Chapter 2

The Basic Theory

2.1 Weierstrass Equations

For most situations in this book, an **elliptic curve** E is the graph of an equation of the form

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

where A and B are constants. This will be referred to as the **Weierstrass equation** for an elliptic curve. We will need to specify what set A, B, x, and y belong to. Usually, they will be taken to be elements of a field, for example, the real numbers $\mathbf{R}$, the complex numbers $\mathbf{C}$, the rational numbers $\mathbf{Q}$, one of the finite fields $\mathbf{F}_p\,(= \mathbf{Z}_p)$ for a prime p, or one of the finite fields $\mathbf{F}_q$, where $q = p^k$ with $k \geq 1$. In fact, for almost all of this book, the reader who is not familiar with fields may assume that a field means one of the fields just listed. If K is a field with $A, B \in K$, then we say that E **is defined over** K. Throughout this book, E and K will implicitly be assumed to denote an elliptic curve and a field over which E is defined.

If we want to consider points with coordinates in some field $L \supseteq K$, we write $E(L)$. By definition, this set always contains the point ∞ defined later in this section:

$$E(L) = \{\infty\} \cup \left\{(x, y) \in L \times L \,|\, y^2 = x^3 + Ax + B\right\}.$$

It is not possible to draw meaningful pictures of elliptic curves over most fields. However, for intuition, it is useful to think in terms of graphs over the real numbers. These have two basic forms, depicted in Figure 2.1.

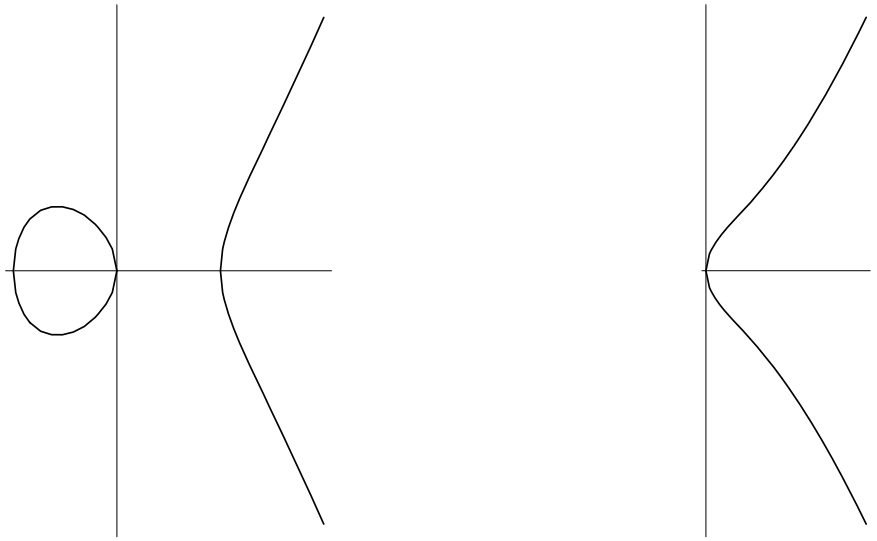
The cubic $y^2 = x^3 - x$ in the first case has three distinct real roots. In the second case, the cubic $y^2 = x^3 + x$ has only one real root.

What happens if there is a multiple root? We don't allow this. Namely, we assume that

$$4A^3 + 27B^2 \neq 0.$$

If the roots of the cubic are r_1, r_2, r_3, then it can be shown that the discriminant of the cubic is

$$((r_1 - r_2)(r_1 - r_3)(r_2 - r_3))^2 = -(4A^3 + 27B^2).$$



(a) $y^2 = x^3 - x$ (b) $y^2 = x^3 + x$

Figure 2.1

Therefore, the roots of the cubic must be distinct. However, the case where the roots are not distinct is still interesting and will be discussed in Section 2.10.

In order to have a little more flexibility, we also allow somewhat more general equations of the form

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \tag{2.1}$$

where $a_1, \ldots, a_6$ are constants. This more general form (we'll call it the **generalized Weierstrass equation**) is useful when working with fields of characteristic 2 and characteristic 3. If the characteristic of the field is not 2, then we can divide by 2 and complete the square:

$$\left(y + \frac{a_1x}{2} + \frac{a_3}{2}\right)^2 = x^3 + \left(a_2 + \frac{a_1^2}{4}\right)x^2 + \left(a_4 + \frac{a_1a_3}{2}\right)x + \left(\frac{a_3^2}{4} + a_6\right),$$

which can be written as

$$y_1^2 = x^3 + a_2'x^2 + a_4'x + a_6',$$

with $y_1 = y + a_1x/2 + a_3/2$ and with some constants a_2', a_4', a_6'. If the characteristic is also not 3, then we can let $x_1 = x + a_2'/3$ and obtain

$$y_1^2 = x_1^3 + Ax_1 + B,$$

for some constants A, B.

