infty(f).$$

Putting the parts together,

$$I = \frac{k}{6} - v_\infty(f).$$

But as we observed,

$$I = \sum_{u \in \mathcal{F}} v_u(f).$$

Thus

$$\sum_{u \in \mathcal{F}} v_u(f) + v_\infty(f) = \frac{k}{6},$$

as required.

It remains to deal with the case where $f(z)$ has one or more poles or zeros on Γ.

1. Suppose $f(z)$ has a pole or zero at $z = z_0 \in A$, where $z_0 \neq -\omega^2$. Then it will also have a pole or zero of the same order at the corresponding point on C, since $f(z-1) = f(z)$.

   Let us make small semi-circular diversions to the west of the pole or zero on both A and C. Then

   $$\int_{A'} + \int_{C'} = 0,$$

   as before; and the pole or zero is included once inside Γ', as required.

2. Suppose $f(z)$ has a pole or zero at $z = z_0 \in B$, where $z_0 \neq -\omega^2$ or i. Then $f(z)$ has a pole or zero of the same order at $Sz_0 \in C$, since

   $$f(Sz) = z^{2k} f(z).$$

   Let us make a small (almost) semicircular diversion δ to the south of z_0. Then $S\delta$ is a similar diversion to the north of Sz_0. It follows from our argument in the main case that

   $$\begin{aligned} \int_{B'} + \int_{C'} &= \frac{1}{2\pi i} \int B' \frac{2k}{z} dz \\ &= \frac{1}{2\pi i} \int B \frac{2k}{z} dz \\ &= \frac{k}{6}, \end{aligned}$$

   since the function $1/z$ is regular at z_0.

3. Now suppose that $f(z)$ has a pole or zero at $-\omega^2$, and so also at $\omega = -\omega^2 - 1$. We make a small diversion around both points, travelling inside FF along circular arcs δ, δ_1 of radius ϵ, so that

   $$B' = B'' + \delta, \quad C' = C'' + \delta_1,$$

   where B'', C'' are slightly curtailed versions of B, C. By our argument in the main case,

   $$\int_{B''} + \int_{C''} = \int_B + \int_C + O(\epsilon) \qquad = \frac{k}{6} + O(\epsilon).$$

   In the neighbourhood of ω,

   $$\frac{f'(z)}{f(z)} = \frac{v_\omega(f)}{z - \omega} + h(z),$$

   where $h(z)$ is holomorphic. The angle between C and D is $\pi/3$, so the arc δ has angle $\pi/3 + O(\epsilon)$, traversed in the negative direction. Hence

   $$\frac{1}{2\pi i} \int_\gamma \frac{f'(z)}{f(z)} dz = -\frac{v_\omega(f)}{6}.$$

   Similarly

   $$\frac{1}{2\pi i} \int_{\gamma_1} \frac{f'(z)}{f(z)} dz = -\frac{v_\omega(f)}{6}.$$

   Also

   $$\int_{A'} + \int_{C'} = 0,$$

   as before. Putting the parts together,

   $$I = \frac{k}{6} - v_\infty(f) - \frac{1}{2} v_\omega(f),$$

   and so

   $$\frac{1}{3} v_\omega(f) + \sum_u v_u(f) = \frac{k}{6},$$

   as required.

4. A pole or zero at i is dealt with similarly, by a small (nearly) semicircular diversion δ of radius ϵ to the north of i. Let D'', E'' denote the curtailed portions of D, E, so that

   $$\Gamma' = A + B + C + D'' + \delta + E''.$$

   Then

   $$\int_{D''} + \int E'' = \frac{k}{6} + O(\epsilon),$$

as in the previous case; while

$$\frac{f'(z)}{f(z)} = \frac{v_i(f)}{z-i} + h(z)$$

in the neighbourhood of i, where $h(z)$ is regular at i, and so

$$\int_\gamma = -\frac{v_i(f)}{2},$$

again as in the previous case. Putting the parts together,

$$\frac{1}{2}v_i(f) + \sum_u v_u(f) = \frac{k}{6},$$

as required.

5. Finally, if $f(z)$ has more than one pole or zero on Γ, eg a pole at $-\omega^2$ (and so also at ω) and a zero at i. Then we make a diversion around each pole or zero, according to the prescription above; and the parts will combine to give the result:

$$\frac{1}{3}v_\omega(f) + \frac{1}{2}v_i(f) + \sum_u v_u(f) = \frac{k}{6}.$$

◄

Proposition 8.24 *There are no modular forms of weight < 0; and the only modular forms of weigth 0 are the constants.*

Proof ► For a modular form $f(z)$, $v_u(f) \geq 0$ for all u. Thus if $f(z)$ were of weight < 0, then the left-hand side of the identity in the Theorem would be ≥ 0, while the right-hand side would be < 0.

Similarly, if $k = 0$ then the only way the identity could be satisfied is if $v_u(f) = 0$ for all u (including ∞). But then $f(z) - f(\infty)$ is a modular form of weight 0 with $v_\infty(f) > 0$, which is a contradiction unless the function is identically zero, ie $f(z) = f(\infty)$ is constant. ◄

Proposition 8.25 *There are no modular forms of weight 2.*

Proof ► Suppose $f(z)$ is such a form. Writing

$$a = v_\omega(f),\ b = v_i(f),\ c = \sum_{u \neq \omega, i} v_u(f),$$

we have

$$\frac{a}{3} + \frac{b}{2} + c = \frac{1}{6},$$

with $a, b \in \mathbb{N}$, which is manifestly impossible. ◄

Proposition 8.26 *The only modular form of weight 4 is $G_2(z)$, up to a scalar multiple.*

Proof ► The only solution of

$$\frac{a}{3} + \frac{b}{2} + c = \frac{1}{3}$$

with $a, b, c \in \mathbb{N}$, is $a = 1,\ b = c = 0$. Thus every modular form $f(z)$ of weight 4 must have a simple zero at ω.

But then

$$f(z) - \frac{f(\infty)}{G_2(\infty)}G_2(z),$$

if non-zero, is a modular form of weight 4 with $v_\infty(f) \geq 1$, which conflicts with our formula. Hence this form vanishes identically, ie

$$f(z) = \rho G_2(z),$$

where $\rho = f(\infty)/G_2(\infty)$. (Recall that $G_2(infty) = \zeta(4) \neq 0$.) ◄

Proposition 8.27 *The only modular form of weight 6 is $G_3(z)$, up to a scalar multiple.*

Proof ► The only solution of

$$\frac{a}{3} + \frac{b}{2} + c = \frac{1}{2}$$

with $a, b, c \in \mathbb{N}$, is evidently $a = 0,\ b = 1,\ c = 0$. Thus every modular form $f(z)$ of weight 6 must have a simple zero at i.

It follows as in the proof of the last Proposition that

$$f(z) = \rho G_3(z),$$

where $\rho = f(\infty)/G_3(\infty)$. ◄

We have proved incidentally the following result.

Proposition 8.28 $G_2(\omega) = 0,\ G_3(i) = 0.$

It is easy enough to prove this directly; since $S\omega = -\omega^2$,

$$G(-\omega^2) = \frac{1}{\omega^4} G(\omega) = \omega^2 G(\omega),$$

while since $-\omega^2 = \omega + 1$,

$$G(-\omega^2) = G(\omega),$$

Similarly, since $Si = i$,

$$G_3(i) = \frac{1}{i^6} G_3(i) = -G_3(i).$$

Recall that the discriminant $\Delta(\mathcal{E})$ of the elliptic curve

$$y^2 = x^3 + bx + c$$

was defined to be

$$\Delta = 2^4 D,$$

where

$$D = -(4b^3 + 27c^2)$$

is the discriminant of the polynomial on the right. (The factor 2^4 was introduced to allow the discriminant of the general Weierstrassian elliptic curve

$$y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

to be defined as a polynomial in c_1, c_2, c_3, c_4, c_6 with integral coefficients.)

It follows that the discriminant of the elliptic curve

$$\mathcal{E}(\mathbb{C}) : y^2 = x^3 - 15g_2x - 35g_3$$

is

$$\Delta(\mathcal{E}) = 2^4 3^3 5^2 (20g_2^3 - 49g_3^2).$$

(The scalar factor is irrelevant for our present purposes, and is only retained for consistency.)

Definition 8.17 *The modular invariant $\Delta(z)$ is defined by*

$$\Delta(z) = 2^4 3^3 5^2 (20G_2^3 - 49G_3^2).$$

Proposition 8.29 *$\Delta(z)$ is a modular form of weight 12. It has a simple zero at ∞, and no other poles or zeros.*

Proof ► It is clear that $\Delta(z)$ is a modular form of weight 12. We know that the elliptic curve

$$\mathcal{E} : y^2 = x^3 - 15g_2x - 35g_3$$

is non-singular. (Recall the argument: If the curve had a singularity, it would be a point $(\alpha, 0)$ on the line of symmetry $y = 0$, where α is a double root of the polynomial on the right. But we have seen that this polynomial has three distinct roots corresponding to the semilattice points of the lattice Λ in question.)

But now our formula gives

$$v_\infty(\Delta) = 1,$$

ie $\Delta(z)$ has a simple zero at ∞. ◄

Remark: A modular form $f(z)$ with $f(\infty) = 0$ is called a *cusp form*.

Proposition 8.30 *The modular forms are generated by $G_2(z)$ and $G_3(z)$. More precisely, a modular form of weight $2k$ is a linear combination of the modular forms*

$$G_2(z)^a G_3(z)^b,$$

where

$$2a + 3b = k.$$

Proof ► We argue by induction on k. We have seen that the result is true for $k = 0, 2, 4, 6$.

Lemma *The only modular form of weight 8 is $G_2(z)^2$, up to a scalar multiple. The only modular form of weight 10 is $G_2(z)G_3(z)$, up to a scalar multiple.*

Proof of Lemma ▷ The only solution of

$$2a + 3b = 4$$

is $a = 2, b = 0$, while the only solution of

$$2a + 3b = 5$$

is $a = 1, b = 1$. The result follows as in Propositions 8.26 and 8.27. ◁

Lemma *The equation*

$$2a + 3b = k \qquad (a, b \in \mathbb{N})$$

has a solution for all $k \geq 2$.

Proof of Lemma ▷ If k is even, $a = k/2,\ b = 0$ is a solution; while if k is odd, $a = (k-3)/2,\ b = 1$ is a solution. ◁

Now suppose $f(z)$ is a modular form of weight $2k$, where $k \geq 6$. By the last Lemma, we can find a, b such that $2a + 3b = k$. Let

$$g(z) = G_2(z)^a G_3(z)^b.$$

Then $g(z)$ is also of weight $2k$; and so is

$$h(z) = f(z) - \rho g(z),$$

where we choose

$$\rho = \frac{f(\infty)}{G_2(\infty)^a G_3(\infty)^b} = \frac{f(\infty)}{\zeta(4)^a \zeta(6)^b}$$

so that

$$h(\infty) = 0.$$

Then $h(z)$ is a modular form of weight $2k$ with $h(\infty) = 0$.

But now

$$k(z) = \frac{h(z)}{\Delta(z)}$$

is a modular form of weight $2k - 12$; for the zero of $h(z)$ at ∞ cancels out the zero of $\Delta(z)$ at ∞, and $\Delta(z)$ has no other zeros.

It follows by our inductive hypothesis that $k(z)$ is a linear combination of the monomial functions

$$G_2(z)^{a'} G_3(z)^{b'} \qquad (2a' + 3b' = k - 6).$$

Hence

$$g(z) = \Delta(z) k(z)$$

is a linear combination of the functions

$$G_2(z)^{a''} G_3(z)^{b''} \qquad (2a'' + 3b'' = k);$$

and so therefore is

$$f(z) = g(z) + G_2(z)^a G_3(z)^b.$$

◀

Proposition 8.31 *The functions $G_2(z)^a G_3(z)^b$ with $2a+3b = k$ form a basis for the modular forms of weight $2k$.*

Proof ▶ Suppose there were a linear relation between these monomial functions. The relation of lowest weight must be of the form

$$\lambda G_2(z)^{3c} + \cdots + \mu G_3(z)^{2c} = 0.$$

(For otherwise we could divide the relation by $G_2(z)$ or $G_3(z)$.)

But now taking $z = i, -\omega^2$,

$$\mu G_3(\omega)^{2c} = 0 \Longrightarrow \mu = 0, \lambda G_2(\omega)^{3c} = 0 \Longrightarrow \lambda = 0.$$

◀

The modular forms constitute a *graded algebra*

$$\mathcal{M} = (\mathcal{M}_k)_{k \in \mathbb{N}},$$

where $\mathcal{M}_k$ is the space of modular forms of weight $2k$. It follows from the Proposition above that this algebra is the polynomial algebra generated by G_2 and G_3:

$$\mathcal{M} = \mathbb{C}[G_2, G_3].$$

8.12 The j-invariant

Definition 8.18 *We set*

$$j(z) = 2^6 3^3 \frac{G_2(z)^3}{\Delta(z)}.$$

Remark: The scalar factor is of no significance for our present purpose. (It is chosen so that $j(z)$ has residue 1 at ∞.)

Proposition 8.32 *$j(z)$ is a modular function of weight 0. It has a simple pole at ∞ and a triple zero at ω mod G, and no other poles or zeros.*

Proof ▶ This follows at once from the properties of $G_2(z)$ and $\Delta(z)$ (Propositions 8.26 and 8.29). ◀

Corollary 12 *For each $c \in \mathbb{C}$ there is just one $z \in \mathcal{F}$ such that*

$$j(z) = c.$$

Proof ▶ The modular function $j(z) - c$ is of weight 0, and has a simple pole at ∞. It follows from the Modular Counting Theorem that $f(z)$ either has a triple zero at $-\omega^2$, or else a simple zero at some other point.

In any case, there is just one zero in $\mathcal{F}$. ◀

Recall that each modular function has an associated lattice function.

Definition 8.19 *For each lattice $\Lambda = \langle \lambda, \mu \rangle$ we set*

$$J(\Lambda) = j(\lambda/\mu).$$

Thus

$$\begin{aligned} J(\Lambda) &= 2^6 3^3 \frac{g_2^3}{\Delta} \\ &= \frac{2^2}{5^2} \frac{g_2^3}{20g_2^3 - 49g_3^2}. \end{aligned}$$

Theorem 8.6 *Each elliptic curve*

$$\mathcal{E}(\mathbb{C}) : y^2 = x^3 + bx + c$$

arises from a unique lattice Λ.

Proof ► We are looking for a lattice Λ with

$$-15g_2(\Lambda) = b, \; -35g_3(\Lambda) = c.$$

For such a lattice

$$j(\Lambda) = 2^2 3^3 5^3 \frac{b^3}{2^2 b^3 - 3^3 c^2} = C,$$

say.

By the Corollary to Proposition 8.32 there is a unique $z_0 \in \mathcal{F}$ such that

$$j(z_0) = C.$$

Let

$$\Lambda_0 = \langle 1, z_0 \rangle;$$

and let

$$\mathcal{E}(\mathbb{C}) : y^2 = x^3 + b_0 x + c_0$$

be the elliptic curve associated to Λ_0. Then

$$\frac{b_0^3}{4b_0^3 - 27c_0^2} = j(z_0) = C = \frac{b^3}{4b^3 - 27c^2}.$$

We know that the denominators do not vanish, since the curves are non-singular. Hence

$$b_0 = 0 \iff b = 0.$$

Suppose for the moment this is not so. Then

$$\frac{4b_0^3 - 27c_0^2}{b_0^3} = \frac{4b^3 - 27c^2}{b^3} \implies \frac{c_0^2}{b_0^3} = \frac{c^2}{b^3}.$$

Evidently

$$c_0 = 0 \iff c = 0.$$

Suppose this too is not so. Then

$$\left(\frac{b}{b_0}\right)^3 = \left(\frac{c}{c_0}\right)^2.$$

Let

$$\frac{b}{b_0} = \beta, \; \frac{c}{c_0} = \gamma, \; \rho = \frac{\gamma}{\beta}.$$

Then $\gamma^2 = \beta^3$, and so

$$\begin{aligned} \rho^2 &= \frac{\gamma^2}{\beta^2} = \frac{\beta^3}{\beta^2} = \beta, \\ \rho^3 &= \frac{\gamma^3}{\beta^3} = \frac{\gamma^3}{\gamma^2} = \gamma. \end{aligned}$$

Thus

$$b = \rho^2 b_0, \; c = \rho^3 c_0.$$

Let $s^2 = \rho$. Then

$$b = s^4 b_0, \; c = s^6 c_0.$$

It follows that the given curve is defined by the lattice

$$\Lambda = s\Lambda_0 = \langle s, sz_0 \rangle.$$

If $b_0 = b = 0$ then the two curves are

$$y^2 = x^3 + c_0, \quad y^2 = x^3 + c.$$

The transformation

$$x \to s^2 x, \; y \to s^3 y$$

will take the first curve into the second provided we choose s so that

$$c = s^6 c_0.$$

Similarly, if $c_0 = c = 0$ then the curves are

$$y^2 = x^3 + b_0 x, \quad y^2 = x^3 + bx,$$

and the transformation will take the first curve into the second provided we choose s so that

$$b = s^4 4_0.$$

Now suppose the given curve $\mathcal{E}$ is also defined by the lattice

$$\Lambda' = \langle \lambda', \mu' \rangle = \mu' \langle 1, \lambda'/\mu' \rangle = \mu \langle 1, z' \rangle,$$

where $z' = \lambda'/\mu'$. Then

$$j(z') = J(\Lambda') = 2^2 3^3 5^3 \frac{b^3}{2^2 b^3 - 3^3 c^2} = J(\Lambda) = j(z_0).$$

Hence, by Proposition 8.32,

$$z = gz$$

for some $g \in G$, say the transformation

$$gz = \frac{az + b}{cz + d}.$$

It follows that the lattices

$$\langle 1, z \rangle, \ \langle 1, z_0 \rangle$$

are the same, and so the lattices Λ', Λ are similar, say

$$\Lambda' = s\Lambda.$$

But then

$$s^b b = b, \ s^6 c = c.$$

If $b, c \neq 0$ this implies that $s^2 = 1$, so that $s = \pm 1$ and the lattices are the same.

If $b = 0$ then

$$s = \pm 1, \ \pm\omega, \ \pm\omega^2.$$

But

$$j(z) = 0 \Longrightarrow z = g\omega.$$

Thus the lattice is similar to

$$\Lambda_0 = \{m + n\omega : m, n \in \mathbb{Z}\},$$

and it is readily verified that

$$\omega\Lambda_0 = \Lambda_0, \quad -\omega^2\Lambda_0 = \Lambda_0,$$

so again the lattice is unique.

Similarly if $c = 0$ then

$$G_3(z_0) = 0 \Longrightarrow z_0 = i,$$

so that the lattice is similar to

$$\Lambda_0 = \{m + ni : m, n \in \mathbb{Z}\},$$

which again is invariant under the transformations given by $s = \pm i$. ◄

Remark: We call $j(z_0) = J(\Lambda)$ the j-*invariant* of the corresponding elliptic curve

$$y^2 = x^3 - 15g_2x - 35g_x.$$

We can extend the definition to all Weierstrassian curves

$$\mathcal{E}(k) : y^2 + c_1xy + c_3y = x^3 + c_2x^2 + c_4x + c_6$$

over all fields k, by exactly the same method by which we extended the definition of the discriminant $\Delta(\mathcal{E})$ to all such curves.

The j-invariant turns out to have an important rôle in the classification of elliptic curves over a general field k. But that is another story.

Chapter 9

Mordell's Theorem

9.1 The Theorem

Our aim in this Chapter is to prove Mordell's Theorem, the central result on the arithmetic of elliptic curves.

Theorem 9.1 (Mordell) *The abelian group on the elliptic curve*

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx + c.$$

is finitely-generated.

In fact we shall find that we have to consider the more general case of an elliptic curve over a *number field* K (rather than $\mathbb{Q}$). This is because our proof requires that the polynomial

$$f(x) \equiv x^3 + ax^2 + bx + c \equiv (x - \alpha)(x - \beta)(x - \gamma)$$

should factorise completely in K, ie that $\alpha, \beta, \gamma \in K$.

If $f(x)$ already factorises in $\mathbb{Q}$ then there is no need to introduce number fields. It is interesting to observe that this is the case with Wiles' proof of Fermat's Last Theorem, which (as we have noted) associates to the solution

$$A^n + B^n = C^n$$

of Fermat's Last Theorem the elliptic curve

$$y^2 = x(x - A^n)(x + B^n)$$

with discriminant

$$\Delta = (ABC)^{2n} :$$

the point being that the discriminant has — in relation to its size — a large number of small factors, which (Wiles shows) leads to a contradiction.

9.2 The Idea of the Proof

Suppose $\mathcal{E}(\mathbb{Q})$ *is* finitely-generated. Then the group

$$\frac{\mathcal{E}(\mathbb{Q})}{2\mathcal{E}(\mathbb{Q})}$$

is finitely-generated, and so is finite. More precisely, we have the following result.

Proposition 9.1 *Suppose A is a finitely-generated abelian group, say*

$$A = F \oplus r\mathbb{Z} = F \oplus \underbrace{\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{r \text{ summands}}$$

where F is finite and $r = \operatorname{rank}(A)$. Suppose there are 2^s elements of order dividing 2 in A. Then

$$A/2A = (r + s)(\mathbb{Z}/(2)) = \underbrace{\mathbb{Z}/(2) \oplus \cdots \oplus \mathbb{Z}/(2)}_{r+s \text{ summands}}.$$

Proof ► It is a readily verified that

$$B = C \oplus D \Longrightarrow B/2B = C/2C \oplus D/2D.$$

It follows that

$$A/2A = F/2F \oplus \mathbb{Z}/(2) \oplus \cdots \oplus \mathbb{Z}/(2),$$

each direct summand $\mathbb{Z}$ in A contributing one copy of $\mathbb{Z}/(2)$. It remains to determine $F/2F$.

Consider the homomorphism

$$\phi : F \to F : x \mapsto 2x.$$

By the First Isomorphism Theorem,

$$\frac{F}{2F} = \frac{F}{\operatorname{im}\phi} \cong \ker\phi = \{x \in F : 2x = 0\}.$$

Thus 2^s is just the number of elements in F of order dividing 2. Since every element in A of finite order is in F, the result follows. ◄

It follows from this Proposition that

$$\mathcal{E}(\mathbb{Q}) \text{ finitely-generated} \Longrightarrow \mathcal{E}(\mathbb{Q})/2\mathcal{E}(\mathbb{Q}) \text{ finite};$$

and moreover,

$$\|\mathcal{E}(\mathbb{Q})/2\mathcal{E}(\mathbb{Q})\| = 2^{r+s}.$$

where r is the rank of $\mathcal{E}(\mathbb{Q})$ and $s = 0, 1$ or 2 according as the cubic $f(x)$ has 0,1 or 3 roots in $\mathbb{Q}$.

The converse, unfortunately, is not true: an abelian group A may have $A/2A$ finite without A being finitely-generated. For example,

$$\mathbb{Q}/2\mathbb{Q} = 0,$$

since every rational is expressible as twice another rational; but $\mathbb{Q}$ is not finitely-generated as an abelian group.

So the condition (that $\mathcal{E}/2\mathcal{E}$ be finite) is necessary but not sufficient. However, it allows us to start a process of "infinite descent", as follows.

Let the points $E_1, \ldots, E_m$ be representatives of the cosets in $\mathcal{E}/2\mathcal{E}$; and suppose $P \in \mathcal{E}$. Then

$$P - E_i \in 2\mathcal{E}$$

for some i, say

$$P - E_{i_0} = 2P_1.$$

We can apply the same argument to P_1:

$$P_1 - E_{i_1} = 2P_2;$$

and we can continue in this way

$$\begin{aligned} P_2 - E_{i_2} &= 2P_3, \\ P_3 - E_{i_3} &= 2P_4. \\ &\ldots \end{aligned}$$

We expect the points $P_1, P_2, \ldots$ defined in this way by successive 'halving' to *descend the curve* in some sense. But what exactly do we mean by 'descend'? When infinite descent is applied to *integral* solutions of an equation, the meaning is clear: the coordinates become smaller. But we are dealing with *rational* points. We need some notion of the simplicity of a rational number $q = m/n$. We therefore define the *height* of $q \in \mathbb{Q}$ to be

$$H(q) = \max(|m|, |n|),$$

if $q = m/n$ in its lowest terms. Now our task is clear; we have to show that the points $P_1, P_2, \ldots$ are descending in the sense that the heights of their coordinates are decreasing.

Actually, we shall find it sufficient, and much simpler, to consider the x-coordinate.

Thus the proof has 2 quite separate parts, which we might call the *algebraic* or group-theoretic part, and the *topological* or valuation-theoretic part.

9.3 When can a Point be 'Halved'?

Recall our mammoth formula for the 'double' of a point $X, Y) \in \mathcal{E}$:

$$\begin{aligned} 2(X,Y) = \Bigg(&\frac{X^4 - 2bX^2 - 8cX + b^2 - 4ac}{4Y^2}, \\ &\frac{X^6 + 2aX^5 + 5bX^4 + 20cX^3 + (8a^2c - 2ab^2 - 4bc)X + b^3 - 4abc + 8c^2}{8Y^3} \Bigg). \end{aligned}$$

If $c = 0$ we may observe that the x-coordinate is a perfect square

$$\left(\frac{X^2 - b}{2Y}\right)^2.$$

At first sight this seems a pure fluke. But it turns out to be the hinge of our argument.

Suppose the line

$$y = mx + d$$

meets $\mathcal{E}$ in the 3 points

$$P = (x_1, y_1),\ Q = (x_2, y_2),\ R = (x_3, y_3).$$

Then x_1, x_2, x_3 are the roots of the cubic

$$(mx + d)^2 = x^3 + ax^2 + bx + c.$$

It follows that

$$\begin{aligned} x_1 + x_2 + x_3 &= m^2 - a, \\ x_2x_3 + x_3x_1 + x_1x_2 &= b - 2md, \\ x_1x_2x_3 &= d^2 - c. \end{aligned}$$

The last of these equations is the one that concerns us now. Suppose again that $c = 0$. Then the equation becomes

$$x_1x_2x_3 = d^2.$$

This has a homomorphic air about it:

$$P + Q + R = 0 \Longrightarrow x_1x_2x_3 = d^2.$$

In particular we recover the 'fluke' above; if we take $Q = R$, so that $P + 2Q = 0$, then we see that

$$x_1x_2^2 = d^2.$$

Remembering that P and $-P$ have the same x-coordinate, it follows that if $P = 2Q$ then the x-coordinate of P is a square.

This was on the assumption that $c = 0$. Geometrically, this means that $(0,0) \in \mathcal{E}$. Now $(0,0)$ is a point of order 2. But any point $(\alpha, 0) \in \mathcal{E}$ of order 2 can be brought to $(0,0)$ by the coordinate-change $x \mapsto x - \alpha$.

Thus the only assumption we are making is that $\mathcal{E}(K)$ possesses a point of order 2. In fact, returning to the original coordinates, we can express the result as follows: Suppose $(\alpha, 0) \in \mathcal{E}$, where α is a root of

$$f(x) = x^3 + ax^2 + bx + c.$$

Then

$$P = (X, Y) \in 2\mathcal{E}(K) \Longrightarrow X - \alpha = \theta^2$$

in K.