In most of this book, we will develop the theory using the Weierstrass equation, occasionally pointing out what modifications need to be made in characteristics 2 and 3. In Section 2.8, we discuss the case of characteristic 2 in more detail, since the formulas for the (nongeneralized) Weierstrass equation do not apply. In contrast, these formulas are correct in characteristic 3 for curves of the form $y^2 = x^3 + Ax + B$, but there are curves that are not of this form. The general case for characteristic 3 can be obtained by using the present methods to treat curves of the form $y^2 = x^3 + Cx^2 + Ax + B$.

Finally, suppose we start with an equation

$$cy^2 = dx^3 + ax + b$$

with $c, d \neq 0$. Multiply both sides of the equation by c^3d^2 to obtain

$$(c^2dy)^2 = (cdx)^3 + (ac^2d)(cdx) + (bc^3d^2).$$

The change of variables

$$y_1 = c^2dy, \quad x_1 = cdx$$

yields an equation in Weierstrass form.

Later in this chapter, we will meet other types of equations that can be transformed into Weierstrass equations for elliptic curves. These will be useful in certain contexts.

For technical reasons, it is useful to add a **point at infinity** to an elliptic curve. In Section 2.3, this concept will be made rigorous. However, it is easiest to regard it as a point (∞, ∞), usually denoted simply by ∞, sitting at the top of the y-axis. For computational purposes, it will be a formal symbol satisfying certain computational rules. For example, a line is said to pass through ∞ exactly when this line is vertical (i.e., x =constant). The point ∞ might seem a little unnatural, but we will see that including it has very useful consequences.

We now make one more convention regarding ∞. It not only is at the top of the y-axis, it is also at the bottom of the y-axis. Namely, we think of the ends of the y-axis as wrapping around and meeting (perhaps somewhere in the back behind the page) in the point ∞. This might seem a little strange. However, if we are working with a field other than the real numbers, for example, a finite field, then there might not be any meaningful ordering of the elements and therefore distinguishing a top and a bottom of the y-axis might not make sense. In fact, in this situation, the ends of the y-axis do not have meaning until we introduce projective coordinates in Section 2.3. This is why it is best to regard ∞ as a formal symbol satisfying certain properties. Also, we have arranged that two vertical lines meet at ∞. By symmetry, if they meet at the top of the y-axis, they should also meet at the bottom. But two lines should intersect in only one point, so the "top ∞" and the "bottom ∞" need to be the same. In any case, this will be a useful property of ∞.

2.2 The Group Law

As we saw in Chapter 1, we could start with two points, or even one point, on an elliptic curve, and produce another point. We now examine this process in more detail.

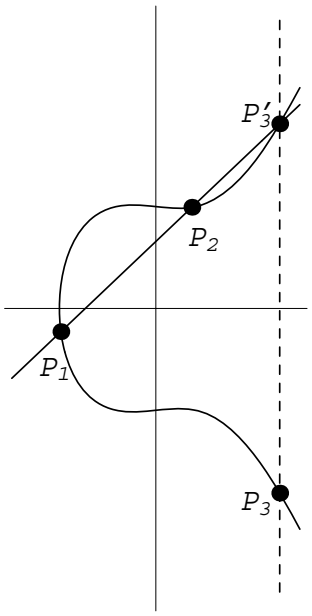


Figure 2.2
Adding Points on an Elliptic Curve

Start with two points

$$P_1 = (x_1, y_1), \quad P_2 = (x_2, y_2)$$

on an elliptic curve E given by the equation $y^2 = x^3 + Ax + B$. Define a new point P_3 as follows. Draw the line L through P_1 and P_2. We'll see below that L intersects E in a third point P_3'. Reflect P_3' across the x-axis (i.e., change the sign of the y-coordinate) to obtain P_3. We define

$$P_1 + P_2 = P_3.$$

Examples below will show that this is not the same as adding coordinates of the points. It might be better to denote this operation by $P_1 +_E P_2$, but we opt for the simpler notation since we will never be adding points by adding coordinates.