But there is nothing special about the root α. Suppose now that all 3 roots $\alpha, \beta, \gamma \in K$. Then our argument shows that

$$P = (X, Y) \in 2\mathcal{E}(K) \Longrightarrow X - \alpha,\ X - \beta,\ X - \gamma \in K^2,$$

that is,

$$X - \alpha = \alpha'^2,\ X - \beta = \beta'^2,\ X - \gamma = \gamma'^2,$$

where $\alpha', \beta', \gamma' \in K$.

This brings us to the main result in the algebraic half of the proof of Mordell's Theorem.

Proposition 9.2 *Suppose*

$$\mathcal{E}(K) : y^2 = x^3 + ax^2 + bx + c$$

is an elliptic curve over the number field K; and suppose

$$f(x) = x^3 + ax^2 + bx + c$$

has roots $\alpha, \beta, \gamma \in K$. Then

$$P = (X, Y) \in 2\mathcal{E}(K) \Longleftrightarrow X - \alpha, X - \beta, X - \gamma \in K^2.$$

Remark: Note that any 2 of these conditions implies the third, since

$$Y^2 = (X - \alpha)(X - \beta)(X - \gamma).$$

Proof ▶ To simplify the presentation, let us make the coordinate-change $x \mapsto x - X$. (This is *not* the same as the earlier coordinate-change making $c = 0$.) The given point P is now $(0, Y)$, and we have to show that

$$P = (0, Y) \in 2\mathcal{E}(K) \Longleftrightarrow -\alpha, -\beta, -\gamma \in K^2,$$

say

$$\alpha = -\alpha'^2, \beta = -\beta'^2, \gamma = -\gamma'^2,$$

where $\alpha', \beta', \gamma' \in K$.

(We have already seen that this condition is necessary. Our argument will re-prove that, and show that the condition is also sufficient.)

By definition, $P = 2Q$ if the tangent to $\mathcal{E}$ at $-Q$ passes through P. Let us therefore determine all the tangents that can be drawn from P to $\mathcal{E}$.

The general line through $P = (0, Y)$ is

$$y = mx + Y.$$

This meets $\mathcal{E}$ where

$$(mx + Y)^2 = x^3 + ax^2 + bx + c.$$

We know that one root of this is $x = 0$ since $P = (0, Y) \in \mathcal{E}$. In other words,

$$Y^2 = c.$$

The other 2 roots of the equation satisfy

$$x^2 + (a - m^2)x + (b - 2mY) = 0.$$

The line will be a tangent if this quadratic has coincident roots. The condition for this is that

$$(a - m^2)^2 = 4(b - 2mY).$$

This is a quartic for m; so in general 4 tangents can be drawn to $\mathcal{E}$ from any point $P \in \mathcal{E}$.

It is easy to see why there are 4 tangents. Let

$$A = (\alpha, 0),\ B = (\beta, 0),\ C = (\gamma, 0)$$

be the 3 points of order 2 on $\mathcal{E}$. If $P = 2Q$ is one 'halving' of P then there are 3 others:

$$P = 2(Q + A),\ P = 2(Q + B),\ P = 2(Q + C).$$

These give rise to the 4 tangents passing through P. In particular we see that if one tangent is defined over K then so are all 4. (Note that the tangents must be distinct, since A, B, C are distinct.) Thus if our quartic has one root in K then all its roots must lie in K.

We should say, that 4 tangents can be drawn *over* $\mathbb{C}$. For there is no reason to suppose that the roots of the quartic will lie in K. In fact, that is exactly what we have to determine.

For if $Q \in \mathcal{E}(K)$ then our line PQ is defined over K, and so $m \in K$.

Conversely, if $m \in K$ and the line is tangent to $\mathcal{E}$ then the point $Q = (\xi, \eta)$ at which it touches has coordinates in K. For the roots of our equation

$$(mx + Y)^2 = x^3 + ax^2 + bx + c$$

are now $0, \xi, \xi$, so that

$$2\xi = m^2 - a \in K,$$

and then

$$\eta = m\xi + Y \in K.$$

Thus $P = 2Q$ if and only if there is a line through P touching $\mathcal{E}$, and defined over K. In other words, $P = 2Q$ if our quartic for m has a root in K.

Recall the classical technique for solving a quartic (or at least reducing it to a cubic): re-cast the quartic in the form

$$Q(x)^2 = L(x)^2,$$

where $Q(x)$ is quadratic and $L(x)$ is linear.

In our case this leads to the equation

$$(m^2 - a + \lambda)^2 = 2\lambda m^2 - 8mY + (\lambda^2 - 2\lambda a + 4b),$$

where we have to choose λ so that the quadratic form in m on the right is a perfect square.

The condition for this is that λ should satisfy the cubic

$$(4Y)^2 = 2\lambda(\lambda^2 - 2\lambda a + 4b).$$

Recalling that $Y^2 = c$, this simplifies to

$$\lambda^3 - 2a\lambda^2 + 4b\lambda - 8c = 0.$$

Miracle! This is almost our original cubic $f(x)$ (in the equation $y^2 = f(x)$). In fact the equation can be written

$$f(-\lambda/2) = 0.$$

It follows that its 3 solutions are

$$\lambda = -2\alpha,\ -2\beta,\ -2\gamma.$$

We can take λ to have any of these values. Suppose we take

$$\lambda = -2\alpha.$$

Then our quartic for m takes the form

$$(m^2 - a + \lambda)^2 = (2\lambda)(m - 2Y/\lambda)^2.$$

Thus if our quartic has a solution in K, which we know is the case if $P = 2Q$, then $\lambda/2 = -\alpha$ must be a square. Similarly, taking the other 2 values for λ, it follows that $-\beta$ and $-\gamma$ must also be squares:

$$-\alpha = \alpha'^2,\ -\beta = \beta'^2,\ -\gamma = \gamma'^2.$$

Conversely suppose that this is the case. Then we can take

$$\lambda = -2\alpha = 4\alpha'^2,$$

and our quartic for m splits into 2 quadratics

$$m^2 - a + \lambda = \pm 2\alpha'(m - 4Y/\lambda).$$

Note that since $\alpha + \beta + \gamma = -a$,

$$\begin{aligned} -a + \lambda &= -\alpha + \beta + \gamma \\ &= \alpha'^2 - \beta'^2 - \gamma'^2. \end{aligned}$$

Furthermore

$$Y^2 = c = -\alpha\beta\gamma = \alpha'^2\beta'^2\gamma'^2,$$

so that

$$Y = \pm\alpha'\beta'\gamma'.$$

We can take the + sign without loss of generality, since the signs of α', β', γ' were arbitrary anyway.

Thus our quadratics become

$$m^2 + \alpha'^2 - \beta'^2 - \gamma'^2 = \pm 2(\alpha' m - 2\beta'\gamma').$$

In other words,

$$(m \pm \alpha')^2 = (\beta' \pm \gamma')^2.$$

We conclude that the 4 tangents through P are $y = mx + Y$, where

$$m = \alpha' + \beta' - \gamma',\ \alpha' - \beta' + \gamma',\ -\alpha' + \beta' - \gamma',\ -\alpha' - \beta' + \gamma'.$$

In particular, we see that if $-\alpha, -\beta, -\gamma$ are perfect squares in K then $m \in K$ and $P = 2Q$. ◄

9.4 The 3 Homomorphisms

Recall that if

$$P = (x_1, y_1),\ Q = (x_2, y_2),\ R = (x_3, y_3) \in \mathcal{E}(K)$$

are 3 points of

$$\mathcal{E}(K) : y^2 = x^3 + ax^2 + bx$$

then

$$P + Q + R = 0 \Longrightarrow x_1x_2x_3 \in K^2.$$

It would have been nicer if we could have said

$$P + Q + R = 0 \Longrightarrow x_1x_2x_3 \in (K^\times)^2,$$

where $K^\times$ denotes the multiplicative group formed by the non-zero elements of K. For then we could say that

$$P + Q + R = 0 \Longrightarrow x_1x_2x_3 \equiv 1 \bmod (K^\times)^2.$$

which would suggest that we had a homomorphism

$$\Theta : \mathcal{E}(K) \to K^\times/(K^\times)^2.$$

Unfortunately, this breaks down if

$$x_1x_2x_3 = 0,$$

ie one of x_1, x_2, x_3 vanishes. This is the case if one or more of the points P, Q, R is equal to $D = (0, 0)$.

Remarkably, the homomorphism can be rescued in this case. Recall that

$$x_2x_3 + x_3x_1 + x_1x_2 = b - 2md.$$

In the case $x_1x_2x_3 = 0$ we have $m = 0$. Let us suppose $x_1 = 0$. Then

$$x_2x_3 = b$$

Thus if we agree to map D onto $b \bmod (K^\times)^2$ rather than 0 then we recover the homomorphic relation.

Proposition 9.3 *Suppose $\mathcal{E}(K)$ is the elliptic curve*

$$y^2 = x^3 + ax^2 + bx + c.$$

Then the map

$$\Theta : \mathcal{E}(K) \to K^\times/(K^\times)^2$$

defined by

$$P \mapsto \begin{cases} X & \text{if } P = (X, Y) \neq (0, 0), \\ b & \text{if } P = (0, 0), \\ 1 & \text{if } P = 0 = [0, 1, 0] \end{cases}$$

is a homomorphism

Proof ► If $P = (X, Y)$ then $-P = (X, -Y)$. Thus

$$P + Q = 0 \Longrightarrow \Theta(P)\Theta(Q) = 1$$

in all cases.

It is sufficient therefore to show that

$$P + Q + R = 0 \Longrightarrow \Theta(P)\Theta(Q)\Theta(R) = 1$$

in all cases. We know that this holds if none of P, Q, R is 0 or D. If one is 0 then the result reduces to the case $P + Q = 0$. It two of P, Q, R are D then the third is 0, so that case has been dealt with.

It only remains to consider the case where just one is D, say $P = D$, and $Q, R \neq 0$. But we have seen that in this case

$$x_2x_3 = b,$$

and so

$$\Theta(Q)\Theta(R) = b.$$

Thus, since $\Theta(D) = b$,

$$\Theta(P)\Theta(Q)\Theta(R) = b^2 = 1$$

in $K^\times/(K^\times)^2$. ◄

We were assuming in this Proposition that $c = 0$. To convert back to the general case, we note that if α is a root of $f(x)$ then the coordinate-change $x \mapsto x - \alpha$ takes $f(x)$ into $x^3 + a'x^2 + b'x$, where

$$a' = a + 3\alpha,\ b' = b + 3\alpha^2 + 2a\alpha.$$

Corollary 13 *Suppose $\mathcal{E}(K)$ is the elliptic curve*

$$y^2 = x^3 + ax^2 + bx + c;$$

and suppose

$$A = (\alpha, 0)$$

is a point of order 2 on $\mathcal{E}(K)$. *Then the map*

$$\Theta_\alpha : \mathcal{E}(K) \to K^\times/(K^\times)^2$$

defined by

$$P \mapsto \begin{cases} X - \alpha & \text{if } P = (X,Y) \neq A, \\ 3\alpha^2 + 2a\alpha + b & \text{if } P = A, \\ 1 & \text{if } P = 0 = [0,1,0], \end{cases}$$

is a homomorphism.

Note that we have 3 homomorphisms, corresponding to the 3 roots α, β, γ of $f(x)$. We can re-state Proposition 9.2 as follows.

Proposition 9.4 *Suppose*

$$\mathcal{E}(K) : y^2 = x^3 + ax^2 + bx + c$$

is an elliptic curve over the number field K*; and suppose*

$$f(x) = x^3 + ax^2 + bx + c$$

has roots $\alpha, \beta, \gamma \in K$. *Then*

$$2\mathcal{E}(K) = \ker\Theta_\alpha \cap \ker\Theta_\beta \cap \ker\Theta_\gamma.$$

Remark: As we noted earlier,

$$\ker\Theta_\alpha \subset \ker\Theta_\beta \cap \ker\Theta_\gamma,$$

and similarly for the other 2 kernels — each is contained in the intersection of the other two. Thus

$$2\mathcal{E}(K) = \ker\Theta_\alpha \cap \ker\Theta_\beta = \ker\Theta_\alpha \cap \ker\Theta_\gamma = \ker\Theta_\alpha \cap \ker\Theta_\beta.$$

Corollary 14 $\mathcal{E}/2\mathcal{E}$ *is finite if and only if* $\operatorname{im}\Theta_\alpha, \operatorname{im}\Theta_\beta, \operatorname{im}\Theta_\gamma$ *are all finite.*

Proof ► By the Proposition (and the following Remark),

$$\begin{aligned} 2\mathcal{E} &= \ker\Theta_\alpha \cap \ker\Theta_\beta \cap \ker\Theta_\gamma \\ &= \ker\Theta_\alpha \cap \ker\Theta_\beta \\ &= \ker\Theta_\alpha \cap \ker\Theta_\gamma \\ &= \ker\Theta_\beta \cap \ker\Theta_\gamma \end{aligned}$$

Lemma 8 *Suppose* B, C *are subgroups of the group* A. *Then*

$$\frac{A}{B \cap C}$$

is finite if and only if

$$\frac{A}{B}, \frac{A}{C},$$

are finite; and then

$$\|\frac{A}{B \cap C}\| \le \|\frac{A}{B}\| \, \|\frac{A}{C}\|.$$

Proof of Lemma ▷ We have

$$\|A/B \cap C\| = \|A/B\| \, \|B/B \cap C\|.$$

Let Φ be the canonical surjective homomorphism

$$\Phi : A \to A/C.$$

If Φ_B is the restriction of Φ to B, then

$$\ker\Phi_B = B \cap C.$$

It follows from the First Isomorphism Theorem that

$$B/B \cap C \cong \operatorname{im}\Phi_B \subset A/C.$$

Hence

$$\|B/B \cap C\| \le \|A/C\|,$$

and the result follows. ◁

Applying the Lemma with $B = \ker\Theta_\alpha$, $C = \ker\Theta_\beta$ we deduce that $\mathcal{E}/2E$ is finite if and only if

$$\mathcal{E}/\ker\Theta_\alpha \cong \operatorname{im}\Theta_\alpha \text{ and } \mathcal{E}/\ker\Theta_\beta \cong \operatorname{im}\Theta_\beta$$

are both finite; and the same is true if α, β are replaced by α, γ or β, γ. ◄

9.5 The Finiteness of the Images

We have to prove that $\operatorname{im}\Theta_\alpha$, $\operatorname{im}\Theta_\beta$, $\operatorname{im}\Theta_\gamma$ (or at least two of them) are finite. It is sufficient to prove the result for one of them; and we can again suppose for simplicity that $c = 0$.

Proposition 9.5 *Let $\mathcal{E}$ be the curve*

$$\mathcal{E}(K) : y^2 = x^3 + ax^2 + bx$$

where b, c are algebraic integers in K. Let Θ be the homomorphism

$$\mathcal{E} \to K^\times/(K^\times)^2 : \ P \mapsto \begin{cases} X & \text{if } P = (X,Y) \neq (0,0), \\ b & \text{if } P = (0,0), \\ 1 & \text{if } P = 0 = [0,1,0] \end{cases}$$

Then $\operatorname{im}\Theta$ is finite.

Proof ► Suppose

$$P = (x, y) \in \mathcal{E},$$

where $y \neq 0$.

Lemma 9 *Suppose $\mathfrak{p}$ is a prime ideal in K such that*

$$\mathfrak{p} \not| b.$$

Then $\mathfrak{p}$ appears to an even power in x:

$$\mathfrak{p}^{2e} \mid\mid x.$$

Proof of Lemma ▷ Suppose

$$p^e \mid\mid x, \ p^f \mid\mid y.$$

If $e < 0$ then the right-hand side is dominated by x^3, and so $f < 0$ and

$$2f = 3e.$$

On the other hand, if $e > 0$ then

$$\mathfrak{p} \not| x^2 + ax + b$$

since we are supposing that $\mathfrak{p} \not| b$. Thus

$$2f = e.$$

In either case (or if $e = 0$) e is even. ◁

Lemma 10 *We can find a finite number of elements $x_1, \dots, x_r \in K$ such that $\bar{x}_k \in \operatorname{im}\Theta$, and for each x with $\bar{x} \in \operatorname{im}\Theta$ we have*

$$\langle x x_k^{-1} \rangle = \mathfrak{a}^2$$

for some $k \in \{1, \dots, r\}$.

Proof of Lemma ▷ By the last Lemme, the only prime ideals $\mathfrak{p}$ appearing to an odd power in x are the finite number dividing b. Suppose these prime ideals are $\mathfrak{p}_1, \dots, \mathfrak{p}_s$. Consider the 2^s ideals

$$\mathfrak{p}_1^{e_1} \cdots \mathfrak{p}_r^{e_r} \quad (e_1, \dots, e_r \in \{0, 1\}),$$

say

$$\mathfrak{a}_1, \dots, \mathfrak{a}_{2^s}.$$

According to the last Lemma, if $\bar{x} \in \operatorname{im}\Theta$ then

$$\langle x \rangle = \mathfrak{a}_k \mathfrak{b}^2$$

for some $k \in \{1, \dots, s\}$.

If such an x exists for the ideal $\mathfrak{a}_k$ let us choose one, say x_k:

$$\langle x_k \rangle = \mathfrak{a}_k \mathfrak{b}^2.$$

If no such element exists let $x_k = 1$.

Then we see that if

$$\langle x \rangle = \mathfrak{a}_k \mathfrak{b}^2$$

then

$$\langle x x_k^{-1} \rangle = \mathfrak{b}^2.$$

◁

If we are working over $\mathbb{Q}$ it follows that

$$x x_k^{-1} = \pm X^2,$$

and so

$$x \equiv \pm x_k \bmod (K^\times)^2$$

for some k. Hence

$$\operatorname{im}\Theta = \{\pm x_1, \dots, \pm x_r\}.$$

Thus $\operatorname{im}\Theta$ is finite, and so the result is established: $\mathcal{E}(\mathbb{Q})/2\mathcal{E}(\mathbb{Q})$ is finite.

For a general number field K we have a little more work to do.

Let

$$S = \langle \bar{x}_1, \dots, \bar{x}_r \rangle$$

be the subgroup of $K^\times/(K^\times)^2$ generated by $\bar{x}_1, \dots, \bar{x}_r$. This subgroup is finite, since each element of $K^\times/(K^\times)^2$ has order 2.

Let T be the subgroup of $K^\times/(K^\times)^2$

$$T = \{\bar{x} \in \operatorname{im}\Theta : \langle x \rangle = \mathfrak{a}^2\}.$$

Then the last Lemma can be re-stated in the form

$$\operatorname{im}\Theta \subset ST.$$

Lemma 11 *Suppose S, T are 2 finite subgroups of the abelian group G. Then ST is finite; and in fact*

$$\|ST\| \text{ divides } \|S\|\,\|T\|.$$

Proof of Lemma ▷ We have

$$\|ST\| = \|ST/T\|\,\|T\|.$$

Let Φ be the canonical surjective homomorphism

$$\Phi : G \to G/T.$$

If Φ_S is the restriction of Φ to S, then

$$\ker \Phi_S = S \cap T,\ \operatorname{im}\Phi_S = ST/T.$$

It follows from the First Isomorphism Theorem that

$$ST/T \cong S/S \cap T,$$

and so

$$\|ST/T\| \text{ divides } \|S\|.$$

◁

Corollary 15 *The group* $\operatorname{im}\Theta$ *is finite if and only if T is finite.*

Recall that T is the subgroup of $\operatorname{im}\Theta$ formed by those $\bar{x}$ with x expressible in the form $x = \mathfrak{a}^2$. Our next Lemma shows that the set of *all* such $\bar{x}$ (not just those in $\operatorname{im}\Theta$ is finite.

Lemma 12 *Let*

$$S = \{x \in K^\times : \langle x \rangle = \mathfrak{a}^2\}.$$

Then $S \supset (K^\times)^2$; and the quotient-group

$$S/(K^\times)^2$$

is finite.

Proof of Lemma ▷ It is evident that $S \supset (K^\times)^2$, since

$$\langle x^2 \rangle = \langle x \rangle^2.$$

By the finiteness of the class number we can find a finite number of ideals $\mathfrak{a}_1, \dots, \mathfrak{a}_h$ such that for any ideal $\mathfrak{a}$ one of the ideals $\mathfrak{a}\mathfrak{a}_i$ is principal, say

$$\mathfrak{a}\mathfrak{a}_i = \langle a \rangle.$$

Now suppose $x \in S$, say

$$\langle x \rangle = \mathfrak{a}^2.$$

Then

$$\mathfrak{a} = \mathfrak{a}_i \langle a \rangle$$

for some i, and so

$$\langle x \rangle = \mathfrak{a}_i^2 \langle a^2 \rangle.$$

It follows that

$$\mathfrak{a}_i^2 = \langle x a^{-2} \rangle = \langle a_i \rangle,$$

say, for some $a_i \in K^\times$. For each $\mathfrak{a}_i (1 \le i \le h)$ let us choose such an a_i if $\mathfrak{a}_i^2$ is principal; otherwise let us set $a_i = 1$.

Now

$$\langle x \rangle = \langle a_i a^2 \rangle.$$

In other words

$$x = \epsilon a_i a^2,$$

where $\epsilon \in U(K)$ is a unit in K.

By Dirichlet's Units Theorem, the group $U(K)$ of units in K is finitely-generated, say

$$U(K) = \langle \epsilon_1, \dots, \epsilon_m \rangle.$$

Then

$$\epsilon = \epsilon_1^{e_1} \cdots \epsilon_m^{e_m} \quad (e_1, \dots, e_m \in \mathbb{Z}).$$

It follows that

$$\epsilon = \epsilon_1^{e_1} \cdots \epsilon_m^{e_m} \eta^2 \quad (e_1, \ldots, e_m \in \{0, 1\}),$$

where $\eta \in U(K)$.

Putting all this together, we have

$$x = \epsilon_1^{e_1} \cdots \epsilon_m^{e_m} a_i (a\eta)^2$$

In other words,

$$x \equiv \epsilon_1^{e_1} \cdots \epsilon_m^{e_m} a_i \bmod (K^\times)^2.$$

There are only a finite number of elements $\epsilon_1^{e_1} \cdots \epsilon_m^{e_m} a_i$. We conclude that the quotient-group

$$S/(K^\times)^2$$

is finite. $\triangleleft$

Since

$$T \subset S/(K^\times)^2$$

it follows that T is finite. Hence

$$\operatorname{im} \Theta = ST$$

is finite. $\blacktriangleleft$

Corollary 16 *$\mathcal{E}(K)/2\mathcal{E}(K)$ is finite.*

Corollary 17 *If $\mathcal{E}(\mathbb{Q})$ is the elliptic curve*

$$y^2 = x^3 + ax^2 + bx + c \quad (a, b, c \in \mathbb{Q})$$

then

$$\mathcal{E}(\mathbb{Q})/2\mathcal{E}(\mathbb{Q})$$

is finite.

9.6 The Height of a Point

We have shown that $\mathcal{E}(\mathbb{Q})/2\mathcal{E}(\mathbb{Q})$ is finite, say

$$\mathcal{E}/2\mathcal{E} = \{\bar{E}_1, \ldots, \bar{E}_n\},$$

where $E_1, \ldots, E_m \in \mathcal{E}$.

Recall our "plan for infinite descent". Suppose $P \in \mathcal{E}$. Then

$$P - E \in 2\mathcal{E}$$

for some $E \in \{E_1, \ldots, E_m\}$, say

$$P - E_{i_0} = 2P_1.$$

Then similarly

$$P_1 - E_{i_1} = 2P_2$$
$$P_2 - E_{i_2} = 2P_3$$
$$\ldots$$

The points $P = P_0, P_1, P_2, \cdots \in \mathcal{E}(\mathbb{Q})$ — derived by repeated halving — represent our infinite descent. But in what sense are they descending? We need some notion of the 'height' of a point on $\mathcal{E}$.

Definition 9.1 *Suppose $q \in \mathbb{Q}$. Let*

$$q = \frac{m}{n}$$

in lowest terms. Then we set

$$H(q) = \max(|m|, |n|), \; h(q) = \log H(q).$$

Lemma 13 *Suppose $x_1, x_2 \in \mathbb{Q}$. Then*

$$h(x_1 x_2) \le h(x_1) + h(x_2), \; h(x_1^n) = nh(x_1), \; h(x_1 + x_2) \le h(x_1) + h(x_2) + \log 2.$$

Also, if $x_1 \neq 0$,

$$h(x^{-1}) = h(x).$$

Proof of Lemma $\triangleright$ Suppose $x_1 = m_1/n_1, \; x_2 = m_2/n_2$. Then

$$x_1 x_2 = \frac{m_1 m_2}{n_1 n_2}, \; x_1^n = \frac{m_1^n}{n_1^n}, \; x_1 + x_2 = \frac{m_1 n_2 + m_2 n_1}{n_1 n_2}, \; x_1^{-1} = \frac{n_1}{m_1}.$$

The result follows at once. $\triangleleft$

If n_1 and n_2 have a large common factor — which will usually be the case for us — the result for $x_1 + x_2$ can be greatly improved, as the following result illustrates.

Lemma 14 *Suppose*

$$X = \frac{f(x)}{g(x)},$$

where $f(x)$ and $g(x)$ are polynomials of degrees d and e. Then

$$h(X) \le \max(d, e) h(x) + C,$$

for some constant C.

Proof of Lemma ▷ We can assume that $d = e$. For suppose $d < e$. Then we can replace $f(x)$ by $f(x) + g(x)$. This replaces X by $X + 1$; but that does not affect the result, since

$$h(X) - C \leq h(X) \leq h(X) + C$$

from the estimate in the last Lemma for $h(x_1 + x_2)$. If $e < d$ we can apply the same argument after replacing X by X^{-1}.

We may also assume that the coefficients of $f(x), g(x)$ are integral, say

$$f(x) = a_0x^d + a_1x^{d-1} + \cdots + a_d,\ g(x) = b_0x^d + b_1x^{d-1} + \cdots + b_d,$$

where $a_i, b_j \in \mathbb{Z}$. Then

$$\begin{aligned} X &= \frac{a_0m^d + a_1m^{d-1}n + \cdots + a_dn^d}{b_0m^d + b_1m^{d-1}n + \cdots + b_dn^d} \\ &= \frac{M}{N}, \end{aligned}$$

say. Thus

$$|M| \leq (|a_0| + \cdots |a_d|)H(x)^d,\ |N| \leq (|b_0| + \cdots |b_d|)H(x)^d,$$

and so

$$h(X) \leq dh(x) + C.$$

◁

We define the height of a point $P = (x, y)$ to be the height of its x-coordinate.

Definition 9.2 *Suppose $P = (x, y) \in \mathcal{E}(\mathbb{Q})$. Then we set*

$$H(P) = H(x),\ h(P) = h(x).$$

We want to show that our infinite descent is descending in the sense that

$$h(P) > h(P_1) > h(P_2) > \cdots,$$

at least until we drop below a specified height.

This will be the conclusion of the following 3 Lemmas, concerning a given elliptic curve $\mathcal{E}(Q)$.

Lemma 15 *For any constant $C > 0$, there are only a finite number of points $P \in \mathcal{E}(\mathbb{Q})$ with*

$$h(P) \leq C.$$

Proof of Lemma ▷ There are at most $4e^{2C} + 1$ rationals with $e(x) \leq C$, since both denominator and numerator must be chosen from $\{-N, -N + 1, \ldots, N-1, N\}$ where $N = [e^C]$.