Assume first that $P_1 \neq P_2$ and that neither point is ∞. Draw the line L through P_1 and P_2. Its slope is

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

If $x_1 = x_2$, then L is vertical. We'll treat this case later, so let's assume that $x_1 \neq x_2$. The equation of L is then

$$y = m(x - x_1) + y_1.$$

To find the intersection with E, substitute to get

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

This can be rearranged to the form

$$0 = x^3 - m^2x^2 + \cdots .$$

The three roots of this cubic correspond to the three points of intersection of L with E. Generally, solving a cubic is not easy, but in the present case we already know two of the roots, namely x_1 and x_2, since P_1 and P_2 are points on both L and E. Therefore, we could factor the cubic to obtain the third value of x. But there is an easier way. As in Chapter 1, if we have a cubic polynomial $x^3 + ax^2 + bx + c$ with roots r, s, t, then

$$x^3 + ax^2 + bx + c = (x - r)(x - s)(x - t) = x^3 - (r + s + t)x^2 + \cdots .$$

Therefore,

$$r + s + t = -a.$$

If we know two roots r, s, then we can recover the third as $t = -a - r - s$.

In our case, we obtain

$$x = m^2 - x_1 - x_2$$

and

$$y = m(x - x_1) + y_1.$$

Now, reflect across the x-axis to obtain the point $P_3 = (x_3, y_3)$:

$$x_3 = m^2 - x_1 - x_2, \qquad y_3 = m(x_1 - x_3) - y_1.$$

In the case that $x_1 = x_2$ but $y_1 \neq y_2$, the line through P_1 and P_2 is a vertical line, which therefore intersects E in ∞. Reflecting ∞ across the x-axis yields the same point ∞ (this is why we put ∞ at both the top and the bottom of the y-axis). Therefore, in this case $P_1 + P_2 = \infty$.

Now consider the case where $P_1 = P_2 = (x_1, y_1)$. When two points on a curve are very close to each other, the line through them approximates a tangent line. Therefore, when the two points coincide, we take the line L through them to be the tangent line. Implicit differentiation allows us to find the slope m of L:

$$2y\frac{dy}{dx} = 3x^2 + A, \quad \text{so} \quad m = \frac{dy}{dx} = \frac{3x_1^2 + A}{2y_1}.$$

If $y_1 = 0$ then the line is vertical and we set $P_1 + P_2 = \infty$, as before. (Technical point: if $y_1 = 0$, then the numerator $3x_1^2 + A \neq 0$. See Exercise 2.5.) Therefore, assume that $y_1 \neq 0$. The equation of L is

$$y = m(x - x_1) + y_1,$$

as before. We obtain the cubic equation

$$0 = x^3 - m^2x^2 + \cdots .$$

This time, we know only one root, namely x_1, but it is a double root since L is tangent to E at P_1. Therefore, proceeding as before, we obtain

$$x_3 = m^2 - 2x_1, \qquad y_3 = m(x_1 - x_3) - y_1.$$

Finally, suppose $P_2 = \infty$. The line through P_1 and ∞ is a vertical line that intersects E in the point P_1' that is the reflection of P_1 across the x-axis. When we reflect P_1' across the x-axis to get $P_3 = P_1 + P_2$, we are back at P_1. Therefore

$$P_1 + \infty = P_1$$

for all points P_1 on E. Of course, we extend this to include $\infty + \infty = \infty$.

Let's summarize the above discussion:

GROUP LAW

Let E be an elliptic curve defined by $y^2 = x^3 + Ax + B$. Let $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ be points on E with $P_1, P_2 \neq \infty$. Define $P_1 + P_2 = P_3 = (x_3, y_3)$ as follows:

1. If $x_1 \neq x_2$, then

$$x_3 = m^2 - x_1 - x_2, \qquad y_3 = m(x_1 - x_3) - y_1, \qquad \text{where } m = \frac{y_2 - y_1}{x_2 - x_1}.$$

2. If $x_1 = x_2$ but $y_1 \neq y_2$, then $P_1 + P_2 = \infty$.

3. If $P_1 = P_2$ and $y_1 \neq 0$, then

$$x_3 = m^2 - 2x_1, \qquad y_3 = m(x_1 - x_3) - y_1, \qquad \text{where } m = \frac{3x_1^2 + A}{2y_1}.$$

4. If $P_1 = P_2$ and $y_1 = 0$, then $P_1 + P_2 = \infty$.

Moreover, define

$$P + \infty = P$$

for all points P on E.

Note that when P_1 and P_2 have coordinates in a field L that contains A and B, then $P_1 + P_2$ also has coordinates in L. Therefore $E(L)$ is closed under the above addition of points.

This addition of points might seem a little unnatural. Later (in Chapters 9 and 11), we'll interpret it as corresponding to some very natural operations, but, for the present, let's show that it has some nice properties.

THEOREM 2.1

The addition of points on an elliptic curve E satisfies the following properties:

1. (commutativity) $P_1 + P_2 = P_2 + P_1$ for all P_1, P_2 on E.
2. (existence of identity) $P + \infty = P$ for all points P on E.
3. (existence of inverses) Given P on E, there exists P' on E with $P + P' = \infty$. This point P' will usually be denoted $-P$.
4. (associativity) $(P_1 + P_2) + P_3 = P_1 + (P_2 + P_3)$ for all P_1, P_2, P_3 on E.

In other words, the points on E form an additive abelian group with ∞ as the identity element.

PROOF The commutativity is obvious, either from the formulas or from the fact that the line through P_1 and P_2 is the same as the line through P_2 and P_1. The identity property of ∞ holds by definition. For inverses, let P' be the reflection of P across the x-axis. Then $P + P' = \infty$.

Finally, we need to prove associativity. This is by far the most subtle and nonobvious property of the addition of points on E. It is possible to define many laws of composition satisfying (1), (2), (3) for points on E, either simpler or more complicated than the one being considered. But it is very unlikely that such a law will be associative. In fact, it is rather surprising that the law of composition that we have defined is associative. After all, we start with two points P_1 and P_2 and perform a certain procedure to obtain a third point $P_1 + P_2$. Then we repeat the procedure with $P_1 + P_2$ and P_3 to obtain $(P_1 + P_2) + P_3$. If we instead start by adding P_2 and P_3, then computing $P_1 + (P_2 + P_3)$, there seems to be no obvious reason that this should give the same point as the other computation.

The associative law can be verified by calculation with the formulas. There are several cases, depending on whether or not $P_1 = P_2$, and whether or not $P_3 = (P_1 + P_2)$, etc., and this makes the proof rather messy. However, we prefer a different approach, which we give in Section 2.4. ∎

Warning: For the Weierstrass equation, if $P = (x, y)$, then $-P = (x, -y)$. For the generalized Weierstrass equation (2.1), this is no longer the case. If $P = (x, y)$ is on the curve described by (2.1), then (see Exercise 2.9)

$$-P = (x,\ -a_1x - a_3 - y).$$

Example 2.1
The calculations of Chapter 1 can now be interpreted as adding points on elliptic curves. On the curve

$$y^2 = \frac{x(x+1)(2x+1)}{6},$$

we have

$$(0,0) + (1,1) = (\frac{1}{2}, -\frac{1}{2}), \quad (\frac{1}{2}, -\frac{1}{2}) + (1,1) = (24, -70).$$

On the curve

$$y^2 = x^3 - 25x,$$

we have

$$2(-4,6) = (-4,6) + (-4,6) = \left(\frac{1681}{144}, -\frac{62279}{1728}\right).$$

We also have

$$(0,0) + (-5,0) = (5,0), \quad 2(0,0) = 2(-5,0) = 2(5,0) = \infty.$$

□

The fact that the points on an elliptic curve form an abelian group is behind most of the interesting properties and applications. The question arises: what can we say about the groups of points that we obtain? Here are some examples.

1. An elliptic curve over a finite field has only finitely many points with coordinates in that finite field. Therefore, we obtain a finite abelian group in this case. Properties of such groups, and applications to cryptography, will be discussed in later chapters.

2. If E is an elliptic curve defined over $\mathbf{Q}$, then $E(\mathbf{Q})$ is a finitely generated abelian group. This is the Mordell-Weil theorem, which we prove in Chapter 8. Such a group is isomorphic to $\mathbf{Z}^r \oplus F$ for some $r \geq 0$ and some finite group F. The integer r is called the **rank** of $E(\mathbf{Q})$. Determining r is fairly difficult in general. It is not known whether r can be arbitrarily large. At present, there are elliptic curves known with rank at least 28. The finite group F is easy to compute using the Lutz-Nagell theorem of Chapter 8. Moreover, a deep theorem of Mazur says that there are only finitely many possibilities for F, as E ranges over all elliptic curves defined over $\mathbf{Q}$.

3. An elliptic curve over the complex numbers $\mathbf{C}$ is isomorphic to a torus. This will be proved in Chapter 9. The usual way to obtain a torus is as $\mathbf{C}/\mathcal{L}$, where $\mathcal{L}$ is a lattice in $\mathbf{C}$. The usual addition of complex numbers induces a group law on $\mathbf{C}/\mathcal{L}$ that corresponds to the group law on the elliptic curve under the isomorphism between the torus and the elliptic curve.