For each such x there are at most 2 values of y such that $(x, y) \in \mathcal{E}$. ◁

Lemma 16 *For each point $P_0 \in \mathcal{E}$ there is a constant $C = C(P_0)$ such that*

$$h(P + P_0) \leq 2h(P) + C.$$

Proof of Lemma ▷ Suppose

$$P + P_0 + Q = 0,$$

ie the line P, P_0 meets $\mathcal{E}$ again at Q.

If $P = (x, y)$ then $-P = (x, -y)$. Hence

$$h(-P) = h(P).$$

Thus it is sufficient to prove the result with Q in place of $P + P_0$.

Let

$$P = (x, y),\ P_0 = (x_0, y_0),\ Q = (X, Y).$$

Suppose the equation of the line PP_0

$$y = mx + d.$$

Then

$$m = \frac{y - y_0}{x - x_0}.$$

The line meets the curve where

$$(mx + d)^2 = x^3 + ax^2 + bx + c.$$

Hence

$$x + x_0 + X = m^2 - a.$$

Thus

$$\begin{aligned} X &= \frac{(y - y_0)^2 - (x + x_0 + a)(x - x_0)^2}{(x - x_0)^2} \\ &= \frac{y^2 - 2y_0y + y_0^2 - x^3 - ax^2 + 2x_0x^2 - 2ax_0x - 3x_0^2x - ax_0^2 - x_0^3}{(x - x_0)^2} \\ &= \frac{-2y_0y + 2x_0x^2 + (b - 2ax_0 - 3x_0^2)x + (c + y_0^2 - ax_0^2 - x_0^3}{(x - x_0)^2}, \end{aligned}$$

since $y^2 = x^3 + ax^2 + bx + c$.

The point is that

$$X = \frac{Ay + Bx^2 + Cx + D}{Ex^2 + Fx + G}$$

for some integers A, B, C, D, E, F depending only on P_0.

If $x = m/n$ then

$$y^2 = \frac{m^3 + am^2n + bmn^2 + cn^3}{n^3}.$$

Thus

$$n^4y^2 = m^3n + am^2n^2 + bmn^3 + cn^4.$$

It follows that $n^2y \in \mathbb{Z}$ and

$$|n^2y| \le (1 + |a| + |b| + |c|)^{1/2} H(x)^2.$$

This allows us to apply the argument in the proof of the last Lemma. We have

$$\begin{aligned} X &= \frac{An^2y + Bm^2 + Cmn + Dn^2}{Em^2 + Fmn + Gn^2} \\ &= \frac{M}{N}, \end{aligned}$$

where

$$\begin{aligned} M &\le \left(|A|(1 + |a| + |b| + |c|)^{1/2} + |B| + |C| + |D|\right) H(x)^2 \\ B &\le (|E| + |F| + |G|)H(x)^2. \end{aligned}$$

It follows that

$$H(X) \le CH(x)^2.$$

from which the result follows. $\triangleleft$

Lemma 17 *There is a constant C such that*

$$h(2P) \ge 4h(P) - C$$

for all $P \in \mathcal{E}$.

Proof of Lemma $\triangleright$ Suppose $P = (x, y)$, $2P = (X, Y)$. Let the tangent at P be

$$y = mx + d.$$

If the elliptic curve $\mathcal{E}(\mathbb{Q})$ has equation

$$y^2 = x^3 + ax^2 + bx + c$$

then

$$2y\frac{dy}{dx} = 3x^2 + 2ax + b = f'(x),$$

and so

$$m = \frac{f'(x)}{2y}.$$

The tangent meets $\mathcal{E}$ where

$$(mx + d)^2 = x^3 + ax^2 + bx + c.$$

This has roots x, x, X. Hence

$$2x + X = m^2 - x;$$

and so

$$\begin{aligned} X &= m^2 - a - 2x \\ &= \frac{f'(x)^2 - (a + 2x)4y^2}{4y^2} \\ &= \frac{f'(x)^2 - 4(a + 2x)f(x)^2}{4f(x)^2}. \end{aligned}$$

It follows from Lemma 14 that

$$h(x) \le 4h(x) + C.$$

But we want a result in the opposite direction!

The essential point is that the numerator and denominator of X have no factor in common, as *polynomials*:

$$\gcd(f'(x)^2 - 4(a + 2x)f(x)^2, 4f(x)^2) = \gcd(f'(x)^2, f(x)) = 1,$$

since $\gcd(f'(x), f(x)) = 1$.

Sublemma *Suppose*

$$X = \frac{f(x)}{g(x)},$$

where $f(x), g(x)$ *are polynomials of degrees* d, e, *with* $\gcd(f(x), g(x)) = 1$. *Then*

$$h(X) \geq \max(d, e) h(x) - C$$

for some constant C.

Proof of Lemma ▷ We may suppose that $d = e$, on replacing $f(x)$ or $g(x)$ by $f(x) + g(x)$, if necessary.

We may also assume that the coefficients of $f(x), g(x)$ are integral, say

$$f(x) = a_0 x^d + a_1 x^{d-1} + \cdots + a_d, \ g(x) = b_0 x^d + b_1 x^{d-1} + \cdots + b_d,$$

where $a_i, b_j \in \mathbb{Z}$.

Let $F(x, z), G(x, z)$ be the corresponding homogeneous forms, ie

$$F(x, z) = a_0 x^d + a_1 x^{d-1} z + \cdots + a_d z^d, \ G(x, z) = b_0 x^d + b_1 x^{d-1} z + \cdots + b_d z^d.$$

If $x = m/n$ then

$$X = \frac{F(m, n)}{G(m, n)}.$$

We have to show that this is almost in its lowest terms.

Since $\gcd(f(x), g(x)) = 1$, we can find polynomials $u(x), v(x) \in \mathbb{Q}[x]$ such that

$$u(x)f(x) + v(x)g(x) = 1.$$

On 'multiplying out' the denominators of the coefficients, and passing to the homogeneous forms, we obtain polynomials $U(x, z), V(x, z) \in \mathbb{Z}[x, z]$ such that

$$U(x, z)F(x, z) + V(x, z)G(x, z) = Az^N$$

where A is a non-zero integer, and $N \in \mathbb{N}$.

In particular,

$$U(m, n)F(m, n) + V(m, n)G(m, n) = An^N$$

It follows that

$$\gcd(F(m, n), G(m, n)) \mid An^N.$$

On the other hand

$$\gcd(F(m, n), n) \mid a_0 m^d.$$

Since $\gcd(m, n) = 1$ this implies that

$$\gcd(F(m, n), n) \mid a_0.$$

It follows that

$$\gcd(F(m, n), n^N) \mid a_0^N,$$

and so

$$\gcd(F(m, n), An^N) \mid Aa_0^N.$$

Hence

$$\gcd(F(m, n), G(m, n)) \mid Aa_0^N.$$

We are nearly there. We have shown that

$$X = \frac{F(m, n)}{G(m, n)} = \frac{M}{N},$$

say, is almost in its lowest terms. It only remains to show that the numerator or denominator is of the correct order of magnitude. This is 'trivial but not obvious'.

Let

$$M(x) = \max(|f(x)|, |g(x)|).$$

Since

$$\frac{f(x)}{x^d} \to a_0 \text{ as } x \to \infty$$

there exist constants $C_1 > 0, C_2 > 0$ such that

$$M(x) \geq C_1 |x|^d$$

for $|x| \geq C_2$.

On the other hand, since $f(x), g(x)$ have no root in common, there is a constant $C_3 > 0$ such that

$$M(x) \geq C_3$$

for $|x| \leq C_2$. It follows that

$$M(x) \geq (C_3 C_2^{-d}) |x|^d$$

for $|x| \leq C_2$.

Putting these together,

$$M(x) \geq C_4 |x|^d$$

for all x, with $C_4 = \min(C_1, C_3 C_2^{-d})$. On setting $x = m/n$, and multiplying out, this gives

$$\max(M, N) = \max(F(m, n), G(m, n)) \ge C_4 |m|^d.$$

By the same argument

$$M(x) \ge C_5$$

for all x, where $C_5 = \min(C_1 C_2^d, C_3) > 0$. This gives

$$\max(M, N) \ge C_5 |n|^d.$$

We conclude that

$$\max(M, N) \ge C_6 H(x)^d,$$

with $C_6 = \min(C_4, C_5)$. Since we know that

$$\gcd(M, N) \le A a_0^N,$$

we conclude that if

$$X = \frac{M'}{N'}$$

in its lowest terms then

$$H(X) = \min(|M'|, |N'|) \ge C_7 H(x)^d,$$

with $C_7 = C_6/(A a_0^N) > 0$; and so finally,

$$h(X) \ge d h(x) - C.$$

◁

In particular, applying this to our formula for $2P$, we have shown that

$$h(2P) \ge 4h(P) - C.$$

◁

9.7 Putting It All Together

Recall that each step of our infinite descent is of the form

$$P_i - E_j = 2P_{i+1},$$

where E_j is one of a fixed (and finite) set of points $E_1, \ldots, E_m$. By Lemma 17,

$$h(P_i - E_j) \ge 4h(P_{i+1}) - c_1.$$

But by Lemma 16 (and the fact that $h(-P) = h(P)$),

$$h(P_i - E_j) \le 2h(P) + c_2.$$

Combining these,

$$2h(P_i) + c_2 \ge 4h(P_{i+1}) - c_1.$$

Hence

$$h(P_{i+1}) \le \frac{1}{2} h(P_i) + c_3$$

with $c_3 = (c_1 + c_2)/4$.

We have shown therefore that

$$h(P_i) > C \Longrightarrow h(P_{i+1}) < h(P_i),$$

for some constant $C > 0$. Let the points of $\mathcal{E}$ with $h(P) \le C$ be

$$P_1, \ldots, P_n.$$

Our infinite descent must lead to one of these points. We see therefore that for any point $P \in \mathcal{E}$ is expressible in the form

$$P = u_1 E_1 + \cdots + u_m E_m + P_i,$$

where $u_1, \ldots, u_r \in \mathbb{N}$.

We conclude that $\mathcal{E}(\mathbb{Q})$ is generated by the points $E_1, \ldots, E_m, P_1, \ldots, P_n$.

9.8 The formula for $\operatorname{rank}(\mathcal{E})$

Since we now know that $\mathcal{E}$ is finitely-generated, it follows from the Structure Theorem for Finitely Generated Abelian Groups that

$$\mathcal{E} = \mathbb{Z} \oplus \cdots \oplus \mathbb{Z} \oplus \mathbb{Z}/(p_1^{e_1}) \oplus \cdots \mathbb{Z}/(p_s^{e_s}),$$

where there are $r = \operatorname{rank}(\mathcal{E})$ copies of $\mathbb{Z}$.

Proposition 9.6 *Let*

$$d = \begin{cases} 0 & \text{if there are 0 points of order 2 on } \mathcal{E}, \\ 1 & \text{if there is 1 point of order 2 on } \mathcal{E}, \\ 2 & \text{if there are 3 points of order 2 on } \mathcal{E}. \end{cases}$$

Then

$$\|\mathcal{E}/2\mathcal{E}\| = 2^s,$$

where

$$s = r + d.$$

Proof ► If

$$A = A_1 \oplus \cdots \oplus A_m$$

then

$$2A = 2A_1 \oplus \cdots \oplus 2A_m$$

and so

$$A/2A = A_1/2A_1 \oplus \cdots \oplus A_m/2A_m.$$

Thus it is sufficient to consider the factors of $\mathcal{E}$.

Evidently the r copies of $\mathbb{Z}$ will give rise to r copies of $\mathbb{Z}/(2)$.

Lemma 18 *If $A = \mathbb{Z}/(2^e)$ then*

$$A/2A = \mathbb{Z}/(2).$$

Proof of Lemma ▷ Let g be a generator of A, so that

$$A = \{0, g, 2g, \ldots, (2^e - 1)g\}$$

Then

$$2A = \{0, 2g, 4g, \ldots, (2^e - 2)g\}.$$

Thus half the elements of A are in $2A$, and so $A/2A$ is of order 2, ie $A/2A = \mathbb{Z}/(2)$. ◁

Lemma 19 *If $A = \mathbb{Z}/(p^e)$, where p is odd,then*

$$A/2A = 0.$$

Proof of Lemma ▷ Consider the map

$$\theta : A \to A : a \mapsto 2a.$$

Then

$$\ker \theta = \{a \in A : 2a = 0\} = 0,$$

since by Lagrange's Theorem there are no elements of order 2 in A. Hence θ is injective, and so surjective, ie $2A = A$, and $A/2A = 0$. ◁

From the two Lemmas it follows that the number of copies of $\mathbb{Z}/(2)$ in

$$\mathcal{E}/2\mathcal{E} = \mathbb{Z}/(2) + \cdots + \mathbb{Z}/(2)$$

is equal to $r + f$, where f is the number of factors of the form $\mathbb{Z}/(2^e)$. It remains to show that $f = d$.

Lemma 20 *The number of elements of order 2 in A is $2^f - 1$, where f is the number of factors of the form $\mathbb{Z}/(2^e)$.*

Proof of Lemma ▷ An element of a direct sum

$$A = A_1 \oplus A_2 \oplus \cdots \oplus A_m$$

is of order 1 or 2 if and only if that is true of each component:

$$2(a_1, a_2, \ldots, a_m) = 0 \iff 2a_1 = 0,\ 2a_2 = 0, \ldots, 2a_m = 0.$$

But there is no element of order 2 in $\mathbb{Z}/(p^e)$ if p is odd, by Lagrange's Theorem; while there is just one element of order 2 in $\mathbb{Z}/(2^e)$, nameley $2^{e-1} \bmod 2^e$.

Thus we have two choices in each factor $\mathbb{Z}/(2^e)$, and one choice in each factor $\mathbb{Z}/(p^e)$ (p odd).

It follows that the number of elements of order 1 or 2 is 2^f where f is the number of factors of the form $\mathbb{Z}/(2^e)$; and so the number of elements of order 2 is $2^f - 1$. ◁

◄

9.9 The square-free part

Each rational $x \in \mathbb{Q}^\times$ is uniquely expressible in the form

$$x = dy^2,$$

where $y \in \mathbb{Q}^\times$ and d is a square-free integer. Explicitly, if

$$x = \pm 2^{e_2} 3^{e_3} 5^{e_5} \cdots$$

then

$$x = \pm 2^{\epsilon_2} 3^{\epsilon_3} 5^{\epsilon_5} \cdots$$

where each $\epsilon_p \in \{0, 1\}$ is given by

$$\epsilon_p \equiv e_p \bmod 2.$$

For example,

$$x = 2/3 \mapsto d = 6, \quad x = -3/4 \mapsto -3.$$

We may call d the *square-free part* of x.

Thus each $\bar{x} \in \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ is represented by a unique square-free integer d, establishing an isomorphism

$$\mathbb{Q}^\times/\mathbb{Q}^{\times 2} \longleftrightarrow D,$$

where D is the group formed by the square-free integers under multiplication modulo squares, eg

$$2 \cdot 6 = 3, \quad -3 \cdot 6 = -2.$$

Let us see how to use this to compute the rank. Recall that

$$\mathcal{E}/\mathcal{E}^2 \cong \operatorname{im} \Theta$$

where

$$\Theta = \theta_\alpha \times \theta_\beta \times \theta_\gamma,$$

with θ_α, for example, given by

$$P = (x, y) \mapsto \begin{cases} \overline{x - \alpha} \text{ if } x \neq \alpha \\ p'(\alpha) \text{ if } x = \alpha \end{cases}$$

If $P = (x, y)$ is on the elliptic curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx + c \qquad (a, b, c \in \mathbb{Z})$$

then

$$x = \frac{m}{t^2}, \quad y = \frac{M}{t^3}$$

where $m, M, t \in \mathbb{Z}$ with $\gcd(m, t) = 1 = \gcd(M, t)$ and $t > 0$.

9.10 An example

Consider the elliptic curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 - x = x(x-1)(x+1).$$

Here

$$\alpha = 0, \ \beta = 1, \ \gamma = -1,$$

so that

$$p'(0) = (0-1)(0+1) = -1, \quad p'(1) = (1-0)(1+1) = 2, \quad p'(-1) = (-1-0)(-1-1) = 2.$$

Thus, from above,

$$\operatorname{im} \Theta \subset S = \{(d, e, f) : d \mid 1, \ e \mid 2, f \mid 2\}.$$

This gives 32 choices:

$$d = \pm 1, \quad e = \pm 1, \pm 2, \quad f = \pm 1, \pm 2.$$

It follows (since $32 = 2^5$) that

$$\|\mathcal{E}/2\mathcal{E}\| \leq 5,$$

and so

$$\operatorname{rank} \mathcal{E} \leq 3.$$

However, we can restrict the range of $\operatorname{im} \Theta$ much more than this. In the first place, since

$$x(x-1)(x+1) = y^2,$$

it follows that def is a perfect square, say

$$def = g^2.$$

This implies firstly that $def > 0$, and secondly that each prime p dividing any of d, e, f must in fact divide just two of them. This reduces the number of cases to 8:

$$(d, e, f) = (1, 1, 1), \ (1, -1, -1), \ (-1, 1, -1), \ (-1, -1, 1), \ (1, 2, 2), \ (1, -2, -2), \ (-1, 2, -2), \ (-1,$$

We can reduce the number still further by observing that since

$$m = du^2, \ m - t^2 = ev^2, \ m + t^2 = fw^2,$$

it follows that

$$d < 0 \Longrightarrow m < 0 \Longrightarrow m - t^2 < 0 \Longrightarrow e < 0,$$

while

$$d > 0 \Longrightarrow m > 0 \Longrightarrow m + t^2 > 0 \Longrightarrow f > 0.$$

This leaves just 4 choices for d, e, f:

$$(d, e, f) = (1, 1, 1),\ (-1, -1, 1),\ (1, 2, 2),\ (-1, -2, 2).$$

Thus

$$\|\mathcal{E}/2\mathcal{E}\| \leq 4$$

Since $d = 2$ (as there are 3 points of order 2),

$$\|\mathcal{E}/2\mathcal{E}\| = 2^{r+d} \geq 4.$$

We conclude that

$$\operatorname{rank} \mathcal{E} = 0.$$

9.11 Another example

Now let us consider the elliptic curve

$$y^2 = x^3 - x = x(x - 2)(x + 2).$$

Here

$$p'(0) = -4,\ p'(2) = 8,\ p'(-2) = 8,$$

and so

$$\mathcal{E}/2\mathcal{E} = \operatorname{im} \Theta \subset \{(d, e, f) : d, e, f \mid 2\}$$

The group on the right contains 2^6 elements, since each of d, e, f can take the values $\pm 1, \pm 2$.

But as before, the condition

$$def = g^2$$

restricts the choice considerably. Firstly,

$$d < 0 \Longrightarrow e < 0. \quad d > 0 \Longrightarrow f > 0.$$

Secondly, the factor 2 occurs in none, or just two, of d, e, f. This reduces the choice to

$$(d, e, f) = (1, 1, 1),\ (-1, -1, 1),\ (1, 2, 2),\ (-1, -2, 2),\ (2, 1, 2),\ (-2, -1, 2),\ (2, 2, 1),\ (-2, -2, 1).$$

Thus the rank is either 0 or 1. Can we reduce the choice further, and reduce the rank to 0? or conversely, can we find a point of infinite order on the curve, and so show that the rank is 1?

Note that it only necessary to eliminate one case; for we know that $\|\mathcal{E}/2E\| = 2^s \geq 4$, since there are 3 points of order 2 (and so $d = 2$).

Suppose

$$(d, e, f) = (-1, -1, 1).$$

In this case,

$$m = -u^2,\ m - 2t^2 = -v^2,\ m + 2t^2 = w^2.$$

Thus

$$u^2 - v^2 = 2t^2 = u^2 + w^2.$$

Now $a^2 \equiv 0$ or 1 mod 4 according as a is even or odd. Since $u^2 - v^2$ is even it followu, v are both even or both odd; and in either case $u^2 - v^2 \equiv 0 \bmod 4$. So t is even, and therefore u, v must both be odd, since $\gcd(m, t) = 1 = \gcd(m - 2t^2, t)$.

9.12 Third example

Consider the elliptic curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x(x - 2)(x + 4) = x^3 + 2x^2 - 8x.$$

The point

$$P = (-1, 3) \in \mathcal{E}.$$

(We chose α, β, γ to give this result.)

The slope at P is

$$\begin{aligned} \frac{dx}{dy} &= \frac{3x^2 + 4x - 8}{2y} \\ &= -\frac{3}{2} \end{aligned}$$

at P. It follows that P is of infinite order (since $2P$ has non-integral coordinates). Thus

$$r = \operatorname{rank}(\mathcal{E}) \geq 1.$$

We have
$$p'(0) = -8, \quad p'(2) = 12, \quad p'(-4) = 24.$$
Thus
$$\operatorname{im}\Theta \subset S\{(d,e,f) : d \mid 2, e \mid 6, f \mid 6; def = g^2\}.$$
Note that any two of d, e, f determine the third since eg $f = de$ (modulo squares).

If
$$P = (m/t^2, M/t^3) \mapsto (d,e,f)$$
then
$$m = du^2, \quad m - 2t^2 = ev^2, \quad m + 4t^2 = fw^2.$$
Thus
$$d > 0 \Longrightarrow m > 0 \Longrightarrow f > 0 \Longrightarrow e > 0,$$
while
$$d < 0 \Longrightarrow m < 0 \Longrightarrow e < 0 \Longrightarrow f > 0.$$
(So $f > 0$ in all cases.)

It follows that
$$\|S\| = 16,$$
with
$$S = \{d = \pm 1, \pm 2, \; f = 1, 2, 3, 6\}.$$
It follows that $s \le 4$, and so
$$\operatorname{rank}(\mathcal{E}) = s - d = s - 2 \le 2.$$
Thus $\operatorname{rank}(\mathcal{E}) = 1$ or 2.

In order to prove that $\operatorname{rank}(\mathcal{E}) = 1$ it is sufficient to show that one of the 16 elements of S does not lie in $\operatorname{im}\Theta$. For $\|S\|$ is a power of 2, so if it is < 16 it must be ≤ 8.

Let us take the element $(-1, -1, 1)$. Suppose this arises from a point $P = (m/t^2, M/t^3)$, where for the moment we assume that P is not of order 2. Then
$$m = -u^2, \quad m - 2t^2 = -v^2, \quad m + 4t^2 = w^2.$$
Thus
$$2t^2 = v^2 - u^2, \quad 4t^2 = u^2 + w^2.$$
From the second equation,
$$u^2 + w^2 \equiv 0 \bmod 4 \Longrightarrow u, w \text{ even},$$

since $a^2 \equiv 0$ or $1 \bmod 4$ according as a is even or odd. It follows that t is odd, since
$$\gcd(m,t) = 1 \Longrightarrow \gcd(u,t) = 1.$$
But then $t^2 \equiv 1 \bmod 4$, and so
$$v^2 - u^2 \equiv 2 \bmod 4,$$
which is impossible.

(Alternatively, adding the two equations,
$$6t^2 = v^2 + w^2.$$
Thus
$$\begin{aligned} v^2 + w^2 \equiv 0 \bmod 3 &\Longrightarrow v \equiv w \equiv 0 \bmod 3 \\ &\Longrightarrow t \equiv 0 \bmod 3 \\ &\Longrightarrow u \equiv 0 \bmod 3, \end{aligned}$$
contradicting $\gcd(m,t) = 1$.)

9.13 Final example

The elliptic curve
$$\mathcal{E}(\mathbb{Q}) : y^2 = x(x+1)(x-14) = x^3 - 13x^2 - 14x$$
is more complicated, but the method is the same.

We have
$$p'(0) = -14, \quad p'(-1) = 15, \quad p'(14) = 14 \cdot 15.$$
Thus
$$\operatorname{im}\Theta \subset S = \{(d,e,f) : d \mid 14, \; e \mid 15, \; f \mid 14 \cdot 15; def = g^2\}.$$
if $P = (m/t^2, M/t^3) \mapsto (d,e,f)$ $(M \ne 0)$ then
$$m = du^2, \quad m + t^2 = ev^2, \quad m - 14t^2 = fw^2.$$
In particular,
$$d > 0 \Longrightarrow e > 0 \Longrightarrow f > 0$$

while

$$d < 0 \Longrightarrow f < 0 \Longrightarrow e > 0$$

(giving $e > 0$ in all cases).

We have

$$d = \pm 1, \pm 2, \pm 7, \pm 14, \quad e = 1, 3, 5, 15.$$

Thus

$$\|S\| = 2^5 \Longrightarrow s \le 5 \Longrightarrow r \le 3.$$

The elements of order 2 give rise to the points

$$\begin{aligned} (0,0) &\mapsto (p'(0), 1, -14) = (-14, 1, -14), \\ (-1,0) &\mapsto (-1, p'(-1), -15) = (-1, 15, -15), \\ (14,0) &\mapsto (14, 15, p'(14)) = (14, 15, 14 \cdot 15), \end{aligned}$$

while of course

$$0 = [0, 1, 0] \mapsto (1, 1, 1).$$

Thus the torsion group gives rise the subgroup

$$D = \{(1,1,1), (-14,1,-14), (-1,15,-15), (14,15,14\cdot 15).$$

We can regard S as a 5-dimensional vector space over $\mathcal{F}_2$, with 5 coordinates defined by: the sign of d, the factor 2 in d, the factor 7 in d, the factor 3 in e, the factor 5 in e. Thus

$$\begin{aligned} (0,0) &\mapsto (-14, 1, -14) \longleftrightarrow (1,1,1,0,0), \\ (-1,0) &\mapsto (-1, 15, -15) \longleftrightarrow (1,0,0,1,1), \\ (14,0) &\mapsto (14, 15, 14 \cdot 15) \longleftrightarrow (0,1,1,1,1). \end{aligned}$$

Our aim is to prove that $\operatorname{rank}(\mathcal{E}) = 0$ by showing that $\operatorname{im}\Theta = D$. At first sight one might think we would have to apply our congruence technique to $2^5 - 2^2 = 28$ cases. However, we can simplify the task by choosing a *complementary subspace* to D – that is, a subspace of $U \subset S$ of dimension 3 such that

$$U \cap D = 0,$$

in which case

$$S = D \oplus U.$$

If now we can show that no elements of U except for $(1,1,1)$ are in $\operatorname{im}\Theta$ then it will follow that

$$S = \operatorname{im}\Theta \oplus U;$$

whence

$$\dim \operatorname{im}\Theta = \dim D \Longrightarrow \operatorname{im}\Theta = D.$$

For our subspace U let us take those vectors with 3rd and 5th components 0, ie

$$U = \{(d,e,f) \in S : d = \pm 1, \pm 2, \ e = 1, 3\}.$$

We see at once that $U \cap D = \{(1,1,1)\}$ (the zero element of our vector space), so U is — as required — complementary to D. It is sufficient therefore to show that no element of U apart from $(1,1,1)$ can be in $\operatorname{im}\Theta$. (This reduces the number of cases to be considered from 28 to 7.)

1. $(-1,1,-1)$: in this case

   $$m = -u^2, \quad m + t^2 = v^2, \quad m - 14t^2 = -w^2,$$

   ie

   $$t^2 = u^2 + v^2, \quad 14t^2 = w^2 - u^2.$$

   From the second equation t must be even, since otherwise $w^2 - u^2 \equiv 2 \bmod 4$, which is impossible.