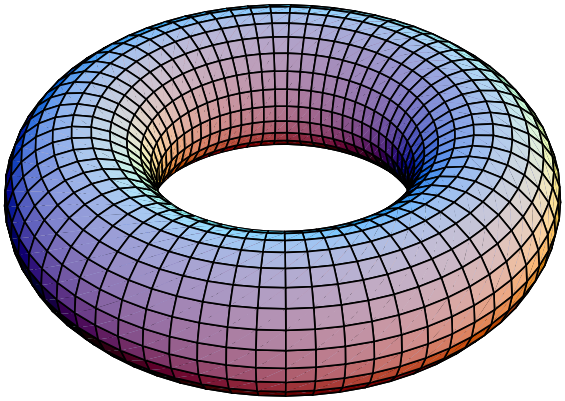


Figure 2.3
An Elliptic Curve over C

4. If E is defined over $\mathbf{R}$, then $E(\mathbf{R})$ is isomorphic to the unit circle S^1 or to $S^1 \oplus \mathbf{Z}_2$. The first case corresponds to the case where the cubic polynomial $x^3 + Ax + B$ has only one real root (think of the ends of the graph in Figure 2.1(b) as being hitched together at the point ∞ to get a loop). The second case corresponds to the case where the cubic has three real roots. The closed loop in Figure 2.1(a) is the set $S^1 \oplus \{1\}$, while the open-ended loop can be closed up using ∞ to obtain the set $S^1 \oplus \{0\}$. If we have an elliptic curve E defined over $\mathbf{R}$, then we can consider its complex points $E(\mathbf{C})$. These form a torus, as in (3) above. The real points $E(\mathbf{R})$ are obtained by intersecting the torus with a plane. If the plane passes through the hole in the middle, we obtain a curve as in Figure 2.1(a). If it does not pass through the hole, we obtain a curve as in Figure 2.1(b) (see Section 9.3).

If P is a point on an elliptic curve and k is a positive integer, then kP denotes $P + P + \cdots + P$ (with k summands). If $k < 0$, then $kP = (-P) + (-P) + \cdots (-P)$, with $|k|$ summands. To compute kP for a large integer k, it is inefficient to add P to itself repeatedly. It is much faster to use **successive doubling**. For example, to compute $19P$, we compute

$$2P, \quad 4P = 2P+2P, \quad 8P = 4P+4P, \quad 16P = 8P+8P, \quad 19P = 16P+2P+P.$$

This method allows us to compute kP for very large k, say of several hundred digits, very quickly. The only difficulty is that the size of the coordinates of the points increases very rapidly if we are working over the rational numbers (see Theorem 8.18). However, when we are working over a finite field, for example $\mathbf{F}_p$, this is not a problem because we can continually reduce mod p and thus keep the numbers involved relatively small. Note that the associative

law allows us to make these computations without worrying about what order we use to combine the summands.

The method of successive doubling can be stated in general as follows:

INTEGER TIMES A POINT

Let k be a positive integer and let P be a point on an elliptic curve. The following procedure computes kP.

1. Start with $a = k$, $B = \infty$, $C = P$.
2. If a is even, let $a = a/2$, and let $B = B$, $C = 2C$.
3. If a is odd, let $a = a - 1$, and let $B = B + C$, $C = C$.
4. If $a \neq 0$, go to step 2.
5. Output B.

The output B is kP (see Exercise 2.8).

On the other hand, if we are working over a large finite field and are given points P and kP, it is very difficult to determine the value of k. This is called the **discrete logarithm problem** for elliptic curves and is the basis for the cryptographic applications that will be discussed in Chapter 6.

2.3 Projective Space and the Point at Infinity

We all know that parallel lines meet at infinity. Projective space allows us to make sense out of this statement and also to interpret the point at infinity on an elliptic curve.

Let K be a field. Two-dimensional **projective space** $\mathbf{P}^2_K$ over K is given by equivalence classes of triples (x, y, z) with $x, y, z \in K$ and at least one of x, y, z nonzero. Two triples (x_1, y_1, z_1) and (x_2, y_2, z_2) are said to be **equivalent** if there exists a nonzero element $\lambda \in K$ such that

$$(x_1, y_1, z_1) = (\lambda x_2, \lambda y_2, \lambda z_2).$$

We write $(x_1, y_1, z_1) \sim (x_2, y_2, z_2)$. The equivalence class of a triple only depends on the ratios of x to y to z. Therefore, the equivalence class of (x, y, z) is denoted $(x : y : z)$.

If $(x : y : z)$ is a point with $z \neq 0$, then $(x : y : z) = (x/z : y/z : 1)$. These are the "finite" points in $\mathbf{P}^2_K$. However, if $z = 0$ then dividing by z should be thought of as giving ∞ in either the x or y coordinate, and therefore the points $(x : y : 0)$ are called the **"points at infinity"** in $\mathbf{P}^2_K$. The point at

infinity on an elliptic curve will soon be identified with one of these points at infinity in $\mathbf{P}^2_K$.

The two-dimensional **affine plane** over K is often denoted

$$\mathbf{A}^2_K = \{(x, y) \in K \times K\}.$$

We have an inclusion

$$\mathbf{A}^2_K \hookrightarrow \mathbf{P}^2_K$$

given by

$$(x, y) \mapsto (x : y : 1).$$

In this way, the affine plane is identified with the finite points in $\mathbf{P}^2_K$. Adding the points at infinity to obtain $\mathbf{P}^2_K$ can be viewed as a way of "compactifying" the plane (see Exercise 2.10).

A polynomial is **homogeneous** of degree n if it is a sum of terms of the form $ax^iy^jz^k$ with $a \in K$ and $i + j + k = n$. For example, $F(x, y, z) = 2x^3 - 5xyz + 7yz^2$ is homogeneous of degree 3. If a polynomial F is homogeneous of degree n then $F(\lambda x, \lambda y, \lambda z) = \lambda^n F(x, y, z)$ for all $\lambda \in K$. It follows that if F is homogeneous of some degree, and $(x_1, y_1, z_1) \sim (x_2, y_2, z_2)$, then $F(x_1, y_1, z_1) = 0$ if and only if $F(x_2, y_2, z_2) = 0$. Therefore, a zero of F in $\mathbf{P}^2_K$ does not depend on the choice of representative for the equivalence class, so the set of zeros of F in $\mathbf{P}^2_K$ is well defined.

If $F(x, y, z)$ is an arbitrary polynomial in x, y, z, then we cannot talk about a point in $\mathbf{P}^2_K$ where $F(x, y, z) = 0$ since this depends on the representative (x, y, z) of the equivalence class. For example, let $F(x, y, z) = x^2 + 2y - 3z$. Then $F(1, 1, 1) = 0$, so we might be tempted to say that F vanishes at $(1 : 1 : 1)$. But $F(2, 2, 2) = 2$ and $(1 : 1 : 1) = (2 : 2 : 2)$. To avoid this problem, we need to work with homogeneous polynomials.

If $f(x, y)$ is a polynomial in x and y, then we can make it homogeneous by inserting appropriate powers of z. For example, if $f(x, y) = y^2 - x^3 - Ax - B$, then we obtain the homogeneous polynomial $F(x, y, z) = y^2z - x^3 - Axz^2 - Bz^3$. If F is homogeneous of degree n then

$$F(x, y, z) = z^n f(\frac{x}{z}, \frac{y}{z})$$

and

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

We can now see what it means for two parallel lines to meet at infinity. Let

$$y = mx + b_1, \qquad y = mx + b_2$$

be two nonvertical parallel lines with $b_1 \neq b_2$. They have the homogeneous forms

$$y = mx + b_1z, \qquad y = mx + b_2z.$$

(The preceding discussion considered only equations of the form $f(x, y) = 0$ and $F(x, y, z) = 0$; however, there is nothing wrong with rearranging these equations to the form "homogeneous of degree n = homogeneous of degree n.") When we solve the simultaneous equations to find their intersection, we obtain

$$z = 0 \quad \text{and} \quad y = mx.$$

Since we cannot have all of x, y, z being 0, we must have $x \neq 0$. Therefore, we can rescale by dividing by x and find that the intersection of the two lines is

$$(x : mx : 0) = (1 : m : 0).$$

Similarly, if $x = c_1$ and $x = c_2$ are two vertical lines, they intersect in the point $(0 : 1 : 0)$. This is one of the points at infinity in $\mathbf{P}^2_K$.

Now let's look at the elliptic curve E given by $y^2 = x^3 + Ax + B$. Its homogeneous form is $y^2z = x^3 + Axz^2 + Bz^3$. The points (x, y) on the original curve correspond to the points $(x : y : 1)$ in the projective version. To see what points on E lie at infinity, set $z = 0$ and obtain $0 = x^3$. Therefore $x = 0$, and y can be any nonzero number (recall that $(0 : 0 : 0)$ is not allowed). Rescale by y to find that $(0 : y : 0) = (0 : 1 : 0)$ is the only point at infinity on E. As we saw above, $(0 : 1 : 0)$ lies on every vertical line, so every vertical line intersects E at this point at infinity. Moreover, since $(0 : 1 : 0) = (0 : -1 : 0)$, the "top" and the "bottom" of the y-axis are the same.