   But then from the first equation, $u^2 + v^2 \equiv 0 \bmod 4$, which implies that u, v are both even, contradicting $\gcd(u,t) = 1$.

2. $(2,1,2)$: in this case

   $$m = 2u^2, \quad m + t^2 = v^2, \quad m - 14t^2 = 2w^2,$$

   ie

   $$t^2 = v^2 - 2u^2, \quad 7t^2 = u^2 - w^2.$$

   From the second equation t must be even, since otherwise $7t^2 \equiv 3 \bmod 4$, and $u^2 - w^2$ cannot be $\equiv 3 \bmod 4$.

   But then from the first equation, v is even and so u is even, contradicting $\gcd(u,t) = 1$.

3. $(-2,1,-2)$: in this case

   $$m = -2u^2, \quad m + t^2 = v^2, \quad m - 14t^2 = -2w^2,$$

ie

$$t^2 = v^2 + 2u^2, \quad 7t^2 = w^2 - u^2.$$

As in the last case, from the second equation t must be even, and then from the first equation, so must v and u, contradicting $\gcd(u, t) = 1$.

4. $(1, 3, 3)$: in this case

$$m = u^2, \quad m + t^2 = 3v^2, \quad m - 14t^2 = 3w^2,$$

ie

$$t^2 = 3v^2 - u^2, \quad 14t^2 = u^2 - 3w^2.$$

From the second equation

$$u^2 - 3w^2 \equiv 0 \bmod 7.$$

Since 3 is a quadratic non-residue mod7, it follows that $u \equiv w \equiv 0 \bmod 7$, which implies (by the second equation) that $7 \mid t$, so again $\gcd(t, u) > 1$.

5. $(-1, 3, -3)$: in this case

$$m = -u^2, \quad m + t^2 = 3v^2, \quad m - 14t^2 = -3w^2,$$

ie

$$t^2 = 3v^2 + u^2, \quad 14t^2 = 3w^2 - u^2.$$

As in the last case, since 3 is not a quadratic residue $\bmod 7$, the second equation implies that $7 \mid u, w, t$, contradicting $\gcd(u, t) = 1$.

6. $(2, 3, 6)$: in this case

$$m = 2u^2, \quad m + t^2 = 3v^2, \quad m - 14t^2 = 6w^2,$$

ie

$$t^2 = 3v^2 - 2u^2, \quad 14t^2 = u^2 - 6w^2.$$

Again, since 6 is not a quadratic residue mod7, this leads to a contradiction.

7. $(-2, 3, -6)$: in this case

$$m = -2u^2, \quad m + t^2 = 3v^2, \quad m - 14t^2 = -6w^2,$$

ie

$$t^2 = 3v^2 + 2u^2, \quad 7t^2 = u^2 - 3w^2.$$

Since 3 is not a quadratic residue mod7, this again leads to a contradiction.

We conclude that

$$\operatorname{im} \Theta = D,$$

ie

$$\operatorname{rank} \mathcal{E} = 0.$$

Chapter 10

Mordell Revisited

10.1 Introduction

There is an alternative way of proving Mordell's Theorem, by 'factorising' the doubling map

$$\mathcal{E} \to \mathcal{E} : P \mapsto 2P;$$

although the factors are not, admittedly, homomorphisms from $\mathcal{E}$ to itself, but involve a 'twin' elliptic curve $\bar{\mathcal{E}}$. The resulting computations are much simpler. Moreover, the use of algebraic numbers is avoided if $f(x)$ has *one* rational root. (In the previous method, algebraic numbers were avoided if *two* — and therefore all three — of the roots of $f(x)$ are rational.)

The only disadvantage of this alternative method is that it requires either an act of faith, in which 'magic' formulae are pulled out of a hat; or else a rather lengthy digression into elliptic curves over $\mathbb{C}$.

10.2 The factors of the doubling map

Suppose

$$\mathcal{E}(\mathbb{C}) = \mathbb{C}/\Lambda$$

is the complex elliptic curve associated to a lattice $\Lambda \subset \mathbb{C}$. Let ω_1, ω_2 be a basis for Λ. Recall that the map

$$z \mapsto (\wp(z), \wp'(z)/2)$$

establishes a one-one correspondence

$$\Phi : \mathbb{C}/\Lambda \leftrightarrow \mathcal{E}(\mathbb{C}),$$

where $\mathcal{E}(\mathbb{C})$ is the curve

$$y^2 = x^3 + bx + c,$$

with coefficients

$$b = -15g_2,\ c = -35g_3,$$

where

$$g_r = \sum_{\omega \in \Lambda, \omega \neq 0} \frac{1}{\omega^{2r}}.$$

Under this correspondence, the 'doubling' homomorphism

$$\Phi : P \mapsto 2P$$

corresponds to the map

$$\phi : z \bmod \Lambda \mapsto 2z \bmod \Lambda.$$

We can express this in the commutative diagram

$$\begin{array}{ccc} \mathcal{E} & \xrightarrow{\Phi} & \mathcal{E} \\ \updownarrow & & \updownarrow \\ \mathbb{C}/\Lambda & \xrightarrow{\phi} & \mathbb{C}/\Lambda \end{array}$$

Let $\bar{\Lambda}$ be the lattice generated by $\frac{1}{2}\omega_1, \omega_2$. so that

$$\tfrac{1}{2}\Lambda \subset \bar{\Lambda} \subset \Lambda$$

(where $\frac{1}{2}\Lambda$ is generated by $\frac{1}{2}\omega_1, \frac{1}{2}\omega_2$). The homomorphism $\Phi : P \mapsto 2P$ can now be split into 2 operations, first doubling in the ω_1-direction, and then in the ω_2-direction. More precisely,

$$\phi = \theta_3\theta_2\theta_1,$$

where

$$\theta_1 : \mathbb{C}/\Lambda \to \mathbb{C}/\bar{\Lambda},\ \theta_2 : \mathbb{C}/\bar{\Lambda} \to \mathbb{C}/\tfrac{1}{2}\Lambda,\ \theta_3 : \mathbb{C}/\tfrac{1}{2}\Lambda \to \mathbb{C}/\Lambda$$

are the homomorphisms

$$\begin{aligned} \theta_1 : z \bmod \Lambda &\to z \bmod \bar{\Lambda}, \\ \theta_2 : z \bmod \bar{\Lambda} &\to z \bmod \tfrac{1}{2}\Lambda, \\ \theta_3 : z \bmod \tfrac{1}{2}\Lambda &\to 2z \bmod \Lambda. \end{aligned}$$

The map θ_3 is just the isomorphism $(x, y) \mapsto (x/4, y/8)$ associated to the similarity $\frac{1}{2}\Lambda \to \Lambda$; it is convenient to combine it with θ_2.

Let $\bar{\mathcal{E}}(\mathbb{C})$ be the elliptic curve associated to the lattice $\bar{\Lambda}$:

$$\bar{\mathcal{E}} = \mathbb{C}/\bar{\Lambda}.$$

Then $\theta_1, \theta_3\theta_2$ define homomorphisms

$$\Theta : \mathcal{E} \to \bar{\mathcal{E}},\ \bar{\Theta} : \bar{\mathcal{E}} \to \mathcal{E}.$$

giving the factorisation

$$\Phi = \bar{\Theta}\Theta$$

of $\Phi : P \mapsto 2P$. This can be expressed in the commutative diagram

$$\begin{array}{ccccc} \mathcal{E} & \xrightarrow{\Theta} & \bar{\mathcal{E}} & \xrightarrow{\bar{\Theta}} & \mathcal{E} \\ \updownarrow & & \updownarrow & & \updownarrow \\ \mathbb{C}/\Lambda & \xrightarrow{\theta_1} & \mathbb{C}/\bar{\Lambda} & \xrightarrow{\theta_3\theta_2} & \mathbb{C}/\Lambda \end{array}$$

Note that

$$\bar{\Phi} = \Theta\bar{\Theta} : \bar{\mathcal{E}} \to \bar{\mathcal{E}}$$

is also a doubling map, this time on $\bar{\mathcal{E}}$, being given by the composition

$$\theta_1\theta_3\theta_2 : z \bmod \bar{\Lambda} \mapsto z \bmod \tfrac{1}{2}\Lambda \mapsto 2z \bmod \Lambda \mapsto 2z \bmod \bar{\Lambda}.$$

All that is straightforward. But how does it translate into geometric terms? What is the elliptic curve $\bar{\mathcal{E}}$? And what are the algebraic formulae for the maps $\Theta, \bar{\Theta}$?

As we just noted, $\mathcal{E}$ is parametrised by

$$(x, y) = (\varphi(z), \varphi'(z)/2).$$

Similary $\bar{\mathcal{E}}$ is parametrised by

$$(x, y) = (\varphi_{\bar{\Lambda}}(z), \varphi'_{\bar{\Lambda}}(z)/2).$$

To determine Θ, we must express $\varphi_{\bar{\Lambda}}(z)$ in terms of $\varphi(z)$.

Proposition 10.1 *Let $\Lambda = \langle\omega_1, \omega_2\rangle$, and let $\bar{\Lambda} = \langle\omega_1/2, \omega_2\rangle$. Then, writing $\varphi(z)$ for $\varphi_\Lambda(z)$,*

$$\varphi_{\bar{\Lambda}}(z) = \frac{\varphi(z)^2 - \alpha\varphi(z) + 3\alpha^2 + b}{\varphi(z) - \alpha},$$

where $\alpha = \varphi(\omega_1/2)$, and b is the coefficient in the functional equation

$$(\varphi'(z)/2)^2 = \varphi(z)^3 + b\varphi(z) + c.$$

Proof ► Since $\bar{\Lambda} \subset \Lambda$, $\varphi_{\bar{\Lambda}}(z)$ is elliptic with respect to Λ. It is also even. Hence it is a rational function of $\varphi(z)$,

$$\varphi_{\bar{\Lambda}}(z) = R(\varphi(z)),$$

where $R(w) = P(w)/Q(w)$ with polynomials $P(w), Q(w)$.

It is easy to see that the function

$$f(z) = \varphi(z) + \varphi(z + \omega_1/2)$$

has periods $\omega_1/2, \omega_2$, and so is elliptic with respect to $\bar{\Lambda}$. Since it has a double pole at 0, and no other poles inside Π_1,

$$f(z) = A\varphi_{\bar{\Lambda}}(z) + B$$

for some constants A, B. In the neighbourhood of $z = 0$,

$$f(z) = \frac{1}{z} + \varphi(\omega_1/2).$$

It follows that

$$f(z) = \varphi_{\bar{\Lambda}}(z) + \varphi(\omega_1/2).$$

Thus

$$\varphi_{\bar{\Lambda}}(z) = \varphi(z) + \varphi(z + \omega_1/2) - \varphi(\omega_1/2).$$

Let

$$\alpha = \varphi(\omega_1/2),\ \beta = \varphi(\omega_2/2),\ \gamma = \varphi(\omega_1/2 + \omega_2/2).$$

Recall that

$$\varphi'(\omega_1/2) = 0,\ \varphi'(\omega_2/2) = 0,\ \varphi'(\omega_1/2 + \omega_2/2) = 0,$$

since $\varphi'(z)$ is an odd function. Thus

$$(\alpha, 0),\ (\beta, 0),\ (\gamma, 0)$$

are just the 3 points of order 2 on $\mathcal{E}$.

We have

$$\varphi_{\bar{\Lambda}}(z) = \varphi(z) + \varphi(z + \omega_1/2) - \alpha;$$

we want to express $\varphi(z+\omega_1/2)$ in terms of $\varphi(z)$. The function $\varphi(z+\omega_1/2)$ is elliptic with respect to Λ, and has a double pole at $\omega_1/2$, and no other poles inside Π. The function $\varphi(z) - \varphi(\omega_1/2) = \varphi(z) - \alpha$ has a double zero at $\omega_1/2$, since $\varphi'(\omega_1/2) = 0$. Thus

$$F(z) = \varphi(z + \omega_1/2)(\varphi(z) - \alpha)$$

has a double pole at the points of Λ, and no other poles. Since $F(z)$ is even, it follows that

$$F(z) = C\wp(z) + D$$

for some constants C, D. To determine these constants we expand $F(z)$ around $z = 0$.

By Taylor's theorem,

$$\wp(z + \omega_1/2) = \wp(\omega_1/2) + \frac{1}{2}\wp''(\omega_1/2)z^2 + \frac{1}{24}\wp''''(\omega_1/2)z^4 + O(z^6).$$

On differentiating the functional equation

$$\wp'(z)^2 = 4\wp(z)^3 + 4b\wp(z) + 4c,$$

we deduce that

$$\wp''(z) = 2(3\wp(z)^2 + b).$$

Differentiating twice more,

$$\wp''''(z) = 12\left(\wp(z)\wp''(z) + \wp'(z)^2\right).$$

In particular,

$$\wp''(\omega_1/2) = 2(3\alpha^2 + b),\ \wp''''(\omega_1/2) = 24\alpha(3\alpha^2 + b).$$

Thus

$$\wp(z + \omega_1/2) = \alpha + (3\alpha^2 + b)z^2 + \alpha(3\alpha^2 + b)z^4 + O(z^6)$$

in the neighbourhood of $z = 0$. It follows that

$$\begin{aligned} F(z) &= \left(\alpha + (3\alpha^2 + b)z^2\right)\left(\frac{1}{z^2} - \alpha\right) + O(z^2) \\ &= \frac{\alpha}{z^2} + (2\alpha^2 + b). \end{aligned}$$

Hence

$$F(z) = \alpha\wp(z) + 2\alpha^2 + b.$$

We conclude that

$$\begin{aligned} \wp_{\bar\Lambda}(z) &= \wp(z) - \alpha + \frac{\alpha\wp(z) + 2\alpha^2 + b}{\wp(z) - \alpha} \\ &= \frac{\wp(z)^2 - \alpha\wp(z) + 3\alpha^2 + b}{\wp(z) - \alpha}. \end{aligned}$$

◄

Corollary 18 *The derivative of $\wp_{\bar\Lambda}(z)$ is given by:*

$$\wp'_{\bar\Lambda}(z) = \frac{\wp(z)^2 - 2\alpha\wp(z) - 2\alpha^2 - b}{(\wp(z) - \alpha)^2}\wp'(z).$$

We see from this that the homomorphism $\Theta : \mathcal{E} \to \bar{\mathcal{E}}$ is given by

$$\Theta : (x, y) \mapsto \left(\frac{x^2 - \alpha x + 3\alpha^2 + b}{x - \alpha}, \frac{x^2 - 2\alpha x - 2\alpha^2 - b}{(x - \alpha)^2}y\right)$$

if $x \neq \alpha$, while

$$\Theta(\alpha, 0) = O.$$

But what is the curve $\bar{\mathcal{E}}$? Recall that

$$\begin{aligned} \wp(z) &= \frac{1}{z^2} + 3g_2z^2 + 5g_3z^4 + O(z^6) \\ &= \frac{1}{z^2} - \frac{b}{5}z^2 - \frac{c}{7}z^4 + O(z^6). \end{aligned}$$

Similarly

$$\wp_{\bar\Lambda}(z) = \frac{1}{z^2} - \frac{b_1}{5}z^2 - \frac{c_1}{7}z^4 + O(z^6).$$

Thus we can determine b_1, c_1 by looking at the expansion of $\wp_{\bar\Lambda}(z)$ around $z = 0$. From above,

$$\begin{aligned} \wp_{\bar\Lambda}(z) &= \wp(z) + \wp(z + \omega_1/2) - \alpha \\ &= \frac{1}{z^2} - \frac{b}{5}z^2 - \frac{c}{7}z^4 + (3\alpha^2 + b)z^2 + \alpha(3\alpha^2 + b)z^4 + O(z^6). \end{aligned}$$

We conclude that

$$\begin{aligned} \bar b &= -15\alpha^2 - 4b, \\ \bar c &= -21\alpha^3 - 7\alpha b + c \\ &= -28\alpha^3 + 8c, \end{aligned}$$

since

$$\alpha^3 + b\alpha + c = 0.$$

The relation between $\bar\Lambda$ and $\frac{1}{2}\Lambda$ is exactly the same as that between Λ and $\bar\Lambda$, except that $\omega_1/2$ is replaced by $\omega_2/2$. More precisely,

$$\alpha = \wp(\omega_1/2)$$

is replaced by

$$\begin{aligned}
\bar{\alpha} &= \wp_{\bar{\Lambda}}(\omega_2/2) \\
&= \wp(\omega_2/2) + \wp(\omega_1/2 + \omega_2/2) - \wp(\omega_1/2) \\
&= \beta + \gamma - \alpha \\
&= -2\alpha;
\end{aligned}$$

for $\alpha + \beta + \gamma = 0$, since α, β, γ are the roots of $x^3 + bx + c$.

It follows that the formula for $\bar{\Theta}$, can be derived from that for Θ by substituting $\bar{b}, \bar{c}, -2\alpha$ for b, c, α, respectively (corresponding to the homomorphism $\theta_2 : \mathbb{C}/\bar{\Lambda} \to \mathbb{C}/\frac{1}{2}\Lambda$) and then dividing the x and y-coordindates by 4 and 8, respectively (corresponding to the homomorphism $\theta_3 : \mathbb{C}/\frac{1}{2}\Lambda \to \mathbb{C}/\Lambda$). Thus

$$\begin{aligned}
\bar{\Theta}(\bar{x}, \bar{y}) &= \left(\frac{1}{4} \cdot \frac{\bar{x}^2 - (-2\alpha)\bar{x} + 3(-2\alpha)^2 + \bar{b}}{\bar{x} - (-2\alpha)}, \frac{1}{8} \cdot \frac{\bar{x}^2 - 2(-2\alpha)\bar{x} - 2(-2\alpha)^2 - \bar{b}}{(\bar{x} - (-2\alpha))^2}\bar{y}\right) \\
&= \left(\frac{1}{4} \cdot \frac{\bar{x}^2 + 2\alpha\bar{x} - 3\alpha^2 - 4b}{\bar{x} + 2\alpha}, \frac{1}{8} \cdot \frac{\bar{x}^2 + 4\alpha\bar{x} + 7\alpha^2 + 4b}{(\bar{x} + 2\alpha)^2}\bar{y}\right).
\end{aligned}$$

if $\bar{x} \neq -2\alpha$, while

$$\bar{\Theta}(-2\alpha, 0) = O.$$

We summarise our results in the following Proposition.

Proposition 10.2 *Suppose*

$$\mathcal{E}(\mathbb{C}) : y^2 = x^3 + bx + c$$

is the elliptic curve associated to a lattice Λ*; and suppose* α *is a root of* $x^3 + bx + c$*. Let* $\tilde{\mathcal{E}}$ *be the elliptic curve*

$$\tilde{\mathcal{E}}(\mathbb{C}) : y^2 = x^3 + (-15\alpha^2 - 4b)x + (-28\alpha^2 + 8c).$$

Then the homomorphism

$$\Phi : \mathcal{E} \to \mathcal{E}$$

under which

$$P \mapsto 2P$$

can be expressed as the product of 2 homomorphisms

$$\Phi = \bar{\Theta}\Theta$$

where

$$\Theta : \mathcal{E} \to \tilde{\mathcal{E}}, \ \bar{\Theta} : \tilde{\mathcal{E}} \to \mathcal{E}$$

are the maps

$$\Theta(x, y) = \left(\frac{x^2 - \alpha x + 3\alpha^2 + b}{x - \alpha}, \frac{x^2 - 2\alpha x - 2\alpha^2 - b}{(x - \alpha)^2}y\right)$$

if $(x, y) \neq (\alpha, 0)$*, while* $\Theta(\alpha, 0) = \Theta(O) = O$*; and*

$$\bar{\Theta}(\bar{x}, \bar{y}) = \left(\frac{1}{4} \cdot \frac{(\bar{x}^2 + 2\alpha\bar{x} - 3\alpha^2 - 4b}{\bar{x} + 2\alpha}, \frac{1}{8} \cdot \frac{\bar{x}^2 + 4\alpha\bar{x} + 7\alpha^2 + 4b}{(\bar{x} + 2\alpha)^2}\bar{y}\right)$$

if $(\bar{x}, \bar{y}) \neq (-2\alpha, 0)$*, while* $\bar{\Theta}(-2\alpha, 0) = \bar{\Theta}(O) = O$*.*

10.3 Tying a neater package

Our formulae become much simpler if we work with elliptic curves in 'constant-free' format

$$\mathcal{E} : y^2 = x^3 + ax^2 + bx.$$

It is not difficult to see why. Our construction starts with an elliptic curve $\mathcal{E}$ *together with a point of order 2 on* $\mathcal{E}$. By taking $\mathcal{E}$ in constant-free form we have a 'built-in' point of order 2, namely $(0, 0)$. Thus we have only 2 constants, a and b, to deal with rather than b, c and α.

To avoid confusion, let us — for the time being — 'dot' the coefficients and variables in the constant-free model:

$$\dot{\mathcal{E}} : \dot{y}^2 = \dot{x}^3 + \dot{a}\dot{x}^2 + \dot{b}\dot{x}.$$

The coordinate-change $x = \dot{x} + \dot{a}/3$ brings this to our earlier 'x^2-free' format

$$\begin{aligned}
y^2 &= (x - \dot{a}/3)^3 + \dot{a}(x - \dot{a}/3)^2 + \dot{b}(x - \dot{a}/3) \\
&= x^3 + (-\dot{a}^2/3 + \dot{b})x + (2\dot{a}^3/27 - \dot{a}\dot{b}/3).
\end{aligned}$$

Thus

$$b = -\dot{a}^2/3 + \dot{b}, \ c = 2\dot{a}^3/27 - \dot{a}\dot{b}/3,$$

and

$$\alpha = \dot{a}/3,$$

since $(0, 0) \mapsto (\dot{a}/3, 0)$.

Hence the associated curve $\bar{\mathcal{E}}$ (in x^2-free format) has coefficients

$$\begin{aligned}
\bar{b} &= -15\alpha^2 - 4b \\
&= -\dot{a}^2/3 - 4\dot{b}, \\
\bar{c} &= -28\alpha^3 + 8c \\
&= -4\dot{a}^3/9 - 8\dot{a}\dot{b}/3.
\end{aligned}$$

We want to transform $\bar{\mathcal{E}}$ into constant-free format. At first sight there might seem some ambiguity in this, since it involves choosing a point of order 2 on $\bar{\mathcal{E}}$. However, we know the point we want: $(\alpha_1, 0) = (-2\alpha, 0) = (-2\dot{a}/3, 0)$. Our transformation must bring this to $(0,0)$, and is therefore

$$\dot{x} = x + 2\alpha = x + 2\dot{a}/3.$$

Thus our new curve $\dot{\bar{\mathcal{E}}}$ has equation

$$\begin{aligned} \dot{y}^3 &= (\dot{x} - 2\dot{a}/3)^3 + b_1(\dot{x} - 2\dot{a}/3) + c_1 \\ &= \dot{x}^3 + \dot{a}_1\dot{x}^2 + \dot{b}_1\dot{x}, \end{aligned}$$

where

$$\begin{aligned} \dot{a}_1 &= -2\dot{a}, \\ \dot{b}_1 &= 4\dot{a}^2/3 + \dot{b}_1 \\ &= \dot{a}^2 - 4\dot{b}, \end{aligned}$$

which is pleasingly simple!

It remains to express Θ and $\bar{\Theta}$ in the new system. We have

$$\dot{\Theta}(\dot{x}, \dot{y}) = (\tilde{x}, \tilde{y}),$$

where

$$\begin{aligned} \tilde{x} &= \frac{(\dot{x} + \dot{a}/3)^2 - \dot{a}(\dot{x} + \dot{a}/3)/3 + 3(\dot{a}/3)^2 - \dot{a}^2/3 + \dot{b}}{(\dot{x} + \dot{a}/3 - \dot{a}/3)^2} + 2\dot{a}/3 \\ &= \frac{\dot{x}^2 + \dot{a}\dot{x} + \dot{b}}{\dot{x}^2}, \\ \tilde{y} &= \frac{(\dot{x} + \dot{a}/3)^2 - 2\dot{a}/3(\dot{x} + \dot{a}/3) - 2(\dot{a}/3)^2 + \dot{a}^2/3 - \dot{b}}{(\dot{x} + \dot{a}/3 - \dot{a}/3)^2}\dot{y} \\ &= \frac{\dot{x}^2 - \dot{b}}{\dot{x}^2}\dot{y}. \end{aligned}$$

We derive $\dot{\bar{\Theta}}$ from this by substituting $\dot{\bar{b}} = \dot{a}^2 - 4\dot{b}$ for $\dot{b}$, and dividing the x- and y-coordinates by 4 and 8, respectively:

$$\dot{\bar{\Theta}}(\dot{x}, \dot{y}) = \left(\frac{\dot{y}^2}{4\dot{x}^2}, \frac{\dot{x}^2 - \dot{a}^2 + 4\dot{b}}{8\dot{x}^2}\dot{y}\right).$$

We summarise our conclusions in the following Definition and Proposition, where we now drop the dots.

Definition 10.1 *To each elliptic curve*

$$\mathcal{E} : y^2 = x^3 + ax^2 + bx.$$

we associated the elliptic curve

$$\tilde{\mathcal{E}} : y^2 = x^3 + \bar{a}x^2 + +\bar{b}x,$$

where

$$\bar{a} = -2a, \; \bar{b} = a^2 - 4b.$$

Theorem 10.1 *Suppose $\mathcal{E}(K)$ is the elliptic curve*

$$\mathcal{E} : y^2 = x^3 + ax^2 + bx$$

over the field K. Let $\bar{\mathcal{E}}(K)$ be the associated elliptic curve, and let the maps

$$\Theta : \mathcal{E} \to \bar{\mathcal{E}}, \; \bar{\Theta} : \bar{\mathcal{E}} \to \mathcal{E}$$

be defined by

$$\Theta(x, y) = \left(\frac{y^2}{x^2}, \frac{x^2 - b}{x^2}y\right)$$

if $x \neq 0$, while $\Theta(O) = \Theta(T) = 0$ for $T = (0,0)$,

$$\bar{\Theta})(x, y) = \left(\frac{\bar{y}^2}{4\bar{x}^2}, \frac{\bar{x}^2 - \bar{b}}{8\bar{x}^2}\bar{y}\right)$$

if $\bar{x} \neq 0$, while $\bar{\Theta}(O) = \bar{\Theta}(\bar{T}) = 0$ for $\bar{T} = (0,0)$. Then $\Theta, \bar{\Theta}$ are homomorphisms; and

$$\bar{\Theta}\Theta : \mathcal{E} \to \mathcal{E}, \; \Theta\bar{\Theta} : \bar{\mathcal{E}} \to \bar{\mathcal{E}}$$

are the doubling maps $P \mapsto 2P$ on $\mathcal{E}$ and $\bar{\mathcal{E}}$.

Proof ▶ Although we established this result on the assumption that $K = \mathbb{C}$, it is readily verified that each part of the result (eg the statement that $\Theta(P + Q) = \Theta(P) + \Theta(Q)$) can be expressed as a number of polynomial identities with integral coefficients, which must remain valid over any field. ◀

At this point we can forget how *the associated elliptic curve $\bar{\mathcal{E}}$ and the homomorphisms $\Theta, \bar{\Theta}$ arose; all we need to know is that the maps given by the formulae above are indeed homomorphisms, and that the doubling map Φ on $\mathcal{E}$ factorises into $\bar{\Theta}\Theta$.*

10.4 Divide and rule

Recall that our main aim is to show that if $K = \mathbb{Q}$ then $[\mathcal{E} : 2\mathcal{E}]$ is finite. The splitting of the doubling map allows us to divide this task.