There are situations where using projective coordinates speeds up computations on elliptic curves (see Section 2.6). However, in this book we almost always work in affine (nonprojective) coordinates and treat the point at infinity as a special case when needed. An exception is the proof of associativity of the group law given in Section 2.4, where it will be convenient to have the point at infinity treated like any other point $(x : y : z)$.

2.4 Proof of Associativity

In this section, we prove the associativity of addition of points on an elliptic curve. The reader who is willing to believe this result may skip this section without missing anything that is needed in the rest of the book. However, as corollaries of the proof, we will obtain two results, namely the theorems of Pappus and Pascal, that are not about elliptic curves but which are interesting in their own right.

The basic idea is the following. Start with an elliptic curve E and points P, Q, R on E. To compute $-((P + Q) + R)$ we need to form the lines $\ell_1 = \overline{PQ}$, $m_2 = \overline{\infty, P + Q}$, and $\ell_3 = \overline{R, P + Q}$, and see where they intersect E. To compute $-((P + (Q + R))$ we need to form the lines $m_1 = \overline{QR}$, $\ell_2 = \overline{\infty, Q + R}$, and $m_3 = \overline{P, Q + R}$. It is easy to see that the points $P_{ij} = \ell_i \cap m_j$

lie on E, except possibly for P_{33}. We show in Theorem 2.6 that having the eight points $P_{ij} \neq P_{33}$ on E forces P_{33} to be on E. Since ℓ_3 intersects E at the points $R, P+Q, -((P+Q)+R)$, we must have $-((P+Q)+R) = P_{33}$. Similarly, $-(P+(Q+R)) = P_{33}$, so

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

which implies the desired associativity.

There are three main technicalities that must be treated. First, some of the points P_{ij} could be at infinity, so we need to use projective coordinates. Second, a line could be tangent to E, which means that two P_{ij} could be equal. Therefore, we need a careful definition of the order to which a line intersects a curve. Third, two of the lines could be equal. Dealing with these technicalities takes up most of our attention during the proof.

First, we need to discuss lines in $\mathbf{P}^2_K$. The standard way to describe a line is by a linear equation: $ax + by + cz = 0$. Sometimes it is useful to give a parametric description:

$$\begin{aligned} x &= a_1u + b_1v \\ y &= a_2u + b_2v \\ z &= a_3u + b_3v \end{aligned} \tag{2.2}$$

where u, v run through K, and at least one of u, v is nonzero. For example, if $a \neq 0$, the line

$$ax + by + cz = 0$$

can be described by

$$x = -(b/a)u - (c/a)v, y = u, z = v.$$

Suppose all the vectors (a_i, b_i) are multiples of each other, say $(a_i, b_i) = \lambda_i(a_1, b_1)$. Then $(x, y, z) = x(1, \lambda_2, \lambda_3)$ for all u, v such that $x \neq 0$. So we get a point, rather than a line, in projective space. Therefore, we need a condition on the coefficients $a_1, \ldots, b_3$ that ensure that we actually get a line. It is not hard to see that we must require the matrix

$$\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{pmatrix}$$

to have rank 2 (cf. Exercise 2.12).

If $(u_1, v_1) = \lambda(u_2, v_2)$ for some $\lambda \in K^\times$, then (u_1, v_1) and (u_2, v_2) yield equivalent triples (x, y, z). Therefore, we can regard (u, v) as running through points $(u : v)$ in 1-dimensional projective space $\mathbf{P}^1_K$. Consequently, a line corresponds to a copy of the projective line $\mathbf{P}^1_K$ embedded in the projective plane.

We need to quantify the order to which a line intersects a curve at a point. The following gets us started.

LEMMA 2.2
Let $G(u,v)$ be a nonzero homogeneous polynomial and let $(u_0 : v_0) \in \mathbf{P}^1_K$. Then there exists an integer $k \geq 0$ and a polynomial $H(u,v)$ with $H(u_0,v_0) \neq 0$ such that

$$G(u,v) = (v_0u - u_0v)^k H(u,v).$$

PROOF Suppose $v_0 \neq 0$. Let m be the degree of G. Let $g(u) = G(u,v_0)$. By factoring out as large a power of $u - u_0$ as possible, we can write $g(u) = (u-u_0)^k h(u)$ for some k and for some polynomial h of degree $m-k$ with $h(u_0) \neq 0$. Let $H(u,v) = (v^{m-k}/v_0^m)h(uv_0/v)$, so $H(u,v)$ is homogeneous of degree $m-k$. Then

$$\begin{aligned} G(u,v) = \left(\frac{v}{v_0}\right)^m g\left(\frac{uv_0}{v}\right) &= \frac{v^{m-k}}{v_0^m}\left(v_0u - u_0v\right)^k h\left(\frac{uv_0}{v}\right) \\ &= (v_0u - u_0v)^k H(u,v), \end{aligned}$$

as desired.