Proposition 10.3 *Suppose $\phi : A \to B$ is a homomorphism of abelian groups, and $S \subset A$ is a subgroup of finite index. Then*

$$[\phi A : \phi S] \le [A : S].$$

Proof ▶ Consider the composition

$$\phi_S : A \to B \to B/\phi S.$$

Evidently

$$\operatorname{im} \phi_S = \phi A/\phi S,$$

while

$$\ker \phi_S \supset S.$$

By the first isomorphism theorem,

$$\phi A/\phi S \cong A/\ker \phi_S.$$

Hence

$$[\phi A : \phi S] = [A : \ker \phi_S] \le [A : S].$$

◀

Proposition 10.4 $[\mathcal{E} : 2\mathcal{E}]$ *and* $[\bar{\mathcal{E}} : 2\bar{\mathcal{E}}]$ *are both finite if and only if* $[\bar{\mathcal{E}} : \operatorname{im} \Theta]$ *and* $[\mathcal{E} : \operatorname{im} \bar{\Theta}]$ *are both finite.*

Proof ▶ We have

$$\begin{aligned}
[\mathcal{E} : 2\mathcal{E}] &= [\mathcal{E} : \bar{\Theta}\Theta\mathcal{E}] \\
&= [\mathcal{E} : \bar{\Theta}\bar{\mathcal{E}}][\bar{\Theta}\bar{\mathcal{E}} : \bar{\Theta}\Theta\mathcal{E}] \\
&\le [\mathcal{E} : \bar{\Theta}\bar{\mathcal{E}}][\bar{\mathcal{E}} : \Theta\mathcal{E}],
\end{aligned}$$

by Proposition 10.3 ◀

10.5 Characterisation of the image

Proposition 10.5 *If $\bar{P} = (\bar{x}, \bar{y}) \in \bar{\mathcal{E}}$ with $\bar{x} \ne 0$ then*

$$\bar{P} \in \operatorname{im} \Theta \iff \bar{x} \in K^2.$$

Similarly if $P = (x, y) \in \mathcal{E}$ with $x \ne 0$ then

$$P \in \operatorname{im} \bar{\Theta} \iff x \in K^2.$$

Proof ▶ Suppose $\bar{P} = \Theta(P)$, where $P = (x, y)$ Then

$$\bar{x} = \frac{y^2}{x^2} \in K^2.$$

Conversely, suppose $(\bar{x}, \bar{y}) \in \bar{\mathcal{E}}$; and suppose

$$\bar{x} = w^2,$$

where $w \in K$. We have to show that there is a point $P = (x, y) \in \mathcal{E}(K)$ with

$$\frac{y^2}{x^2} = w^2.$$

We may suppose that

$$y = wx,$$

on taking $-P$ if $y = -wx$.

Substituting $y = wx$ in the equation for $\mathcal{E}$,

$$w^2x^2 = x^3 + ax^2 + bx = 0.$$

One solution is $x = 0$; the other two are given by

$$x^2 + (a - w^2)x + b = 0.$$

This will have a solution in K if and only if

$$(a - w^2)^2 - 4b \in K^2,$$

ie

$$w^4 - 2aw^2 + (a^2 - 4b) \in K^2,$$

ie

$$\bar{x}^2 + \bar{a}\bar{x} + \bar{b} \in K^2.$$

But since $(\bar{x}, \bar{y}) \in \bar{\mathcal{E}}$,

$$\bar{y}^2 = \bar{x}(\bar{x}^2 + \bar{a}\bar{x} + \bar{b}).$$

By hypothesis, $\bar{x} \in K^2$. Hence

$$\bar{x}^2 + \bar{a}\bar{x} + \bar{b} \in K^2,$$

which as we saw is the condition for $(\bar{x}, \bar{y}) \in \operatorname{im}\Theta$.

The proof of the corresponding result for $\bar{\Theta}$ is identical, the factor $1/4$ in the x-coordinate of $\bar{\Theta}(\bar{x}, \bar{y})$ making no difference, since we are working modulo squares. ◄

10.6 The associated homomorphism

Proposition 10.6 *The map*

$$\chi : \mathcal{E}(K) \to K^\times / K^{\times 2}$$

under which

$$P \mapsto \begin{cases} x \bmod K^{\times 2} & \text{if } P = (x, y) \text{ with } x \neq 0 \\ b \bmod K^{\times 2} & \text{if } P = T = (0, 0) \\ 1 \bmod K^{\times 2} & \text{if } P = O \end{cases}$$

is a homomorphism.

Proof ► Trivially,

$$\chi(-P) = \chi(P) = 1/\chi(P),$$

since $x = 1/x$ for all $x \in K^\times/K^{\times 2}$ (ie all elements are of order 1 or 2).

Now suppose

$$P + Q + R = 0,$$

ie P, Q, R are collinear. We have to show that

$$\chi(P)\chi(Q)\chi(R) = 1.$$

If one of P, Q, R is O, say $P = O$, this reduces to the result just proved:

$$Q + R = 0 \Longrightarrow \chi(Q)\chi(R) = 1.$$

Suppose none of the points is O. Let the line PQR be $y = mx + d$. This line meets $\mathcal{E}$ where

$$(mx + d)^2 = x^3 + ax^2 + bx.$$

The roots of this are the x-coordinates of P, Q, R, say x_1, x_2, x_3. Thus

$$x_1 x_2 x_3 = d^2.$$

If none of x_1, x_2, x_3 is zero, then

$$\chi(P)\chi(Q)\chi(R) = x_1 x_2 x_3 \equiv 1 \bmod K^{\times 2},$$

as required.

Finally, suppose one of x_1, x_2, x_3 is 0, say $x_1 = 0$, ie $P = T = (0, 0)$. Then $d = 0$, and the remaining two points satisfy the quadratic

$$m^2 x = x^2 + ad + b = 0.$$

Thus

$$x_2 x_3 = b.$$

Now $\chi(T) = b$ (by what may have seemed an arbitrary definition, but whose purpose is now apparent); so

$$\chi(P)\chi(Q)\chi(R) = b x_2 x_3 = b^2 \equiv 1 \bmod K^{\times 2}.$$

Thus in all cases

$$P + Q + R = 0 \Longrightarrow \chi(P)\chi(Q)\chi(R) = 1.$$

Hence χ is a homomorphism. ◄

Now we can re-state Proposition 10.5 as

Proposition 10.7 *We have*

$$\operatorname{im}\Theta = \ker\bar{\chi}, \ \operatorname{im}\bar{\Theta} = \ker\chi.$$

Equivalently, the two sequences

$$\mathcal{E} \xrightarrow{\Theta} \bar{\mathcal{E}} \xrightarrow{\bar{\chi}} K^\times/K^{\times 2}, \quad \bar{\mathcal{E}} \xrightarrow{\bar{\Theta}} \mathcal{E} \xrightarrow{\chi} K^\times/K^{\times 2}$$

are exact.

Proposition 10.8 $[\mathcal{E} : 2\mathcal{E}]$ *and* $\bar{\mathcal{E}} : 2\bar{\mathcal{E}}]$ *are both finite if and only if* $\operatorname{im}\chi$ *and* $\operatorname{im}\bar{\chi}$ *are both finite.*

Proof ► This follows at once from Proposition 10.4, since

$$\mathcal{E}/\operatorname{im}\bar{\Theta} \cong \operatorname{im}\chi, \quad \bar{\mathcal{E}}/\operatorname{im}\Theta \cong \operatorname{im}\bar{\chi},$$

◄

10.7 The rational case

So far we have been working over a general field K. Now let us turn to the rational case $K = \mathbb{Q}$. Note that since $T = (0,0) \in \mathcal{E}$, we are assuming that our elliptic curve contains a rational point of order 2.

Proposition 10.9 *Let $\mathcal{E}$ be the elliptic curve*

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx \qquad (a, b \in \mathbb{Z});$$

and let

$$\chi : \mathcal{E} \to \mathbb{Q}^\times / \mathbb{Q}^{\times 2}$$

be the associated homomorphism under which

$$P = (x, y) \mapsto x \bmod \mathbb{Q}^{\times 2}.$$

Then each element of $\operatorname{im} \chi$ *is of the form*

$$b_1 \bmod \mathbb{Q}^{\times 2}$$

where $b_1 \mid b$.

Proof ► Suppose $P = (x, y) \in \mathcal{E}$. We know that x, y can be expressed in the form

$$x = \frac{m}{e^2},\ y = \frac{n}{e^3},$$

where $e, m, n \in \mathbb{Z}$ and $\gcd(m, e) = \gcd(n, e) = 1$.

From the equation of the curve,

$$n^2 = m(m^2 + ae^2 m + be^4).$$

Let

$$b_1 = \gcd(m, m^2 + ae^2 m + be^4)$$

Then

$$\begin{aligned} b_1 &= \gcd(m, be^4) \\ &= \gcd(m, b), \end{aligned}$$

since $\gcd(m, e) = 1$. In particular, $b_1 \mid b$. Let

$$b = b_1 b_2,\ m = b_1 m_1,$$

where we choose the sign of b_1 so that $m_1 \geq 0$. Then

$$n^2 = b_1^2 m_1 (b_1 m_1^2 + ae^2 m_1 + b_2 e^4).$$

Hence $b_1^2 \mid n^2$, and so $b_1 \mid n$, say

$$n = b_1 n_1.$$

Thus

$$n_1^2 = m_1 (b_1 m_1^2 + ae^2 m_1 + b_2 e^4).$$

The two factors on the right are co-prime, since we took out their common factor. Hence

$$m_1 = U^2,\ b_1 m_1^2 + ae^2 m_1 + b_2 e^4 = V^2.$$

For future reference we note that this implies

$$b_1 U^4 + ae^2 U^2 + b_2 e^4 = V^2,$$

with $e > 0,\ \gcd(U, V) = 1$.

But for our present purpose we simply need the fact that

$$\begin{aligned} x &= \frac{b_1 m_1}{e^2} \\ &= b_1 \frac{U^2}{e^2} \\ &\equiv b_1 \bmod \mathbb{Q}^{\times 2}. \end{aligned}$$

◄

Corollary 19 *Each element of* $\operatorname{im} \bar{\chi}$ *is of the form*

$$\overline{b_1} \bmod \mathbb{Q}^{\times 2}$$

where $\overline{b_1} \mid \bar{b}$.

Theorem 10.2 *The group $\mathcal{E}/2\mathcal{E}$ is finite.*

Proof ► By Proposition 10.4,

$$\begin{aligned} [\mathcal{E} : 2\mathcal{E}] &\leq [\mathcal{E} : \operatorname{im} \bar{\Theta}][\bar{\mathcal{E}} : \operatorname{im} \Theta] \\ &= \| \operatorname{im} \chi \| \cdot \| \operatorname{im} \bar{\chi} \|. \end{aligned}$$

But these two images are finite, by Proposition 10.9 and its Corollary. ◄

10.8 Determining the rank of $\mathcal{E}$

We know that

$$\mathcal{E} = F \oplus \mathbb{Z}^r,$$

where F is the torsion subgroup of $\mathcal{E}$, and r is its rank. It follows that

$$\mathcal{E}/2\mathcal{E} = F/2F \oplus (\mathbb{Z}/(2))^r.$$

Note that if A is an abelian group then $A/2A$ is of exponent 2, ie $2\bar{a} = 0$ for all $\bar{a} \in A/2A$. Thus if A is finitely-generated, it follows from the Structure Theorem that

$$A/2A = (\mathbb{Z}/(2))^d$$

for some d. (Alternatively, $A/2A$ can be regarded as a vector space over the finite field $\mathcal{F}_2 = \{0, 1\}$; and d is the dimension of this vector space.)

In our case,

$$2^r = \frac{[\mathcal{E} : 2\mathcal{E}]}{[F : 2F]}.$$

It is easy to determine $[F : 2F]$; so computation of the rank r reduces to the determination of $[\mathcal{E} : 2\mathcal{E}]$. For this we need to sharpen a little our earlier proof that $[\mathcal{E} : 2\mathcal{E}]$ is finite.

But first let us consider the torsion subgroup F. Suppose A is a finite abelian group. By the Structure Theorem,

$$\begin{aligned} A &= \mathbb{Z}/(p_1^{e_1} \oplus \cdots \oplus \mathbb{Z}/(p_r^{e_r}) \\ &= C_1 \oplus \cdots \oplus C_r, \end{aligned}$$

say, where $C_i = \mathbb{Z}/(p_i^{e_i})$. Thus

$$A/2A = C_1/2C_1 \oplus \cdots \oplus C_r/2C_r.$$

Proposition 10.10 *Suppose $A = \mathbb{Z}/(p^e)$. Then*

$$[A : 2A] = \begin{cases} \mathbb{Z}/(2) & \text{if } p = 2 \\ 0 & \text{if } p \neq 2. \end{cases}$$

Proof ► Consider the map $\phi : A \to A$ under which

$$a \mapsto 2a.$$

Then

$$\ker \phi = \{a \in A : 2a = 0\}.$$

If $p \neq 2$ then A has no elements of order 2, by Lagrange's Theorem. Hence $\ker \phi = 0$, and so

$$2A = A,$$

ie every element $a \in A$ is of the form $a = 2b$ for some $b \in A$.

On the other hand, if $p = 2$ then $\mathbb{Z}/(2^e)$ has just one element of order 2, namely $2^{e-1} \bmod 2^e$. Thus $\| \ker \phi \| = 2$; and so

$$[A : 2A] = 2.$$

◄

Corollary *If*

$$A = \mathbb{Z}/(p_1^{e_1}) \oplus \cdots \oplus \mathbb{Z}/(p_r^{e_r})$$

then

$$[A : 2A] = 2^d,$$

where d is the number of factors with $p_i = 2$.

Corollary *If A is a finite abelian group with*

$$[A : 2A] = 2^d,$$

then the number of elements of order 2 in A is $2^e - 1$.

Proof ► As we saw above, the factor $\mathbb{Z}/(p^e)$ contains just one element of order 2 if $p = 2$ and none otherwise. But the element

$$a = a_1 \oplus \cdots \oplus a_r$$

is of order 1 or 2 if and only if that is true of each a_i. Thus the number of such elements is 2^d by Corollary 1; and the result follows on subtracting the one element of order 1. ◄

We apply this result to our elliptic curve $\mathcal{E}$. We know that $\mathcal{E}$ has at least one point of order 2, namely $T = (0, 0)$. We know too that if it has more than one point of order 2 then it must have just three.

Proposition 10.11 *Suppose F is the torsion subgroup of*

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx.$$

Then

$$[F : 2F] = \begin{cases} 4 & \text{if } \bar{b} \in \mathbb{Q}^2 \\ 2 & \text{if } \bar{b} \notin \mathbb{Q}^2 \end{cases}$$

Proof ► $P \in \mathcal{E}$ is of order 2 if $P = (\alpha, 0)$, where α is a root of

$$x^3 + ax^2 + bx = 0.$$

One root is $\alpha = 0$; the other two are the roots of the quadratic

$$x^2 + ax + b = 0.$$

This has rational roots if and only if

$$a^2 - 4b = \bar{b} \in \mathbb{Q}^2.$$

Thus $\mathcal{E}$ has 3 or 1 points of order 2, and so $[F : 2F] = 4$ or 2, according as $\bar{b}$ is or is not a perfect square. ◄

We proved that $[\mathcal{E} : 2\mathcal{E}]$ is finite by showing that

$$\begin{aligned}[\mathcal{E} : 2\mathcal{E}] &= [\mathcal{E} : \bar{\Theta}\Theta\mathcal{E}] \\ &= [\mathcal{E} : \bar{\Theta}\bar{\mathcal{E}}][\bar{\Theta}\bar{\mathcal{E}} : \bar{\Theta}(\Theta\mathcal{E})] \\ &\le [\mathcal{E} : \bar{\Theta}\bar{\mathcal{E}}][\bar{\mathcal{E}} : \Theta\mathcal{E}].\end{aligned}$$

But now we need a slightly more precise result in place of Proposition 10.3.

Proposition 10.12 *Suppose $\phi : A \to B$ is a homomorphism of abelian groups, and $S \subset A$ is a subgroup of finite index. Then*

$$[A : S] = [\phi A : \phi S][\ker \phi : \ker \phi \cap S].$$

Proof ► With the same notation as in the earlier proof,

$$\ker \phi_S = S + \ker \phi.$$

For

$$\begin{aligned}\phi_S a = 0 &\Longrightarrow \phi a = \phi s \\ &\Longrightarrow a = s + k,\end{aligned}$$

where $k \in \ker \phi$; while conversely $\phi(s + k) = \phi s \in \phi S$. Thus

$$[A : S] = [\phi A : \phi S][\ker \phi + S : S].$$

But by the Second Isomorphism Theorem, if $S, T \subset A$ then

$$(S + T)/S \cong T/(S \cap T).$$

In particular,

$$[\ker \phi + S : S] = [\ker \phi : \ker \phi \cap S].$$

◄

Corollary 20 *We have*

$$[\mathcal{E}\ 2\mathcal{E}] = \frac{[\mathcal{E} : \operatorname{im} \bar{\Theta}][\bar{\mathcal{E}} : \operatorname{im} \Theta]}{[\ker \bar{\Theta} \cap \operatorname{im} \Theta]}$$

Proof ► This follows on applying the Proposition with $A = \bar{\mathcal{E}},\ B = \mathcal{E},\ \phi = \bar{\Theta},\ S = \operatorname{im} \Theta$. ◄

The subgroups $\ker \Theta$ and $\ker \bar{\Theta}$ are (almost) trivial.

Proposition 10.13 *We have*

$$\ker \Theta = \{O, T\}, \quad \ker \bar{\Theta} = \{O, \bar{T}\},$$

where $T = (0, 0) \in \mathcal{E},\ \bar{T} = (0, 0) \in \bar{\mathcal{E}}$.

Proof ► This follows at once from the definitions of $\Theta, \bar{\Theta}$, since $\Theta(x, y)$ is finite (ie $Z \ne 0$) if $x \ne 0$; and $\bar{\Theta}(\bar{x}, \bar{y})$ is finite if $\bar{x} \ne 0$. ◄

Proposition 10.14 *We have*

$$[\mathcal{E} : 2\mathcal{E}] = \frac{[\mathcal{E} : \operatorname{im} \bar{\Theta}][\bar{\mathcal{E}} : \operatorname{im} \Theta]}{d}$$

where

$$d = \begin{cases} 2 & \text{if } \bar{b} \in \mathbb{Q}^2, \\ 1 & \text{if } \bar{b} \notin \mathbb{Q}^2 \end{cases}$$

Proof ► After Proposition 10.13 we simply have to determine whether or not

$$\bar{T} \in \operatorname{im} \Theta.$$

Suppose $T = \Theta(P)$, where $P = (x, y)$. Then $y = 0$ from the definition of Θ. On the other hand $P \ne T$, since $\Theta(T) = O$, by definition.

In other words, $\Theta(P) = T$ if and only if $P \in \mathcal{E}$ is a point of order 2 other than T. But, as we saw in the proof of Proposition 10.11, there are two such points if $\bar{b} \in \mathbb{Q}^2$, and no such points otherwise.

Thus

$$\| \ker \bar{\Theta} \cap \operatorname{im} \Theta \| = \begin{cases} 2 & \text{if } \bar{b} \in \mathbb{Q}^2, \\ 1 & \text{if } \bar{b} \notin \mathbb{Q}^2; \end{cases}$$

and the result follows. ◄

Theorem 10.3 *If the rank of the elliptic curve*

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx$$

is r then

$$2^r = \frac{\| \operatorname{im} \chi \| \cdot \| \operatorname{im} \bar{\chi} \|}{4}.$$

Proof ► If

$$\mathcal{E} = F \oplus \mathbb{Z}^r$$

then as we saw

$$2^r = \frac{[\mathcal{E} : 2\mathcal{E}]}{[F : 2F]}$$

The result now follows at once from Propositions 10.14 and 10.11. ◄

10.9 An example

Consider the elliptic curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + x.$$

The associated curve is

$$\bar{\mathcal{E}}(\mathbb{Q}) : y^2 = x^3 - 4x.$$

Thus

$$b = 1, \ \bar{b} = -4.$$

If the rank of $\mathcal{E}$ is r then

$$2^r = \frac{\| \operatorname{im} \chi \| \cdot \| \operatorname{im} \bar{\chi} \|}{4}$$

by Theorem 10.3. We have to determine $\| \operatorname{im} \chi \|, \| \operatorname{im} \bar{\chi} \|$.

Let us consider

$$\chi : \mathcal{E} \to \mathbb{Q}^\times / \mathbb{Q}^{\times 2}$$

first. We know that the elements of $\operatorname{im} \chi$ are of the form

$$b_1 \bmod \mathbb{Q}^{\times 2},$$

where $b_1 \mid b$. In this case $b = 1$, and so

$$b_1 = \pm 1.$$

Certainly $1 = \chi(O) \in \operatorname{im} \chi$. We have to determine if $-1 \in \operatorname{im} \chi$.

We saw in the proof of Proposition 10.9 that if this is so then we can find e, U, V with $e \geq 1, \gcd(U, V) = 1$ satisfying

$$b_1 U^4 + b_2 e^4 = V^2,$$

ie

$$-U^4 - e^4 = V^2,$$

which is clearly impossible. Hence $-1 \notin \operatorname{im} \chi$, and so

$$\operatorname{im} \chi = \{1\}.$$

Turning to $\bar{\chi}$, we have $\bar{b} = -4$, and so $b_1 = \pm 1, \pm 2$. (We can omit $b_1 = \pm 4$, since we are working modulo squares.) We know that $1 \in \operatorname{im} \bar{\chi}$. Also

$$\chi(\bar{T}) = \bar{b} = -4 \equiv -1,$$

where $\bar{T} = (0, 0)$. Thus $-1 \in \operatorname{im} \bar{\chi}$.

It remains to determine if $b_1 = \pm 2 \in \operatorname{im} \bar{\chi}$. (Note that if one is in the image then so is the other, since $\operatorname{im} \bar{\chi}$ is a subgroup containing -1.) For $b_1 = 2, b_2 = -2$, we have to solve the equation

$$2U^4 - 2e^4 = V^2.$$

This has the trivial solution $(e, U, V) = (1, 1, 0)$ (corresponding to the point $P = (2, 0) \in \bar{\mathcal{E}}$).

We conclude that

$$\operatorname{im} \bar{\chi} = \{\pm 1, \pm 2\}.$$

(Note that once we knew that $\operatorname{im} \chi = \{1\}$, it followed from Theorem 10.3 that $\| \operatorname{im} \bar{\chi} \| \geq 4$; so in fact it was clear that $\operatorname{im} \bar{\chi} = \{\pm 1, \pm 2\}$.)

Hence

$$2^r = \frac{1 \cdot 4}{4} = 1,$$

ie $\mathcal{E}$ is of rank 0, that is, $\mathcal{E}(\mathbb{Q})$ is finite.

Now we can find $\mathcal{E} = F$ easily, by the Nagell-Lutz Theorem. We have

$$D = -4.$$

Hence $y = 0, \pm 1, \pm 2$. But the equations

$$x^3 + x - 1 = 0, \quad x^3 + x - 4$$

have no solutions. Hence the only rational points on $\mathcal{E}$ are the 3 points of order 2,

$$\mathcal{E} = \{O, (0, 0), (2, 0), (-2, 0)\}.$$

10.10 Another example

If b is not a perfect square, then $1 = \Theta(O)$, $b = \Theta(T)$ are distinct elements of $\operatorname{im}\chi$. Similarly, if $\bar{b}$ is not a perfect square, then $1, \bar{b}$ are distinct elements of $\operatorname{im}\bar{\chi}$.

Thus if neither b nor $\bar{b}$ is a perfect square then these elements alone contribute 4 to $\|\operatorname{im}\chi\| \cdot \|\operatorname{im}\bar{\chi}\|$; so by Theorem 10.3 any further element in either of these images ensures that the rank is ≥ 1.

Consider the elliptic curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + 3x.$$

The associated curve is

$$\bar{\mathcal{E}}(\mathbb{Q}) : \bar{y}^2 = \bar{x}^3 - 12\bar{x}.$$

Thus

$$b = 3,\ \bar{b} = -12.$$

We know that $3 = \chi(T) \in \operatorname{im}\chi$. On the other hand $-1, -3 \notin \operatorname{im}\chi$, since

$$b_1 U^4 + b_2 e^4 < 0$$

in these cases. Thus

$$\operatorname{im}\chi = \{1, 3\}.$$

Similarly $-3 \equiv -12 \in \operatorname{im}\bar{\chi}$. We have to determine which other factors b_1 of 12 are in $\operatorname{im}\bar{\chi}$ — remembering that since this is a 2-group it contains either 2, 4 or 8 elements. The candidates are: $-1, \pm 2, 3, \pm 6$.

If $-1 \in \operatorname{im}\bar{\chi}$ then the equation

$$-U^4 + 12e^4 = V^2$$

has a solution with $e \geq 1$, $\gcd(U, V) = 1$. U must be odd, since otherwise U, V are both even. But then

$$-U^4 \equiv 1 \bmod 4,$$

and so

$$-U^4 + 12e^4 \equiv -1 \bmod 4.$$

Since -1 is not a square $\bmod 4$, the equation has no solution, and $-1 \notin \operatorname{im}\bar{\chi}$. Thus $\|\operatorname{im}\bar{\chi}\| = 2$ or 4.

The equation for $b_1 = -2$, $b_2 = 6$ is

$$-2U^4 + 6e^4 = V^2,$$

which has the obvious solution $(e, U, V) = (1, 1, 2)$. Thus $-2 \in \operatorname{im}\bar{\chi}$. It follows that

$$\operatorname{im}\bar{\chi} = \{1, -2, -3, 6\}.$$

In particular $\|\operatorname{im}\bar{\chi}\| = 4$, and so

$$\operatorname{rank}(\mathcal{E}) = 1.$$

We can determine the torsion subgroup $F \subset \mathcal{E}$ by the Nagell-Lutz theorem, in the usual way. The discriminant of $x^3 + 3x$ is $-4 \cdot 3^2 = -36$ Thus if $P = (x, y) \in \mathcal{E}$ is of finite order, then $x, y \in \mathbb{Z}$ and $y = 0$ or $y^2 \mid 36$. Hence

$$y = 0, \pm 1, \pm 2, \pm 3, \pm 6.$$

Also

$$y^2 = x^3 + 3x = x(x^2 + 3) \Longrightarrow x \geq 0.$$

It is readily verified that the only possible points of finite order are: O, $(0,0)$, $(1, \pm 2)$, $(3, \pm 6)$.

We can use the 'factors of double' to simplify computation of $2P$. (Alternatively, we could find where the tangent at P meets the curve again, in the usual way.) Let $S = (1, 2)$. Then

$$\Theta(S) = \left(\frac{2^2}{1^2}, \frac{1^2 - 3}{1^2} \cdot 2\right) = (4, -4),$$

and so

$$2S = \bar{\Theta}\Theta(S) = \bar{\Theta}(4, -4) = \left(\frac{1}{4} \cdot \frac{16}{16}, -\frac{1}{8} \cdot \frac{4^2 + 12}{4^2} \cdot 4\right) = \left(\frac{1}{4}, -\frac{7}{8}\right).$$

Since $2S$ has non-integral coordinates, it is of infinite order; and so therefore is S.

Since

$$\Theta(3, 6) = (4, 4) = -\Theta(S) = \Theta(-S)$$

it follows that

$$(3, 6) + S \in \ker\Theta = \{O, T\}$$

and so

$$(3, 6) = T - S.$$

Thus
$$F = \{O, T\}.$$

It is an interesting — if long-winded — exercise to show that T and S together generate $\mathcal{E}$:

$$\mathcal{E}(\mathbb{Q}) = \langle T \rangle \oplus \langle S \rangle \cong \mathbb{Z}/(2) \oplus \mathbb{Z}.$$

In other words, each point $P \in \mathcal{E}$ is uniquely expressible in the form

$$P = nS \text{ or } P = T + nS.$$

Note that the subgroup $\mathbb{Z}$ is not unique; if T, S generate $\mathcal{E}$ then so do $T, T + S$.

To show that $\mathcal{E} = \langle T, S \rangle$ we would apply the Method of Infinite Descent; where now each step $P \mapsto 2P$ could be divided into two steps: $P \mapsto \bar{P} = \Theta P \in \bar{\mathcal{E}}$ and $\bar{P} \mapsto \bar{\Theta}\bar{P} = 2P \in \mathcal{E}$.

We leave this as an exercise to the reader, merely observing that even when the rank is known it can be a difficult problem to find free generators, ie to find a $\mathbb{Z}$-basis for $\mathcal{E}/F$.

10.11 Computing the rank — II

Recall that we associate to the elliptic curve

$$\mathcal{E} : y^2 = x^3 + ax^2 + bx$$

a second elliptic curve

$$\mathcal{E}_1 : y^2 = x^3 + a_1x^2 + b_1x,$$

where
$$a_1 = -2a, \; b_1 = a^2 - 4b.$$

The map $\mathcal{E} \to \mathcal{E} : P \mapsto 2P$ factorises into two homomorphisms

$$\Theta : \mathcal{E} \to \mathcal{E}_1, \; \Phi : \mathcal{E}_1 \to \mathcal{E},$$

defined by

$$\Theta(x, y) = \left(\frac{x^2 + ax + b}{x}, \frac{x^2 - b}{x^2} y \right), \Phi(x_1, y_1) = \left(\frac{x_1^2 + a_1x_1 + b_1}{4x_1}, \frac{x_1^2 - b_1}{8x_1^2} y_1 \right),$$

except that in each case the point $(0, 0)$ of order 2 maps to 0. (Thus each homomorphism has kernel $\{0, (0, 0)\}$, since every affine point apart from $(0, 0)$ maps to an affine point.)

It follows (by a little elementary group theory) that

$$\begin{aligned} [\mathcal{E} : 2\mathcal{E}] &= [\mathcal{E} : \operatorname{im}\Phi] \, [\operatorname{im}\Phi : \operatorname{im}\Phi\Theta] \\ &= \frac{[\mathcal{E} : \operatorname{im}\Phi] \, [\mathcal{E}_1 : \operatorname{im}\Theta]}{[\ker\Phi : \ker\Phi \cap \operatorname{im}\Theta]} \end{aligned}$$

Our basic Lemma (corresponding to Mordell's Lemma in the earlier approach) states that $P_1 = (x_1, y_1) \in \mathcal{E}_1$ lies in $\operatorname{im}\Theta$ if and only if x_1 is a perfect square; and similarly $P = (x, y) \in \mathcal{E}$ lies in $\operatorname{im}\Phi$ if and only if x is a perfect square.

Thus if we introduce the auxiliary homomorphisms

$$\chi : \mathcal{E} \to \mathbb{Q}^\times/\mathbb{Q}^{\times 2}, \quad \chi_1 : \mathcal{E}_1 \to \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$$

defined by

$$\begin{aligned} \chi(x, y) &= \bar{x} \; (x \neq 0), \quad \chi(0, 0) = \bar{b} \\ \chi_1(x_1, y_1) &= \overline{x_1} \; (x_1 \neq 0), \quad \chi_1(0, 0) = \bar{b_1}. \end{aligned}$$

then

$$\operatorname{im}\Theta = \ker\chi_1, \quad \operatorname{im}\Phi = \ker\chi.$$

It follows that

$$[\mathcal{E} : 2\mathcal{E}] = \frac{\|\operatorname{im}\chi\|\ \|\operatorname{im}\chi_1\|}{e},$$

where

$$e = \begin{cases} 1 \text{ if } b_1 \text{ is a perfect square,} \\ 2 \text{ otherwise.} \end{cases}$$

Since $r = \operatorname{rank}\mathcal{E}$ is given by

$$2^{r+d} = [\mathcal{E} : 2\mathcal{E}],$$

where $d = 1$ or 2 according as $x^3 + ax^2 + bx$ has 1 rational root or 3, the rank is completely determined once we know $\|\operatorname{im}\chi\|$ and $\|\operatorname{im}\chi_1\|$.

Recall that if

$$P = (x, y) \in \mathcal{E} : y^2 = x^3 + ax^2 + bx + c,$$

where $a, b, c \in \mathbb{Z}$ then x, y take the forms

$$x = \frac{m}{t^2},\ y = \frac{M}{t^3},$$

with $\gcd(m, t) = 1 = \gcd(M, t)$.

As in the earlier method, we represent each rational $x \in \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$ by its square-free part d. Thus if

$$m = du^2$$

where d is square-free then we may take d as the representative of $\bar{x} \in \mathbb{Q}^\times/\mathbb{Q}^{\times 2}$.

Proposition 10.15 *Suppose*

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + ax^2 + bx$$

is an elliptic curve with $a, b \in \mathbb{Z}$. If $d \in \operatorname{im}\chi$ (where d is square-free) then $d \mid b$. Moreover, if $b = dd'$ then $d \in \operatorname{im}\chi$ if and only if there exist u, v, t with $\gcd(u, t) = 1 = \gcd(v, t)$ such that

$$du^4 + au^2t^2 + d't^4 = v^2.$$

Conversely, any solution u, v, t of this equation with $\gcd(u, t) = 1$ arises in this way from a point on $\mathcal{E}$.

Proof ► Suppose

$$P = \left(\frac{du^2}{t^2}, \frac{M}{t^3}\right) \in \mathcal{E}.$$

Then

$$\frac{M^2}{t^6} = \frac{du^2}{t^2}\left(\frac{d^2u^4}{t^4} + a\frac{du^2}{t^2} + b\right).$$

Thus

$$\begin{aligned} M^2 &= du^2(d^2u^4 + adu^2t^2 + bt^4) \\ &= d^2u^2(du^4 + au^2t^2 + d't^4). \end{aligned}$$

. It follows that $du^4 + au^2t^2 + d't^4$ is a perfect square, say

$$du^4 + au^2t^2 + d't^4 = v^2.$$

Conversely, if u, v satisfy this equation then

$$P = \left(\frac{du^2}{t^2}, \frac{duv}{t^3}\right) \in \mathcal{E}.$$

Finally, $\gcd(v, t) = 1$, since

$$p \mid v, t \Longrightarrow p^2 \mid du^2 \Longrightarrow p \mid u,$$

contradicting $\gcd(u, t) = 1$. ◄

10.12 Example

Consider the elliptic curve

$$y^2 = x^3 + 1.$$

over the rationals. There is one point of order 2 on the curve, namely $D = (-1, 0)$.

(The point $P = (2, 3)$ is also on the curve. Since

$$\begin{aligned} \frac{dy}{dx} &= \frac{3x^2}{2y} \\ &= \frac{12}{6} = 2 \end{aligned}$$

at this point, the tangent at P cuts $\mathcal{E}$ again at (X, Y), where

$$2 + 2 + X = 2^2,$$

ie

$$X = 0.$$

It follows that $2P = -D = D$, so that P is of order 4.)

The transformation $x' = x + 1$, ie $x = x' - 1$ (taking the point of order 2 to $(0,0)$) brings the curve to our preferred form

$$\mathcal{E} : x^3 - 3x^2 + 3x$$

Thus

$$a = -3,\ b = 3,$$

and so

$$a_1 = 6,\ b_1 = -3,$$

ie the associated curve is

$$\mathcal{E}_1 : y^2 = x^3 + 6x^2 - 3x.$$

Since there is just one point of order 2 on $\mathcal{E}$, and b_1 is not a perfect square,

$$2^{r+1} = \frac{\|\operatorname{im}\chi\|\ \|\operatorname{im}\chi_1\|}{2},$$

We start by computing $\|\operatorname{im}\chi\|$. Since $d \mid 3$,

$$\operatorname{im}\chi \subset \{\pm 1, \pm 3\}.$$

Since $(0,0) \mapsto 3$,

$$\operatorname{im}\chi = \{1,3\} \text{ or } \{\pm 1, \pm 3\}.$$

Suppose $d = -1$. Then $d' = -3$, and we are looking for solutions of

$$-u^4 - 3u^2t^2 - 3t^4 = v^2.$$

Since the left-hand side is negative while the right-hand side is positive, there is no such solution. Hence

$$\operatorname{im}\chi = \{1,3\}.$$

Turning to $\operatorname{im}\chi_1$, we again have $d \mid 3$, and so

$$\operatorname{im}\chi_1 \subset \{\pm 1, \pm 3\}.$$

But now $(0,0) \mapsto -3$. Thus

$$\operatorname{im}\chi = \{1,-3\} \text{ or } \{\pm 1, \pm 3\}.$$

Again, consider $d = -1$. Now $d' = 3$, and we are looking for solutions of

$$-u^4 + 6u^2t^2 + 3t^4 = v^2.$$

This implies that

$$-u^4 \equiv v^2 \bmod 3.$$

and therefore

$$3 \mid u, v$$

since the quadratic residues mod3 are $\{0,1\}$. But then

$$\begin{aligned} 3^2 \mid u^4, u^2t^2, v^2 &\Longrightarrow 3^2 \mid 3t^4 \\ &\Longrightarrow 3 \mid t, \end{aligned}$$

contradicting the condition $\gcd(u,t) = 1$.

We conclude that

$$\operatorname{im}\chi_1 = \{1, -3\}.$$

Hence

$$2^{r+1} = \frac{2 \cdot 2}{2},$$

ie

$$\operatorname{rank}\mathcal{E} = r = 0.$$

10.13 Another example

Let us re-visit the curve

$$\mathcal{E} : y^2 = x^3 - x,$$

which we already saw has rank 0 (in the last chapter).

The associated curve is

$$\mathcal{E}_1 : y^2 = x^3 + 4x,$$

Since $b_1 = 4$ is a perfect square, while the original equation has three points of order 2,

$$2^{r+2} = \|\operatorname{im}\chi\|\ \|\operatorname{im}\chi_1\|.$$

If $d \in \operatorname{im}\chi$ then $d \mid b = -1$. Thus

$$\operatorname{im}\chi \subset \{\pm 1\}.$$

In fact, since $(0,0) \mapsto -1$,
$$\operatorname{im}\chi = \{\pm 1\}.$$

Turning to $\operatorname{im}\chi_1$, since $d \mid 4 \Longrightarrow d \mid 2$ (as d is square-free),

$$\operatorname{im}\chi_1 \subset \{\pm 1, \pm 2\}.$$

We observe that $(2,4) \in \mathcal{E}_1$. Thus $2 \in \operatorname{im}\chi_1$, and so

$$\operatorname{im}\chi_1 = \{1,2\} \text{ or } \{\pm 1, \pm 2\}.$$

Suppose $d = -1$. Then $d' = -4$, and we are looking for solutions of

$$-u^4 - 4t^4 = v^2,$$

which is impossible, since the left-hand side is negative, while the right-hand side positive. Thus
$$\operatorname{im}\chi_1 = \{1,2\}.$$

We conclude that

$$2^{r+2} = 2 \cdot 2,$$

whence

$$\operatorname{rank}\mathcal{E} = r = 0.$$

10.14 A third example

Finally, let us look again at the curve

$$\mathcal{E}(\mathbb{Q}) : y^2 = x(x-2)(x+4) = x^3 + 2x^2 - 8x,$$

which we already saw (in the last Chapter) has rank 1, with the point $P = (-1,3)$ having infinite order.

Since
$$a_1 = -2a = -4,\ b_1 = a^2 - 4b = 36,$$
the associated curve is

$$\mathcal{E}_1 : y^2 = x^3 - 4x^2 + 36x.$$

Since $b_1 = 36$ is a perfect square,

$$2^{r+2} = \|\operatorname{im}\chi\|\,\|\operatorname{im}\chi_1\|.$$

If $d \in \operatorname{im}\chi$ then $d \mid -8$. Thus

$$\operatorname{im}\chi \subset \{\pm 1, \pm 2\}.$$

Since $(2,0) \mapsto 2$, while $(-4,0) \mapsto -1$, we deduce that

$$\operatorname{im}\chi = \{\pm 1, \pm 2\}.$$

Turning to $\operatorname{im}\chi_1$, we have $d \mid 36$. Thus

$$\operatorname{im}\chi_1 \subset \{\pm 1, \pm 2, \pm 3, \pm 6\}.$$

The point $(0,0) \mapsto 1$ (since $36 \equiv 1$ modulo squares), which is not much help.

Consider $d = -1$. In this case $d' = -36$, and we have to solve the equation

$$-u^4 - 4u^2t^2 - 36t^4 = v^2.$$

Since the left-hand side is < 0, we conclude that $-1 \notin \operatorname{im}\chi_1$.

In fact, any $d < 0$ will lead to a contradiction in the same way. We conclude that
$$\operatorname{im}\chi_1 \subset \{1,2,3,6\}.$$

Suppose $d = 3$. Then $d' = 12$, and the equation reads

$$3u^4 - 4u^2t^2 + 12t^4 = v^2.$$

But this implies that
$$-u^2t^2 \equiv v^2 \bmod 3.$$

Thus $3 \mid v$ and $3 \mid u$ or t. But

$$3 \mid u, v \Longrightarrow 3^2 \mid 12t^4 \Longrightarrow 3 \mid t$$

while

$$3 \mid v, t \Longrightarrow 3^2 \mid 3u^4 \Longrightarrow 3 \mid u,$$

and in either case $\gcd(u,t) > 1$, contrary to assumption.

We conclude that $3 \notin \operatorname{im}\chi_1$; and therefore

$$2^{r+2} \le \frac{4 \cdot 2}{\Longrightarrow} r \le 1.$$

However, we recall that the point $(-1,3) \in \mathcal{E}$ is of infinite order, and so

$$\operatorname{rank}\mathcal{E} = 1.$$

Chapter 11

The modular group

Recall that

$$\mathrm{SL}(2,\mathbb{R}) = \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{R},\ ad - bc = 1\}.$$

By analogy we set

$$\mathrm{SL}(2,\mathbb{Z}) = \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbb{Z},\ ad - bc = 1\}.$$

Proposition 11.1 *The centre of* $\mathrm{SL}(2,R)$ *is* $\{\pm I\}$.

Proof ► Suppose

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Z\left(\mathrm{SL}(2,\mathbb{Z})\right).$$

Let

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},\ T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Then

$$SX = XS \Longrightarrow \begin{pmatrix} -c & -d \\ a & b \end{pmatrix} = \begin{pmatrix} b & -a \\ d & -c \end{pmatrix} \Longrightarrow a = d,\ b = -c;$$

while

$$TX = XT \Longrightarrow \begin{pmatrix} a+c & b+d \\ c & d \end{pmatrix} = \begin{pmatrix} a & a+b \\ c & c+d \end{pmatrix} \Longrightarrow c = 0.$$

Thus

$$b = c = 0 \Longrightarrow X = \pm I.$$

◄

Definition 11.1 *The* modular group Γ *is the quotient-group*

$$\Gamma = \mathrm{SL}(2,R)/\{\pm I\}.$$

Thus each element $g \in \Gamma$ corresponds to two matrices $\pm X$. We write $g = \bar{X}$, or even $g = X$, if that causes no confusion.

The modular group Γ acts on the upper complex plane

$$\mathcal{H} = \{z \in \mathbb{C} : \Im(z) > 0\}$$

by

$$gz = \frac{az+b}{cz+d}$$

if $g = \bar{X}$, where

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

This action is faithful, ie $g \in \Gamma$ acts trivially only if $g = e$. This allows us to identify $g \in \Gamma$ with the corresponding transformation of $\mathcal{H}$.

Definition 11.2 *We define $s, t, u \in \Gamma$ as the elements corresponding to the matrices*

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},\ T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},\ U = ST = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

Proposition 11.2 Γ *is generated by s, t:*

$$\Gamma = \langle s, t \rangle.$$

Proof ► It is sufficient to show that $\mathrm{SL}(2,\mathbb{Z})$ is generated by S, T.

Suppose

$$X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Our strategy is to act on X with S and T on either side so as to minimize $|b| + |c|$. We implement this through the following steps.

Step A Observe that

$$SXS^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

If at any stage $|c| > |b|$ then we can replace X by SXS^{-1}; Thus we may assume that

$$|c| \le |b|.$$

Step B We have

$$XT^r = \begin{pmatrix} a & b+ra \\ c & d+rc \end{pmatrix}.$$

We can choose r so that

$$|b+ra| \le |a|/2.$$

Thus we may assume that

$$|b| \le |a|/2.$$

Step C We have

$$T^r X = \begin{pmatrix} a+rc & b+rd \\ c & d \end{pmatrix}.$$

We can choose r so that

$$|b+rd| \le |d|/2.$$

Thus we may assume that

$$|b| \le |d|/2.$$

Note that in each of these steps, $|b|+|c|$ is either reduced or at worst left unchanged. We may suppose therefore that we reach a stage where none of the steps leads to any "improvement", ie our matrix entries satisfy

$$|c| \le |b|,\ |b| \le |a|/2,\ |b| \le |d|/2.$$

Hence

$$|bc| \le |ad|/4.$$

But

$$\begin{aligned} ad - bc = 1 &\implies |ad| - 1 \le |bc| \\ &\implies |ad| - 1 \le |ad|/4 \\ &\implies |ad| \le 4/3 \\ &\implies |ad| = 1 \\ &\implies |bc| \le 1/4 \\ &\implies |bc| = 0 \\ &\implies b = c = 0. \end{aligned}$$

Thus our final matrix is $\pm I$.

Accordingly, we have found 'wordw' W_1, W_2 in S, T, T^{-1} such that

$$W_1 X W_2 = \pm I.$$

It follows that

$$X = \pm W_1^{-1} W_2^{-1}.$$

Since $-I = S^2$, we have expressed X as a word in S, T, T^{-1}. Thus S, T generate $\mathrm{SL}(2,\mathbb{Z})$; and so s, t generate Γ. ◄

Corollary 21 *Γ is generated by s, u:*

$$\Gamma = \langle s, t \rangle.$$

Theorem 11.1 *Γ is freely-generated by the subgroups $C_2 = \langle s \rangle$, $C_3 = \langle u \rangle$, ie each $g \in \Gamma$ is uniquely expressible in the form*

$$g = u^{i_0} s u^{i_1} \cdots u_{n-1}^{i_{n-1}} s u_n^{i_n},$$

where

$$0 \le i_0, i_n \le 2,\ 1 \le i_j, i_n \le 2 \quad (0 < j < n).$$

Proof ► After the last Corollary, it only remains to prove uniqueness.

Let $\Gamma^+ \subset \Gamma$ correspond to the matrices with *non-negative* entries:

$$\Gamma^+ = \{\bar{X} : X = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : ad - bc = 1,\ a, b, c, d \ge 0.\}$$

Evidently

$$g, h \in \Gamma^+ \implies gh \in \Gamma^+.$$

Now

$$SU = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$$

◄

11.1 Congruence subgroups

If $X, Y \in \mathrm{Mat}(n, \mathbb{Z})$ we write

$$X \equiv Y \pmod{m}$$

as a shorthand for

$$X_{ij} \equiv Y_{ij} \pmod{m}$$

for all i, j.

It is easy to see that

$$X_1 \equiv Y_1 \pmod{m}, X_2 \equiv Y_2 \pmod{m} \Longrightarrow X_1+X_2 \equiv Y_1+Y_2 \pmod{m}, X_1X_2 \equiv Y_1Y_2 \quad (\mathrm{mod}$$

In other words, the map

$$\Theta(n) : \mathrm{Mat}(n, \mathbb{Z}) \to \mathrm{Mat}(n, \mathbb{Z}/(n))$$

under which

$$X \mapsto X \bmod n$$

is a ring-homomorphism.

Definition 11.3 *For each $n \in \mathbb{N} (n \neq 0)$ we define the subgroup $\Gamma(n) \subset \Gamma$ by*

$$\Gamma(n) = \{\bar{X} : X \equiv I \bmod n\}.$$

In other words, $\Gamma(n)$ consists of the transformations

$$z \mapsto \frac{az+b}{cz+d}$$

with

$$a \equiv d \equiv 1 \pmod{n}, \quad b \equiv c \equiv 0 \pmod{n}.$$

Proposition 11.3 $\Gamma(n) = \overline{\ker \Theta_n}$.

Theorem 11.2 *$\Gamma(n)$ is a normal subgroup of Γ, of finite index*

$$[\Gamma : \Gamma(n)] = \prod_{p|n} (p^3 - p).$$

Proof ►

◄

Appendix A

The Structure of Finitely-Generated Abelian Groups

A.1 Finitely-generated abelian groups

Definition A.1 *The abelian group A is said to be generated by the subset $S \subset A$ if each element $a \in A$ is expressible in the form*

$$a = n_1 s_1 + \cdots + n_r s_r \qquad (n_i \in \mathbb{Z}, s_i \in S).$$

A is said to be finitely-generated *if it is generated by a finite set $\{a_1, \ldots, a_r\} \subset A$. We write $A = \langle a_1, \ldots, a_r \rangle$ in this case.*

Proposition A.1 *If*

$$0 \to A \to B \to C \to 0$$

is an exact sequence of abelian groups then B is finitely-generated if and only if A and C are both finitely-generated.

Proof ► Suppose B is generated by $\{b_1, \ldots, b_r\}$. Then the quotient-group C is generated by the images $\{\overline{b_1}, \ldots, \overline{b_r}\}$.

To see that any subgroup $A \subset B$ is also finitely-generated, we argue by induction on r, the number of generators. The quotient-group $B/\langle b_1 \rangle$ is generated by $r-1$ elements. Hence by induction

$$A/A \cap \langle b_1 \rangle \subset B/\langle b_1 \rangle$$

is finitely-generated, by $\{a_1, \ldots, a_s\}$ say. But a subgroup of a cyclic group is cyclic; and so

$$A \cap \langle b_1 \rangle = \langle a \rangle,$$

say. Thus

$$A = \langle a, a_1, \ldots, a_s \rangle.$$

Conversely, suppose A is generated by $\{a_1, \ldots, a_r\}$. and C is generated by $\{\overline{b_1}, \ldots, \overline{b_s}\}$, where $b_1, \ldots, b_s \in B$. Then it is readily verified that B is generated by $\{a_1, \ldots, a_r, b_1, \ldots, b_s\}$. ◀

A.2 Torsion groups

Recall that an abelian group is said to be a *torsion group* if every element is of finite order; conversely, it is said to be *torsion-free* if 0 is the only element of finite order. Evidently a finite abelian group is a torsion group; while a torsion-free group is necessarily infinite.

Proposition A.2 *The elements of finite order in an abelian group A form a subgroup*

$$F = \{a \in A : na = 0 \textit{ for some } n \in \mathbb{N},\ n \neq 0\}.$$

Proof ► To see that F is a subgroup, note that

$$a, b \in F \Longrightarrow ma = 0, nb = 0 \Longrightarrow mn(a+b) = 0 \Longrightarrow a + b \in F.$$

◀

Definition A.2 *We call F the* torsion subgroup *of A. The elements of F, ie the elements of finite order in A, are called* torsion elements *of A.*

Proposition A.3 *The torsion group F of a finitely-generated abelian group A is finite.*

Proof ► This follows at once from Propositions A.1 and the following

Lemma 21 *A finitely-generated torsion group is necessarily finite.*

Proof of Lemma ▷ Suppose A is generated by $\{a_1, \ldots, a_r\}$, and suppose a_i is of order d_i for $1 \leq i \leq r$. Then each element $a \in A$ is expressible in the form

$$a = n_1 a_1 + \cdots + n_r a_r$$

where

$$0 \leq n_i < d_i.$$

Thus

$$\|A\| \leq d_1 \cdots d_r.$$

◁ ◀

Proposition A.4 *The quotient-group A/F is torsion-free.*

Proof ► Suppose $n\bar{a} = 0$, where $a \in A$. Then $na \in F$. Hence $mna = 0$ for some m. Thus a is of finite order, ie $a \in F$. In other words, $\bar{a} = 0$. ◀

The last 2 Propositions allow us to divide our task — the study of finitely-generated abelian groups — into 2 parts: finite abelian groups, and finitely-generated torsion-free abelian groups. These are the subjects of the next 2 sections.

A.3 Finite Abelian Groups

Proposition A.5 *Suppose A is an abelian group. For each prime p, the elements of order p^n in A for some $n \in N$ form a subgroup*

$$A_p = \{a \in A : p^n a = 0 \textit{ for some } n \in \mathbb{N}\}.$$

Proof ► Suppose $a, b \in A_p$. Then

$$p^m a = 0,\ p^n b = 0,$$

for some m, n. Hence

$$p^{m+n}(a+b) = 0,$$

and so $a + b \in A_p$. ◀

Definition A.3 *We call A_p the p-*component *of A.*

Proposition A.6 *Suppose A is an abelian group. Then the torsion subgroup F is the direct sum of the A_p:*

$$F = \oplus_p A_p.$$

Proof ► Suppose $a \in F$, say $na = 0$. Let

$$n = p_1^{e_1} \cdots p_r^{e_r};$$

and set

$$m_i = n / e_i^{p_i}.$$

Then $\gcd(m_1, \ldots, m_r) = 1$, and so we can find $n_1, \ldots, n_r$ such that

$$m_1 n_1 + \cdots + m_r n_r = 1.$$

Thus
$$a = a_1 + \cdots + a_r,$$
where
$$a_i = m_i n_i a.$$
But
$$p_i^{e_i} a_i = (p_i^{e_i} m_i) n_i a = n n_i a = 0$$
(since $na = 0$). Hence
$$a_i \in A_{p_i}.$$
Thus A is the sum of the subgroups A_p.

To see that this sum is direct, suppose
$$a_1 + \cdots + a_r = 0,$$
where $a_i \in A_{p_i}$, with distinct primes $p_1, \ldots, p_r$. Suppose
$$p_i^{e_i} a_i = 0.$$
Let
$$m_i = p_1^{e_1} \cdots p_{i-1}^{e_{i-1}} p_{i+1}^{e_{i+1}} \cdots p_r^{e_r}.$$
Then
$$m_i a_j = 0 \text{ if } i \neq j.$$
Thus (multiplying the given relation by m_i),
$$m_i a_i = 0.$$
But $\gcd(m_i, p_i^{e_i}) = 1$. Hence we can find m, n such that
$$m m_i + n p_i^{e_i} = 1.$$
But then
$$a_i = m(m_i a_i) + n(p_i^{e_i} a_i) = 0.$$
We conclude that A is the direct sum of its p-components A_p. ◂

Proposition A.7 *If A is a finitely-generated abelian group then $A_p = 0$ for almost all p, ie for all but a finite number of p.*

Proof ▸ The torsion subgroup $F \subset A$ is finite, by Proposition reffinite. Thus the orders of all the elements of F have only a finite number of prime factors. If p is not among these primes then evidently $A_p = 0$. ◂

Theorem A.1 *Suppose A is a finite abelian p-group (ie each element is of order p^e for some e). Then A can be expressed as a direct sum of cyclic p-groups:*
$$A = \mathbb{Z}/(p^{e_1}) \oplus \cdots \oplus \mathbb{Z}/(p^{e_r}).$$
Moreover the powers $p^{e_1}, \ldots, p^{e_r}$ are uniquely determined by A.

Proof ▸ We argue by induction on $\|A\| = p^n$. We may assume therefore that the result holds for the subgroup
$$pA = \{pa : a \in A\}.$$
For pA is stricty smaller than A, since
$$pA = A \Longrightarrow p^n A = A,$$
while we know from Lagrange's Theorem that $p^n A = 0$.

Suppose
$$pA = \langle pa_1 \rangle \oplus \cdots \oplus \langle pa_r \rangle.$$
Then the sum
$$\langle a_1 \rangle + \cdots + \langle a_r \rangle = B,$$
say, is direct. For suppose
$$n_1 a_1 + \cdots + n_r a_r = 0.$$
If $p \mid n_1, \ldots, n_r$, say $n_i = p m_i$, then we can write the relation in the form
$$m_1(pa_1) + \cdots + m_r(pa_r) = 0,$$
whence $m_i p a_i = n_i a_i = 0$ for all i.

On the other hand, if p does not divide all the n_i then
$$n_1(pa_1) + \cdots + n_r(pa_r) = 0,$$
and so $p n_i a_i = 0$ for all i. But if $p \nmid n_i$ this implies that $pa_i = 0$. (For the order of a_i is a power of p, say p^e; while $p^e \mid n_i p$ implies that $e \leq 1$.) But this contradicts our choice of pa_i as a generator of a direct summand of pA. Thus the subgroup $B \subset A$ is expressed as a direct sum
$$B = \langle a_1 \rangle \oplus \cdots \oplus \langle a_r \rangle.$$

Let
$$K = \{a \in A : pa = 0\}.$$

Then

$$A = B + K.$$

For suppose $a \in A$. Then $pa \in pA$, and so

$$pa = n_1(pa_1) + \cdots + n_r(pa_r)$$

for some $n_1, \ldots, n_r \in \mathbb{Z}$. Thus

$$p(a - n_1a_1 - \cdots - n_ra_r) = 0,$$

and so

$$a - n_1a_1 - \cdots - n_ra_r = k \in K.$$

Hence

$$a = (n_1a_1 + \cdots + n_ra_r) + k \in B + K.$$

If $B = A$ then all is done. If not, then $K \not\subset B$, and so we can find $k_1 \in K, k_1 \notin B$. Now the sum

$$B_1 = B + \langle k_1 \rangle$$

is direct. For $\langle k_1 \rangle$ is a cyclic group of order p, and so has no proper subgroups. Thus

$$B \cap \langle k_1 \rangle = \{0\},$$

and so

$$B_1 = B \oplus \langle k_1 \rangle$$

If now $B_1 = A$ we are done. If not we can repeat the construction, by choosing $k_2 \in K, k_2 \notin B_1$. As before, this gives us a direct sum

$$B_2 = B_1 \oplus \langle k_2 \rangle = B \oplus \langle k_1 \rangle \oplus \langle k_2 \rangle.$$

Continuing in this way, the construction must end after a finite number of steps (since A is finite):

$$\begin{aligned} A = B_s &= B \oplus \langle k_1 \rangle \oplus \cdots \oplus \langle k_s \rangle \\ &= \langle a_1 \rangle \oplus \cdots \oplus \langle a_r \rangle \oplus \langle k_1 \rangle \oplus \cdots \oplus \langle k_s \rangle. \end{aligned}$$

It remains to show that the powers $p^{e_1}, \ldots, p^{e_r}$ are uniquely determined by A. This follows easily by induction. For if A has the form given in the theorem then

$$pA = \mathbb{Z}/(p^{e_1-1}) \oplus \cdots \oplus \mathbb{Z}/(p^{e_r-1}).$$

Thus if $e > 1$ then $\mathbb{Z}/(p^e)$ occurs as often in A as $\mathbb{Z}/(p^{e-1})$ does in pA. It only remains to deal with the factors $\mathbb{Z}/(p)$. But the number of these is now determined by the order $\|A\|$ of the group. ◄

Remark: It is important to note that if we think of A as a direct sum of cyclic *subgroups*, then the orders of these subgroups are uniquely determined, by the theorem; but *the actual subgroups themselves are not in general uniquely determined.* In fact the only case in which they are uniquely determined (for a finite p-group A) is if A is itself cyclic,

$$A = \mathbb{Z}/(p^e),$$

in which case of course there is just one summand.

To see this, it is sufficient to consider the case of 2 summands:

$$A = \mathbb{Z}/(p^e) \oplus \mathbb{Z}/(p^f).$$

We may suppose that $e \geq f$. Let a_1, a_2 be the generators of the 2 summands. Then it is easy to see that we could equally well take $a_1' = a_1 + a_2$ in place of a_1:

$$A = \langle a_1 + a_2 \rangle \oplus \langle a_2 \rangle.$$

For certainly these elements $a_1 + a_2, a_2$ generate the group; and the sum must be direct, since otherwise there would not be enough terms $m_1a_1' + m_2a_2$ to give all the p^{e+f} elements in A.

A.4 Torsion-free Abelian Groups

Definition A.4 *To each abelian group A we associate the vector space $V = V(A)$ over $\mathbb{Q}$ given by*

$$V = A \otimes_{\mathbb{Z}} \mathbb{Q}.$$

Remarks:

1. Concretely, we construct V from A as follows. Each element $v \in V$ is of the form

   $$v = \lambda a \quad (\lambda \in \mathbb{Q}, a \in A).$$

   Two elements

   $$v = \lambda a, \ w = \mu b.$$

   are equal if we can find m, n, N such that

   $$\lambda = \frac{m}{N}, \ \mu = \frac{n}{N}, \ ma = nb.$$

In other words, a linear relation

$$\lambda_1 v_1 + \cdots + \lambda_r v_r = 0$$

holds in V if when multiplied by some integer N with $N\lambda_1, \ldots, N\lambda_r \in \mathbb{Z}$ it yields a relation that holds in A.

2. We can put this in a more general setting. Recall that a *module* M over a ring R (not necessarily commutative, but with identity element 1) is defined by giving an abelian group A on which R acts so that

   (a) $\lambda(\mu m) = (\lambda\mu)m$;

   (b) $(\lambda + \mu)m = \lambda m + \mu m$;

   (c) $\lambda(m + n) = \lambda m + \lambda n$;

   (d) $1m = m$.

   There are 2 special cases of importance. Firstly, a module over a field k is just a vector space over k. Thus the concept of a module may be seen as a natural generalisation of that of a vector space, in which the scalars are allowed to form a ring.

   Secondly, a module over the integers $\mathbb{Z}$ is just an abelian group.

   Suppose
   $$\phi : R \to S$$
   is a ring-homomorphism. Then each R-module M gives rise to an S-module N, where
   $$N = S \otimes_R M.$$
   Concretely, each element $n \in N$ is expressible as a sum
   $$n = s_1 m_1 + \cdots + s_r m_r,$$
   with addition and scalar multiplication being defined in the natural way. We have a natural map
   $$M \to N : m \mapsto 1 \cdot m.$$
   Our case arises in this way from the natural injection
   $$i : \mathbb{Z} \to \mathbb{Q}.$$
   It is a special case in so far as each element of V is expressible as a single element λa rather than a sum of such elements. As we just observed, we have a natural group homomorphism
   $$A \to V : a \mapsto 1 \cdot a.$$

3. In the language of categories and functors, we have a covariant functor
   $$\mathcal{F} : \mathcal{A} \to \mathcal{V}$$
   from the category $\mathcal{A}$ of abelian groups to the category $\mathcal{V}$ of vector spaces over $\mathbb{Q}$.

Definition A.5 *The* rank $r(A)$ *of the abelian group A is defined to be the dimension of V:*
$$r(A) = \dim_{\mathbb{Q}} V.$$

Proposition A.8 *A finitely-generated abelian group has finite rank.*

Proof ► If $A = \langle a_1, \ldots, a_n \rangle$ then $1 \cdot a_1, \ldots, 1 \cdot a_n$ span V, and so
$$r(A) \leq n.$$
◄

Proposition A.9 *Suppose A is an abelian group. Then the map*
$$A \to V : a \mapsto 1 \cdot a$$
is a homomorphism of abelian groups, with kernel F.

Proof ► Suppose $a \mapsto 0$, ie $1 \cdot a = 0$ in V. By definition this means that $Na = 0$ for some $N \in \mathbb{N}$ $(N \neq 0)$. In other words, $a \in F$. ◄

Corollary 1 *An abelian group A is of rank 0 if and only if it is a torsion group.*

Corollary 2 *A torsion-free abelian group A can be embedded in a vector space V over $\mathbb{Q}$:*
$$A \subset V.$$

Theorem A.2 *A finitely-generated torsion-free abelian group A is necessarily free, ie A is expressible as a direct sum of copies of the integers $\mathbb{Z}$:*
$$A = r\mathbb{Z} = \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}.$$

Proof ► We have seen that $A \subset V$, where V is a finite-dimensional vector space over $\mathbb{Q}$. Suppose $a_1, \ldots, a_n$ generate A. Then these elements span V. Hence we can choose a basis for V from among them. After re-ordering we may suppose the $a_1, \ldots, a_r$ form a basis for V.

We derive a $\mathbb{Z}$-basis $b_1, \ldots, b_r$ for A as follows. Choose b_1 to be the smallest positive multiple of a_1 in A:

$$b_1 = \lambda_1 a_1 \in A.$$

(It is easy to see that $\lambda_1 = 1/m_1$ for some $m \in \mathbb{N}$.)

Now choose b_2 to be an element of A in the vector subspace $\langle a_1, a_2 \rangle$ with smallest positive second coefficient

$$b_2 = \mu_1 a_1 + \lambda_2 a_2 \in A.$$

(Again, it is easy to see that $\lambda_2 = 1/m_2$ for some $m \in \mathbb{N}$.)

Continuing in this way, choose b_i to be an element of A in the vector subspace $\langle a_1, \ldots, a_i \rangle$ with smallest positive ith coefficient

$$b_i = \mu_1 a_1 + \cdots + \mu_{i-1} a_{i-1} + \lambda_i a_i \in A.$$

(Once again, it is easy to see that $\lambda_i = 1/m_i$ for some $m \in \mathbb{N}$.)

Finally, we choose b_r to be an element of A with smallest positive last coefficient

$$b_r = \mu_1 a_1 + \cdots + \mu_{r-1} a_{i-1} + \lambda_r a_i \in A.$$

We assert that $b_1, \ldots, b_r$ forms a $\mathbb{Z}$-basis for A. For suppose $a \in A$. Let

$$a = \rho_{r,1} a_1 + \cdots + \rho_{r,r} a_r,$$

where $\rho_1, \ldots, \rho_r \in \mathbb{Q}$. The last coefficient $\rho_{r,r}$ must be an integral multiple of λ_r,

$$\rho_{r,r} = n_r \lambda_r.$$

For otherwise we could find a combination $ma + nb_r$ with last coefficient positive but smaller than λ_r.

But now

$$a - n_r b_r \in \langle a_1, \ldots, a_{r-1} \rangle,$$

say

$$a - n_r b_r = \rho_{r-1,1} a_1 + \cdots + \rho_{r-1,r-1} a_{r-1}.$$

By the same argument, the last coefficient $\rho_{r-1,r-1}$ is an integral multiple of λ_{r-1}.

$$\rho_{r-1,r-1} = n_{r-1} \lambda_{r-1},$$

and so

$$a - n_r b_r - n_{r-1} b_{r-1} \in \langle a_1, \ldots, a_{r-2} \rangle.$$

Continuing in this fashion, we find finally that

$$a = n_r b_r + n_{r-1} b_{r-1} + n_1 b_1,$$

with $n_r, \ldots, n_1 \in \mathbb{Z}$. Thus $b_1, \ldots, b_r$ forms a $\mathbb{Z}$-basis for A, and

$$A = \mathbb{Z} b_1 \oplus \cdots \oplus \mathbb{Z} b_r \equiv r\mathbb{Z}.$$

◄

Remark: We can think of the summands $\mathbb{Z}$ as subgroups of A. But it should be noted that these subgroups are not unique, unless $A = \mathbb{Z}$. For there are many ways of splitting $\mathbb{Z} \oplus \mathbb{Z}$ into 2 direct summands. In fact, if the generators of these summands are e, f,

$$A = \mathbb{Z}e \oplus \mathbb{Z}f,$$

then we can take as generators any pair

$$n_{11}e + n_{12}f, \; n_{21}e + n_{22}f,$$

(where $n_{11}, n_{12}, n_{21}, n_{22} \in \mathbb{Z}$) provided

$$\det \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix} = \pm 1,$$

that is, the matrix must be *unimodular.*

The corresponding result holds for $r\mathbb{Z}$: any unimodular transformation will give us a new expression for the group as a direct sum of subgroups isomorphic to $\mathbb{Z}$.

Theorem A.3 *Every finitely-generated abelian group A is the direct sum of its torsion group F and a torsion-free group P:*

$$A = F \oplus P.$$

Proof ► Let F be the torsion subgroup of A.

Lemma 22 *The quotient-group*

$$Q = A/F$$

is torsion-free.

Proof of Lemma ▷ For suppose $\bar{a} \in Q$ (where $a \in A$) has finite order, say $n\bar{a} = 0$, for some $n > 0$. In other words, $na \in F$. But then $m(na) = 0$ for some $m > 0$. Thus a is of finite order, ie $a \in F$, and so $\bar{a} = 0$. ◁

It follows from Proposition A.2 that Q is a direct sum of copies of $\mathbb{Z}$:

$$Q = \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}.$$

Choose elements $a_1, \ldots, a_r$ in A mapping onto the elements $(1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 1)$ in Q; and let

$$P = \langle a_1, \ldots, a_r \rangle.$$

We shall show that $A = F \oplus P$.

Recall that the abelian group A is the direct sum of the subgroups B and C,

$$A = B \oplus C,$$

if and only if

1. $B \cap C = \{0\}$;
2. $A = B + C$, ie each element $a \in A$ is expressible in the form $a = b + c$, with $b \in B, c \in C$.

We apply this with $B = F$, $C = P$. Firstly, $F \cap P = \{0\}$. For suppose $a \in F \cap P$. Since $a \in P$,

$$a = n_1 a_1 + \cdots + n_r a_r$$

for some $n_1, \ldots, n_r \in \mathbb{Z}$. Since $a \in F$, we have $na = 0$ for some $n > 0$. Thus

$$nn_1 a_1 + \cdots + nn_r a_r = 0.$$

It follows — going over to the quotient group Q — that

$$nn_1 e_1 + \cdots + nn_r e_r = 0.$$

But that implies that $nn_1 = \cdots = nn_r = 0$, since $e_1, \ldots, e_r$ form a $\mathbb{Z}$-basis for Q. Thus $n_1 = \cdots = n_r = 0$, and so $a = 0$, ie $F \cap P = \{0\}$.

Secondly, suppose $a \in A$. Then $\bar{a} \in Q$ can be expressed in the form

$$\bar{a} = m_1 e_1 + \cdots + m_r e_r,$$

for some $m_1, \ldots, m_r \in \mathbb{Z}$. But then

$$a - m_1 a_1 - \cdots - m_r a_r = f \in F.$$

Thus

$$a = f + m_1 a_1 + \cdots + m_r a_r \in F + P.$$

It follows that

$$A = F \oplus P.$$

◀

Corollary *Every finitely-generated abelian group A is the direct sum of a finite group F and a number of copies of $\mathbb{Z}$:*

$$A = F \oplus \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}.$$

Remark: While F is unique — it is the torsion subgroup of A — the subgroups corresponding to the copies of $\mathbb{Z}$ not in general unique.

In fact the only cases in which the subgroups are unique is if either the group is finite (so that $A = F$) or else $A = \mathbb{Z}$ (so that $F = 0$ and there is just one copy of Z). For we can split

$$A = F \oplus \mathbb{Z}$$

in many ways if $F \neq \{0\}$. In fact if e is a generator of $\mathbb{Z}$,

$$A = F \oplus \langle e \rangle,$$

then we can replace e by $e + f$, where f is any element of F:

$$A = F \oplus \langle e + f \rangle,$$

For $e + f$ has infinite order, and so every non-zero element of $\langle e + f \rangle$ also has infinite order. Hence

$$F \cap \langle e + f \rangle = \{0\},$$

and so the sum is direct.

A.5 The Structure Theorem

Putting together the results of the last 3 sections, we derive the Structure Theorem for Finitely-Generated Abelian Groups.

Theorem A.4 *Every finitely-generated abelian group A is expressible as a direct sum of cyclic groups (including $\mathbb{Z}$):*

$$A = \mathbb{Z}/(p^{e_1}) \oplus \cdots \oplus \mathbb{Z}/(p^{e_s}) \oplus \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}.$$

Moreover the prime-powers $p_1^{e_1}, \ldots, p_s^{e_s}$ and the number of copies of $\mathbb{Z}$ are uniquely determined by A.

Remark: If we think of the Theorem as expressing A as a direct sum of cyclic subgroups, then in general these subgroups will not be unique, although their orders (p^e or ∞) will be.

The only case in which the expression will be unique is if A is cyclic. For if that is so then either $A = \mathbb{Z}$ or else A is a finite cyclic group $\mathbb{Z}/(n)$. In this last case each p-component A_p is also cyclic, since every subgroup of a cyclic abelian group is cyclic. Thus the expression for A as a direct sum in the Theorem is just the splitting of A into its p-components A_p; and we know that this is unique.

Conversely, if A is not cyclic, then either

1. A has at least 2 $\mathbb{Z}$ summands; or
2. A has a component $\mathbb{Z}$ and $F \neq \{0\}$; or
3. some component A_p is not cyclic.

In each of these cases we have seen above that the splitting is not unique.

Appendix B

Fermat's Last Theorem when $n = 4$

B.1 The Case $n = 2$

The equation

$$x^2 + y^2 = z^2$$

certainly has solutions, eg $(3, 4, 5)$ and $(5, 12, 13)$. This does not contradict Fermat's Last Theorem, of course, since that only asserts there is no solution *if* $n > 2$.

Pythagoras already knew that this equation (with $n = 2$) had an infinity of solutions; and Diophantus later found all the solutions, following the technique below.

In the first place, we may assume that

$$\gcd(x, y, z) = 1.$$

We may also assume that $x, y, z > 0$. We shall use the term *Pythagorean triple* for a solution with these properties.

Note that modulo 4

$$x^2 = \begin{cases} 0 \bmod 4 & \text{if } x \text{ is even,} \\ 1 \bmod 4 & \text{if } x \text{ is odd.} \end{cases}$$

It follows that x and y cannot both be odd; for then we would have $z^2 = 2 \bmod 4$, which is impossible. Thus just one of x and y is even; and so z must be odd. We can assume without loss of generality that x is even, say $x = 2X$. Our equation can then be written

$$4X^2 = z^2 - y^2 = (z + y)(z - y).$$

We know that $2 \mid z+y,\ 2 \mid z-y$, since y, z are both odd. On the other hand no other factor can divide $z+y$ and $z-y$:

$$\gcd(z+y, z-y) = 2.$$

For

$$d \mid z+y,\ z-y \Longrightarrow d \mid 2y,\ 2z.$$

It follows that

$$z+y = 2u^2, \quad z-y = 2v^2, \quad x = 2uv.$$

Thus

$$(x, y, z) = (2uv, u^2 - v^2, u^2 + v^2).$$

where $\gcd(u, v) = 1$. Note that just one of u, v must be odd; for if both were odd, x, y, z would all be even.

Every Pythagorean triple arises in this way from a unique pair (u, v) with $\gcd(u, v) = 1$, $u > v > 0$, and just one of u, v odd. The uniqueness follows from the fact that

$$(u+v)^2 = z + x, \quad (u-v)^2 = z - x.$$

For this shows that x, y, z determine $u+v$ and $u-v$, and therefore u and v.

B.2 The Case $n = 4$

The only case of his "Theorem" that Fermat actually proved, as far as we know, was the case $n = 4$:

$$x^4 + y^4 = z^4.$$

His proof was based on a technique which he invented: *the Method of Infinite Descent*. Basically, this consists in showing that from any solution of the equation in question one can construct a second, smaller, solution.

Actually, we are going to apply this to the Diophantine equation

$$x^4 + y^4 = z^2.$$

If we can show that this has no solution in non-zero integers, then the same will be true *a fortiori* of Fermat's equation with $n = 4$.

Suppose (x, y, z) is a solution of this equation. As before we may and shall suppose that $x, y, z > 0$ and $\gcd(x, y.z) = 1$. Evidently (x^2, y^2, z) is

then a Pythagorean triple, and so can be expressed in the form (swapping x, y if necessary)

$$x^2 = 2ab,\ y^2 = a^2 - b^2,\ z = a^2 + b^2,$$

where a, b are positive integers with $\gcd(a, b) = 1$. Since x is even, $4 \mid x^2$, and therefore just one of a and b must be even.

If a were even and b were odd, then $a^2 - b^2 = 3 \bmod 4$, so the second equation $y^2 = a^2 - b^2$ would be untenable. Thus b is even, and so from the first equation $x^2 = 2ab$ we can write

$$a = u^2,\ b = 2v^2,\ x = 2uv,$$

where $\gcd(u, v) = 1$, and $u, v > 0$.

The second equation now reads

$$y^2 = u^4 - 4v^4.$$

Thus

$$4v^4 + y^2 = u^4,$$

and so $(2v^2, y, u^2)$ is a Pythagorean triple. It follows that we can write

$$2v^2 = 2st,\ y = s^2 - t^2,\ u^2 = s^2 + t^2,$$

where $\gcd(s, t) = 1$. From the first equation we can write

$$s = X^2,\ t = Y^2,\ v = XY,$$

where $\gcd(X, Y) = 1$, and $X, Y > 0$; and so on writing Z for u the third equation reads

$$X^4 + Y^4 = Z^2,$$

which is just the equation we started from. So from any solution (x, y, z) of the equation

$$x^4 + y^4 = z^2$$

with $\gcd(x, y, z) = 1$, $x, y > 0$ and x even, we obtain a second solution (X, Y, Z) with $\gcd(X, Y, Z) = 1$, $X, Y > 0$ and X even, where

$$\begin{aligned} x &= 2uv = 2XYZ, \\ y &= s^2 - t^2 = X^4 - Y^4, \\ z &= a^2 + b^2 = u^4 + v^4 = Z^4 + X^4Y^4. \end{aligned}$$

The new solution is evidently smaller than the first in every sense. In particular,

$$Z < z^{1/4};$$

so our infinite chain must (rapidly) lead to a contradiction, and Fermat's Last Theorem is proved for $n = 4$.

Appendix C

Fermat's Last Theorem when $n = 3$

Having proved Fermat's Last Theorem for $n = 4$, it only (?) remains to prove it for odd primes $3, 5, 7, 11, \ldots$. It is convenient in this case to take the equation in symmetric form

$$x^p + y^p + z^p = 0$$

(on replacing z by $-z$).

Our proof for $p = 3$ is based, like that for $n = 4$, on Fermat's Method of Infinite Descent. But now we have to mix in a little algebraic number theory.

C.1 Algebraic numbers

Definition C.1 *A number $\alpha \in \mathbb{C}$ is said to be* algebraic *if it satisfies a polynomial equation*

$$f(x) = x^n + a_1 x^{n-1} + \cdots + a_n = 0$$

with rational coefficients $a_i \in \mathbb{Q}$.

For example, $\sqrt{2}$ and i are algebraic.

A number is said to be *transcendental* if it is not algebraic. Both e and π are transcendental. It is in general extremely difficult to prove a number transcendental, and there are many open problems in this area, eg it is not known if π^e is transcendental.

Proposition C.1 *The algebraic numbers form a field $\bar{\mathbb{Q}} \subset \mathbb{C}$.*

Proof ► If α satisfies the equation $f(x) = 0$ then $-\alpha$ satisfies $f(-x) = 0$, while $1/\alpha$ satisfies $x^n f(1/x) = 0$ (where n is the degree of $f(x)$). It follows that $-\alpha$ and $1/\alpha$ are both algebraic. Thus it is sufficient to show that if α, β are algebraic then so are $\alpha + \beta, \alpha\beta$.

Suppose α satisfies the equation

$$f(x) \equiv x^m + a_1 x^{m-1} + \cdots + a_m = 0,$$

and β the equation

$$g(x) \equiv x^n + b_1 x^{n-1} + \cdots + b_n = 0.$$

Consider the vector space

$$V = \langle \alpha^i \beta^j : 0 \le i < m,\ 0 \le j < n \rangle$$

over $\mathbb{Q}$ spanned by the mn elements $\alpha^i \beta^j$. Evidently

$$\alpha + \beta, \alpha\beta \in V.$$

But if $\theta \in V$ then the $mn + 1$ elements

$$1, \theta, \theta^2, \ldots, \theta^{mn}$$

are necessarily linearly dependent (over $\mathbb{Q}$), since $\dim V \le mn$. In other words θ satisfies a polynomial equation of degree $\le mn$. Thus each element $\theta \in V$ is algebraic. In particular $\alpha + \beta$ and $\alpha\beta$ are algebraic. ◄

C.2 Algebraic integers

Definition C.2 *A number $\alpha \in \mathbb{C}$ is said to be an* algebraic integer *if it satisfies a polynomial equation*

$$f(x) = x^n + a_1 x^{n-1} + \cdots + a_n = 0$$

with integral coefficients $a_i \in \mathbb{Z}$.

Proposition C.2 *The algebraic integers form a ring $\bar{\mathbb{Z}} \subset \bar{\mathbb{Q}}$. That is, if α, β are algebraic integers, then so are $\alpha + \beta$, $\alpha - \beta$ and $\alpha\beta$.*

Proof ► If α is a root of the monic polynomial $f(x)$ then $-\alpha$ is a root of the monic polynomial $f(-x)$. It follows that if α is an algebraic integer then so is $-\alpha$. Thus it is sufficient to show that if α, β are algebraic integers then so are $\alpha + \beta, \alpha\beta$.

Suppose α satisfies the equation

$$f(x) \equiv x^m + a_1 x^{m-1} + \cdots + a_m = 0 \quad (a_1, \ldots, a_m \in \mathbb{Z}),$$

and β the equation

$$g(x) \equiv x^n + b_1 x^{n-1} + \cdots + b_n = 0 \quad (b_1, \ldots, b_n \in \mathbb{Z}).$$

Consider the abelian group (or $\mathbb{Z}$-module)

$$M = \langle \alpha^i \beta^j : 0 \le i < m,\ 0 \le j < n \rangle$$

generated by the mn elements $\alpha^i \beta^j$. Evidently

$$\alpha + \beta, \alpha\beta \in V.$$

As a finitely-generated torsion-free abelian group, M is isomorphic to $\mathbb{Z}^d$ for some d. Moreover M is *noetherian*, ie every increasing sequence of subgroups of M is stationary: if

$$S_1 \subset S_2 \subset S_3 \cdots \subset M$$

then for some N,

$$S_N = S_{N+1} = S_{N+2} = \cdots.$$

Suppose $\theta \in M$. Consider the increasing sequence of subgroups

$$\langle 1 \rangle \subset \langle 1, \theta \rangle \subset \langle 1, \theta, \theta^2 \rangle \subset \cdots.$$

This sequence must become stationary; that is to say, for some N

$$\theta^N \in \langle 1, \theta, \ldots, \theta^{N-1} \rangle.$$

In other words, θ satisfies an equation of the form

$$\theta^N = a_1 \theta^{N-1} + a_2 \theta^{N-2} + \cdots.$$

Thus every $\theta \in M$ is an algebraic integer. In particular $\alpha + \beta$ and $\alpha\beta$ are algebraic integers. ◄

Proposition C.3 *A rational number $c \in \mathbb{Q}$ is an algebraic integer if and only if it is a rational integer:*

$$\bar{\mathbb{Z}} \cap \mathbb{Q} = \mathbb{Z}.$$

Proof ► Suppose $c = m/n$, where $\gcd(m, n) = 1$; and suppose c satisfies the equation

$$x^d + a_1 x^{d-1} + \cdots + a_d = 0 \quad (a_i \in \mathbb{Z}).$$

Then

$$m^d + a_1 m^{d-1} n + \cdots + a_d n^d = 0.$$

Since n divides every term after the first, it follows that $n \mid m^d$. But that is incompatible with $\gcd(m, n) = 1$, unless $n = 1$, ie $c \in \mathbb{Z}$. ◄

Definition C.3 *A number $\alpha \in \mathbb{C}$ is said to be a* unit *if both α and $1/\alpha$ are algebraic integers.*

Any root of unity, ie any number satisfying $x^n = 1$ for some n, is a unit. But these are not the only units; for example, $\sqrt{2} - 1$ is a unit. The units form a multiplicative subgroup of $\bar{\mathbb{Q}}^\times$.

C.3 The field $\mathbb{Q}(\omega)$

Let

$$\omega = e^{2\pi i/3}.$$

Then $\omega^3 = 1$; more precisely,

$$\omega^2 + \omega + 1 = 0.$$

Proposition C.4 *The numbers of the form*

$$a + \omega b \quad (a, b \in \mathbb{Q})$$

form a field.

Proof ► $\mathbb{Q}(\omega)$ is closed under addition, subtraction and multiplication. It only remains to show that it is closed under division. Suppose $\theta \in \mathbb{Q}(\omega)$, $\theta \neq 0$. Since $\mathbb{Q}(\omega)$ is a vector space of dimension 2 over $\mathbb{Q}$, the elements $1, \theta, \theta^2$ are linearly dependent over $\mathbb{Q}$, ie θ satisfies an equation of degree 1 or 2 over $\mathbb{Q}$.

If θ satisfies an equation of degree 1 over $\mathbb{Q}$ then $\theta \in \mathbb{Q}$, and so $1/\theta \in \mathbb{Q} \subset \mathbb{Q}(\omega)$.

Suppose θ satisfies the equation

$$\theta^2 + b\theta + c = 0.$$

We may suppose $c \neq 0$ (or else divide the equation by θ). Then

$$\theta^{-1} = -c^{-1}\theta - c^{-1}b \in \mathbb{Q}(\omega).$$

◄

C.3.1 Automorphisms and norms

The conjugacy automorphism

$$z \mapsto \bar{z} : \mathbb{C} \to \mathbb{C}$$

of the complex numbers induces an automorphism of $\mathbb{Q}(\omega)$, under which

$$\omega \mapsto \bar{\omega} = \omega^2,$$

and more generally

$$a + \omega b \mapsto a + \omega^2 b = (a - b) - b\omega.$$

If $\xi = a + \omega b$, we call $\bar{\xi} = a + \omega^2 b$ the *conjugate* of ξ in $\mathbb{Q}(\omega)$. If ξ satisfies a polynomial equation $f(x) =$ with coefficients in $\mathbb{Q}$, then so does its conjugate $\bar{\xi}$. (This follows on applying the automorphism to the equation $f(\xi) = 0$. The coefficients of f will be left untouched, since they lie in $\mathbb{Q}$, while each power ξ^n will be replaced by $\bar{\xi}^n$.) In particular, if ξ is an algebraic integer, then so is $\bar{\xi}$

The product

$$N(\xi) = \xi\bar{\xi} = |\xi|^2$$

is called the *norm* of ξ. Clearly the norm is multiplicative:

$$N(\alpha\beta) = N(\alpha)N(\beta).$$

C.4 The ring $\mathbb{Z}[\omega]$

Which numbers in $\mathbb{Q}(\omega)$ are algebraic integers? The answer is not obvious.

Certainly ω is an algebraic integer, since it satisfies $x^3 - 1 = 0$; and so are all the numbers in the set $\mathbb{Z}[\omega]$ consisting of numbers of the form

$$a + \omega b \quad (a, b \in \mathbb{Z})$$

since the algebraic integers are closed under addition and multiplication.

Proposition C.5 *The algebraic integers in $\mathbb{Q}(\omega)$ are just the elements of $\mathbb{Z}[\omega]$.*

Proof ► Suppose

$$\xi = a + \omega b \quad (a, b \in \mathbb{Q})$$

is an algebraic integer. Then so is its conjugate

$$\bar{\xi} = a + \omega^2 b = (a - b) - \omega b.$$

Hence

$$\xi + \bar{\xi} = 2a - b$$

is an algebraic integer. Since this number is rational, it follows that

$$2a - b \in \mathbb{Z}.$$

Similarly

$$\omega\xi = -b + \omega(a - b)$$

is an algebraic integer, and so by the previous argument

$$-2b - (a - b) = -a - b \in Z.$$

We deduce that

$$3a, 3b \in \mathbb{Z};$$

say

$$a = \frac{r}{3}, \ b = \frac{s}{3},$$

where $r, s \in \mathbb{Z}$.

But we also know that

$$N(\xi) = \xi\bar{\xi} = a^2 - ab + b^2$$

is an algebraic integer, and therefore a rational integer. This means that

$$r^2 - rs + s^2 = 0 \bmod 9.$$

It is readily verified that this is only soluble if $r, s = 0 \bmod 3$, ie if $a, b \in \mathbb{Z}$.
◄

C.5 Units in $\mathbb{Z}[\omega]$

Proposition C.6 *There are just 6 units in $\mathbb{Z}[\omega]$:*

$$\pm 1, \pm\omega, \pm\omega^2.$$

Proof ► Suppose ϵ is a unit. Then

$$N(\epsilon)N(\epsilon^{-1}) = 1.$$

It follows that

$$N(\epsilon) = 1.$$

Conversely, if $N(\epsilon) = 1$ then ϵ is a unit, since

$$N(\epsilon) = \epsilon\bar{\epsilon} = 1 \Longrightarrow \epsilon^{-1} = \bar{\epsilon} \in \mathbb{Z}[\omega].$$

Thus we have to find all $\epsilon = a + \omega b$ with $a, b \in \mathbb{Z}$ satisfying

$$N(\epsilon) = a^2 - ab + b^2 = 1.$$

This equation can be re-written:

$$(2a - b)^2 + 3b^2 = 4.$$

Evidently $b = 0, \pm 1$. It is a trivial matter to consider these cases separately, and deduce that the only solutions are the 6 listed above. ◄

We say that $\pi \in \mathbb{Z}[\omega]$ is a *prime* if for every factorisation

$$\pi = \alpha\beta \quad (\alpha, \beta \in \mathbb{Z}[\omega])$$

either α or β is a unit.

If π is a prime then so is $\epsilon\pi$ for any unit ϵ. Two primes that differ only by a unit factor are said to be *equivalent*, and we write

$$\pi \equiv \pi' = \epsilon\pi.$$

In general, we do not distinguish between equivalent primes.

C.6 Unique Factorisation in $\mathbb{Z}[\omega]$

Let us recall the main steps in the proof of unique factorisation in $\mathbb{Z}$ (or $\mathbb{N}$):

Division with Remainder Suppose $a, b \in Z$, with $b \neq 0$. Then we can find $q \in Z$ such that

$$a = bq + r,$$

where

$$|r| < |b|.$$

The Euclidean Algorithm This is a procedure for determining the greatest common divisor $\gcd(a, b) = d$ of $a, b \in \mathbb{Z}$. We start by dividing a by b:

$$a = q_1 b + r_1,$$

where $|r_1| < |b|$. Now we divide b by the remainder r_1:

$$b = q_2 r_1 + r_2,$$

where $|r_2| < |r_1|$. We continue in this way, successively dividing remainders:

$$\begin{aligned} r_1 &= q_3 r_2 + r_3, \\ r_2 &= q_4 r_3 + r_4, \\ &\dots \end{aligned}$$

At some point, the process must terminate when an exact division occurs (with zero remainder):

$$r_{n-1} = q_{n+1} r_n.$$

For the remainders have been getting steadily smaller:

$$|b| > |r_1| > |r_2| > \dots$$

and so must ultimately vanish.

The last non-zero remainder is the sought-for gcd:

$$d = \gcd(a, b) = r_n.$$

For $d \mid r_{n-1}$, from the last line of the algorithm. Hence $d \mid r_{n-2}$ from the previous line; and so, working up the algorithm,

$$d \mid r_{n-3}, r_{n-4}, \dots, r_1, b, a.$$

On the other hand, if $e \mid a, b$ then working down the algorithm,

$$e \mid a, b, r_1, r_2, \dots, r_n.$$

Thus

$$e \mid a, b \Longrightarrow e \mid d.$$

$au + bv = d$ The Euclidean Algorithm has one important consequence that is not immediately obvious. Let us say that e is *expressed linearly in terms of* c, d if we have an expression

$$e = cx + dy$$

with $x, y \in Z$.

The last line but one of the algorithm expresses $d = r_n$ linearly in terms of r_{n-1} and r_{n-2}, say

$$d = r_{n-1}x_1 + r_{n-2}y_1.$$

The previous line expresses r_{n-1} in terms of r_{n-2} and r_{n-3}, allowing us to express d linearly in terms of r_{n-2} and r_{n-3}, say

$$d = r_{n-2}x_2 + r_{n-3}y_2.$$

Continuing in this way, we obtain expressions

$$\begin{aligned} d &= r_{n-3}x_3 + r_{n-4}y_3 \\ \ldots \\ d &= r_2x_{n-2} + r_1y_{n-2} \\ d &= r_1x_{n-1} + byn - 1 \end{aligned}$$

and finally

$$d = bx_n + ay_n.$$

Thus d is expressed linearly in terms of a, b:

$$d = au + bv$$

for some $u, v \in Z$.

The Lemma Suppose p is a prime number. Then

$$p \mid ab \Longrightarrow p \mid a \text{ or } p \mid b.$$

We take the classic definition of a prime number: a number that has no factors other than 1 and itself. If $p \not| a$ then $\gcd(p, a) = 1$, and so by the Euclidean Algorithm we can find $u, v \in \mathbb{Z}$ such that

$$pu + av = 1.$$

Similarly if $p \not| b$ then we can find $x, y \in Z$ such that

$$px + by = 1.$$

Multiplying these relations together

$$\begin{aligned} 1 &= (pu + av)(px + by) \\ &= p(puv + uby + avx) + abvy \end{aligned}$$

Now if $p \mid ab$ then p divides all the terms on the right, and we deduce that $p \mid 1$, which is absurd.

Unique Factorisation Firstly, we can prove by induction that any $n \in \mathbb{N}$ is expressible as a product of primes. For if n is not prime then we can write $n = ab$, where $|a|, |b| < |n|$. By our inductive hypothesis we can express a, b as products of primes; and these combine to give such an expression for n.

We can prove by induction on n that this expression is unique up to order. For suppose

$$n = p_1^{e_1} \ldots p_r^{e_r} = q_1^{f_1} \ldots q_s^{f_s}.$$

By repeated use of the lemma above, the first factor p_1 on the left must occur on the right. Dividing both sides by p_1, we can apply the inductive hypothesis to show that the the factors, with one p_1 removed, are the same up to order. Hence they are the same with the p_1 restored to both sides.

Now we see that the entire argument rests upon Division with Remainder. Wherever this exists we will have unique factorisation.

One place where this holds is the ring $k[x]$ of polynomials over a field k, since we can divide one polynomial by another,

$$f(x) = g(x)q(x) + r(x),$$

leaving a remainder $r(x)$ *of lower degree* than $g(x)$. It follows by our argument that there is unique factorisation into prime (or irreducible) polynomials in $k[x]$. Note that the degree in this case plays the rôle of the absolute value $|n|$ in the case of $\mathbb{Z}$ above. The essential point is that it must be a positive integer, to ensure that our reduction process ends.

Proposition C.7 *Given $\alpha, \beta \in \mathbb{Z}[\omega]$ (with $\beta \neq 0$), we can find $\gamma, \delta \in \mathbb{Z}[\omega]$ such that*

$$\alpha = \beta\gamma + \delta,$$

where

$$N(\delta) < N(\beta).$$

Proof ► We can certainly divide α by β in $\mathbb{Q}(\omega)$, say

$$\frac{\alpha}{\beta} = r + \omega s \quad (r, s \in \mathbb{Z}).$$

Now let us choose m, n to be the nearest integers to r, s, so that

$$|r - m| \le \frac{1}{2},\ |s - n| \le \frac{1}{2}.$$

Set

$$\gamma = m + \omega n \in \mathbb{Z}[\omega];$$

and let

$$\theta = (r - m) + \omega(s - n) \in \mathbb{Q}(\omega).$$

Then

$$\begin{aligned} N(\theta) &= (r-m)^2 - (r-m)(s-n) + (s-n)^2 \\ &\le \frac{1}{4} + \frac{1}{4} + \frac{1}{4} < 1, \end{aligned}$$

and so

$$\alpha = \beta\gamma + \delta,$$

where

$$\delta = \gamma\theta,$$

and

$$N(\delta) = \mathbb{N}(\gamma)N(\theta) < N(\gamma).$$

◄

Corollary *There is unique factorisation into primes (up to equivalence and order) in* $\mathbb{Z}[\omega]$.

C.7 Fermat's Last Theorem in $\mathbb{Z}[\omega]$

It is convenient to take Fermat's equation (for $n = 3$) in the symmetric form

$$x^3 + y^3 + z^3 = 0.$$

Suppose first (x, y, z) is a solution in $\mathbb{Z}$. As usual we assume that $\gcd(x, y, z) = 1$.

Suppose that $x = 1 \bmod 3$, say $x = 1 + 3a$. Then

$$\begin{aligned} x^3 &= (1+3a)^3 \\ &= 1 + 3^2 a + 3^3 a^2 + 3^3 a^3 \\ &= 1 \bmod 3^2. \end{aligned}$$

Similarly

$$x = -1 \bmod 3 \Longrightarrow x^3 = -1 \bmod 3^2.$$

It follows that one (and just one) of x, y, z must be divisible by 3, since otherwise we would have an impossible congruence

$$\pm 1 \pm 1 \pm 1 = 0 \bmod 3^2.$$

Our aim is to extend this idea to solutions in $\mathbb{Z}[\omega]$, with the prime Π playing the rôle of 3 (recalling that $\Pi^2 \equiv 3$).

We note in the first place that there are just 3 residue classes in $\mathbb{Z}[\omega]$ modulo Π, representated by $0, 1$, and -1. (For the number of residues modulo α is $N(\alpha)$, and $N(\Pi) = 3$.)

Lemma *If* $x = 1 \bmod \Pi$ *then*

$$x^3 = 1 \bmod \Pi^4.$$

Proof ► Suppose

$$x = 1 + \Pi\alpha.$$

Then

$$\begin{aligned} x^3 &= (1 + \Pi\alpha)^3 \\ &= 1 + 3\Pi\alpha + 3\Pi^2\alpha^2 + \Pi^3\alpha^3 \\ &= 1 - \omega^2\Pi^3\alpha + \Pi^3\alpha^3 \bmod \Pi^4, \end{aligned}$$

since $3 = -\omega^2\Pi^2$, while $\Pi^4 \mid 3\Pi^2$. Thus

$$\begin{aligned} x^3 - 1 &= \alpha(-\omega^2 + \alpha^2)\Pi^3 \bmod \Pi^4 \\ &= \alpha(\alpha + \omega)(\alpha - \omega)\Pi^3 \bmod \Pi^4. \end{aligned}$$

Now $0, \omega, -\omega$ are in the 3 different residue classes modulo Π; and so therefore are $\alpha, \alpha + \omega, \alpha - \omega$. It follows that just one of these must be divisible by Π; and so

$$x^3 = 1 \bmod \Pi^4.$$

◄

Corollary *If $x = -1 \bmod \Pi$ then*

$$x^3 = -1 \bmod \Pi^4.$$

This follows at once from the lemma on replacing x by $-x$.

Let us turn to Fermat's equation

$$x^3 + y^3 + z^3 = 0,$$

where we are now looking for solutions in $\mathbb{Z}[\omega]$ (although this will, of course, include solutions in $\mathbb{Z}$). We assume as usual that $\gcd(x, y, z) = 1$.

One of x, y, z must be divisible by Π. For otherwise, by the Lemma and Corollary above, we will have an impossible congruence

$$\pm 1 \pm 1 \pm 1 = 0 \bmod \Pi^4.$$

In fact we can go further; one of x, y, z must be divisible by Π^2. For otherwise we would have

$$\Pi^3 \alpha^3 \pm 1 \pm 1 = 0 \bmod \Pi^4,$$

where $\Pi \not| \alpha$.

We may thus suppose that $x = \Pi^2 x'$, so that

$$\begin{aligned} \Pi^6 x'^3 &= -(y^3 + z^3) \\ &= -(y + z)(y + \omega z)(y + \omega^2 z). \end{aligned}$$

How can the prime-power Π^6 be distributed among the 3 factors on the right? Evidently one factor must be divisible by Π^2 at least. On replacing z by ωz or $\omega^2 z$, if necessary, we may assume that $\Pi^2 \mid (y + z)$. But

$$(y + \omega z) - (y + z) = (\omega - 1)z \equiv \Pi z.$$

Thus

$$\Pi^2 \mid y + z \Longrightarrow \Pi \parallel y + \omega z,$$

where $\pi^e \parallel \alpha$ means that $\pi^e \mid \alpha$ but $\pi^{e+1} \not| \alpha$. Similarly

$$\Pi^2 \mid y + z \Longrightarrow \Pi \parallel y + \omega^2 z.$$

It follows that

$$\Pi^4 \mid y + z, \ \Pi \parallel y + \omega z, \ \Pi \parallel y + \omega^2 z.$$

Thus it follows from unique factorisation that

$$y + z \equiv \Pi^4 X^3, \ y + \omega z \equiv \Pi Y^3, \ y + \omega^2 z \equiv \Pi Z^3,$$

where $\gcd(\Pi X, Y, Z) = 1$. But

$$(y + z) + \omega(y + \omega z) + \omega^2(y + \omega^2 z) = 0.$$

This yields a relation of the form

$$\epsilon_1 \Pi^3 X^3 + \epsilon_2 Y^3 + \epsilon_3 Z^3 = 0,$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are units, and $\gcd(\Pi X, Y, Z) = 1$. We can assume that $\epsilon_2 = 1$. Since $\Pi \not| Y, Z$, we have $Y^3, Z^3 = \pm 1 \bmod \Pi^3$. Thus

$$\pm 1 \pm \epsilon_3 = 0 \bmod \Pi^3.$$

This congruence can only be satisfied if $\epsilon_3 = \pm 1$. After replacing Z by $-Z$ if required, we may therefore assume that $\epsilon_3 = 1$. Thus the equation reads

$$\epsilon \Pi^3 X^3 + Y^3 + Z^3 = 0.$$

Proposition C.8 *The equation*

$$\epsilon \Pi^3 x^3 + y^3 + z^3 = 0$$

has no solution (x, y, z) in $\mathbb{Z}[\omega]$ with $\gcd(\Pi x, y, z) = 1$ for any unit ϵ.

Proof ▶ Since $\Pi \not| y, z$,

$$y^3, \ z^3 = \pm 1 \bmod \Pi^4.$$

Thus

$$\epsilon \Pi^3 x^3 \pm 1 \pm 1 = 0 \bmod \Pi^4.$$

The only way this congruence can be satisfied is if $\Pi \mid x$, say $x = \Pi x'$. Then

$$\begin{aligned} \epsilon \Pi^6 x'^3 &= -(y^3 + z^3) \\ &= -(y + z)(y + \omega z)(y + \omega^2 z). \end{aligned}$$

Our earlier argument still holds — the introduction of the unit ϵ makes no difference. After replacing z by ωz or $\omega^2 z$, if necessary, we have

$$y + z \equiv \Pi^4 X^3, \quad y + \omega z \equiv \Pi Y^3, \quad y + \omega^2 z \equiv \Pi Z^3,$$

where $\gcd(\Pi X, Y, Z) = 1$. As before, we deduce that

$$\epsilon_1 \Pi^3 X^3 + \epsilon_2 Y^3 + \epsilon_3 Z^3 = 0,$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are units. Dividing by ϵ_2 we have

$$\epsilon \Pi^3 X^3 + Y^3 + \epsilon' Z^3 = 0.$$

This is only soluble modulo Π^3 if $\epsilon' = \pm 1$; and we may assume that $\epsilon' = 1$, on replacing Z by $-Z$ if necessary. Thus we are led to a new solution of our equation

$$\epsilon \Pi^3 X^3 + Y^3 + Z^3 = 0,$$

with $\gcd(\Pi X, Y, Z) = 1$.

It remains to show that this solution is 'smaller', in some sense, than the first. To this end, note that

$$x = \Pi XYZ.$$

Thus

$$N(x) = 3\mathbb{N}(X)N(Y)N(Z),$$

and so

$$\max(N(x), N(y), N(z)) > \max(N(X), N(Y), N(Z)).$$

◀

Corollary *Fermat's Last Theorem holds for $n = 3$.*

Appendix H

Elliptic Curve Factorisation

Lenstra's Elliptic Curve Factorisation (ECF) technique is an analogue of Pollard's so-called '$p-1$ method', in which the group $\mathbb{Z}/p)^\times$ is replaced by the group on an elliptic curve $\mathcal{E}(\mathcal{F}_p)$ over a finite field. So we start by describing Pollard's method.

H.1 The Pollard "$p-1$ method"

We want to factorise a large number n.

It is a straightforward matter to determine whether n is prime, using the Miller-Rabin algorithm. We may therefore suppose that n is composite.

Suppose p is a prime factor of n. By Fermat's Little Theorem, if $p \nmid a$ then

$$a^{p-1} \equiv 1 \bmod p.$$

Hence

$$a^k \equiv 1 \bmod p$$

if $p - 1 \mid k$.

It follows that

$$d = \gcd(a^k - 1, n) > 1$$

since p is a factor of both numbers.

It would be very bad luck if we found a factor d of n in this way, and then discovered that $d = n$. We may therefore suppose in this case that we have a proper factor of n.

But how do we choose k? We make the assumption at this point that the prime-factors of $p - 1$ are all (relatively) small.

H.2 Elliptic curve factorisation

Let n, as before, be a large composite integer that we wish to factorise.

Suppose p is a prime factor of n. Let

$$\mathcal{E}(\mathbb{Q}) : y^2 = x^3 + bx + c \qquad (b, c \in \mathbb{Z})$$

be an elliptic curve over $\mathbb{Q}$. Unless we are very unlucky (or very lucky) p will be a good prime for $\mathcal{E}$, ie the curve

$$\mathcal{E}(\mathcal{F}_p) : y^2 = x^3 + bx + c$$

over the finite field $\mathcal{F}_p$ is still elliptic. (We say lucky because p is a bad prime if and only if

$$p \mid \Delta = -(4b^3 + 27c^2).$$

Thus if p is a bad prime,

$$d = \gcd(\Delta, n) > 1;$$

so if we wished we could compute this gcd at the outset. However, the probability of p being bad is so small that this is probably not worth considering.)

Suppose the curve $\mathcal{E}(\mathcal{F}_p)$ contains N points. By Hasse's Theorem,

$$p + 1 - 2\sqrt{p} < N < p + 1 + 2\sqrt{p}.$$

Suppose N is b-smooth. As before, let

$$k = \prod_{q \le b} q^{e(q)}.$$

Then

$$N \mid k.$$

Suppose $P \in \mathcal{E}(\mathbb{Q})$. We express P in homogeneous coordinates:

$$P = [X, Y, Z],$$

where $X, Y, Z \in \mathbb{Z}$.

It is a straightforward matter to find a formula for the sum of two points:

$$[X_1, Y_1, Z_1] + [X_2, Y_2, Z_2] = [X_3, Y_3, Z_3],$$

where X_3, Y_3, Z_3 are polynomials in $X_1, Y_1, Z_1, X_2, Y_2, Z_2$ with integer coefficients:

$$X_3, Y_3, Z_3 \in \mathbb{Z}[X_1, Y_1, Z_1, X_2, Y_2, Z_2].$$

In effect, we simply have to dress up our usual computation

$$x_1 + x_2 + x_3 = m^2, \qquad y_3 = mx_3 + c$$

in homogeneous form.

As a special case, this gives a formula for the double of a point:

$$2[X, Y, Z] = [X_1, Y_1, Z_1],$$

where X_1, Y_1, Z_1 are polynomials over $\mathbb{Z}$ in X, Y, Z.

Using these formulae we can compute

$$rP = [X_r, Y_r, Z_r]$$

for any $r \in \mathbb{N}$.

Now let

$$P_p = [X \bmod p, Y \bmod p, Z \bmod p]$$

be the point of $\mathcal{E}(\mathcal{F}_p)$ corresponding to $P \in \mathcal{E}(\mathbb{Q})$. By Lagrange's Theorem,

$$NP_p = 0,$$

and therefore

$$kP_p = 0.$$

But kP_p is just the point we get from

$$kP = [X_k, Y_k, Z_k]$$

by reduction mod$p$. It follows that

$$Z_k \equiv 0 \bmod p.$$

(We also have $X_k \equiv 0 \bmod p$. However, this follows from the result for Z_k since the only point of $\mathcal{E}(\mathcal{F}_p)$ on the line at infinity $Z = 0$ is $O = [0, 1, 0]$.)

It follows that

$$d = \gcd(Z_k, n) > 1;$$

and unless we are very unlucky this will give us a proper factor of n.

Note that in constructing Z_k for this purpose we can work throughout mod$n$.

This method has one very large advantage over Pollard's $p - 1$ method; by changing the coefficients b, c in the elliptic curve we change N, which probably ranges at random over the interval $(p+1-2\sqrt{p}, p+1+2\sqrt{p})$. This allows us many chances of finding a 'smooth' N, while Pollard's method only gives us the one chance $p-1$.

Analysis shows that if we have some idea of the size of p then it pays to choose b of order $\sqrt{p}$, and move on to another elliptic curve if this fails.

Incidentally, it is easier to choose the point $P=[X,Y,Z]$ first, and then find b,c so that the elliptic curve contains this point, rather than choosing the curve and then looking for a rational point on it.