If $v_0 = 0$, then $u_0 \neq 0$. Reversing the roles of u and v yields the proof in this case. ∎

Let $f(x,y) = 0$ (where f is a polynomial) describe a curve C in the affine plane and let

$$x = a_1t + b_1, y = a_2t + b_2$$

be a line L written in terms of the parameter t. Let

$$\tilde{f}(t) = f(a_1t + b_1, a_2t + b_2).$$

Then L intersects C when $t = t_0$ if $\tilde{f}(t_0) = 0$. If $(t-t_0)^2$ divides $\tilde{f}(t)$, then L is tangent to C (if the point corresponding to t_0 is nonsingular. See Lemma 2.5). More generally, we say that L intersects C to order n at the point (x,y) corresponding to $t = t_0$ if $(t-t_0)^n$ is the highest power of $(t-t_0)$ that divides $\tilde{f}(t)$.

The homogeneous version of the above is the following. Let $F(x,y,z)$ be a homogeneous polynomial, so $F = 0$ describes a curve C in $\mathbf{P}^2_K$. Let L be a line given parametrically by (2.2) and let

$$\tilde{F}(u,v) = F(a_1u + b_1v,\ a_2u + b_2v,\ a_3u + b_3v).$$

We say that L **intersects** C **to order** n at the point $P = (x_0 : y_0 : z_0)$ corresponding to $(u : v) = (u_0 : v_0)$ if $(v_0u - u_0v)^n$ is the highest power of $(v_0u - u_0v)$ dividing $\tilde{F}(u,v)$. We denote this by

$$\mathrm{ord}_{L,P}(F) = n.$$

If $\tilde{F}$ is identically 0, then we let $\text{ord}_{L,P}(F) = \infty$. It is not hard to show that $\text{ord}_{L,P}(F)$ is independent of the choice of parameterization of the line L. Note that $v = v_0 = 1$ corresponds to the nonhomogeneous situation above, and the definitions coincide (at least when $z \neq 0$). The advantage of the homogeneous formulation is that it allows us to treat the points at infinity along with the finite points in a uniform manner.

LEMMA 2.3
Let L_1 and L_2 be lines intersecting in a point P, and, for $i = 1, 2$, let $L_i(x, y, z)$ be a linear polynomial defining L_i. Then $\text{ord}_{L_1,P}(L_2) = 1$ unless $L_1(x, y, z) = \alpha L_2(x, y, z)$ for some constant α, in which case $\text{ord}_{L_1,P}(L_2) = \infty$.

PROOF When we substitute the parameterization for L_1 into $L_2(x, y, z)$, we obtain $\tilde{L}_2$, which is a linear expression in u, v. Let P correspond to $(u_0 : v_0)$. Since $\tilde{L}_2(u_0, v_0) = 0$, it follows that $\tilde{L}_2(u, v) = \beta(v_0 u - u_0 v)$ for some constant β. If $\beta \neq 0$, then $\text{ord}_{L_1,P}(L_2) = 1$. If $\beta = 0$, then all points on L_1 lie on L_2. Since two points in $\mathbf{P}^2_K$ determine a line, and L_1 has at least three points ($\mathbf{P}^1_K$ always contains the points $(1 : 0), (0 : 1), (1 : 1)$), it follows that L_1 and L_2 are the same line. Therefore $L_1(x, y, z)$ is proportional to $L_2(x, y, z)$. ∎

Usually, a line that intersects a curve to order at least 2 is tangent to the curve. However, consider the curve C defined by

$$F(x, y, z) = y^2 z - x^3 = 0.$$

Let

$$x = au, \quad y = bu, \quad z = v$$

be a line through the point $P = (0 : 0 : 1)$. Note that P corresponds to $(u : v) = (0 : 1)$. We have $\tilde{F}(u, v) = u^2(b^2 v - a^3 u)$, so every line through P intersects C to order at least 2. The line with $b = 0$, which is the best choice for the tangent at P, intersects C to order 3. The affine part of C is the curve $y^2 = x^3$, which is pictured in Figure 2.7. The point $(0, 0)$ is a singularity of the curve, which is why the intersections at P have higher orders than might be expected. This is a situation we usually want to avoid.

DEFINITION 2.4 A curve C in $\mathbf{P}^2_K$ defined by $F(x, y, z) = 0$ is said to be **nonsingular** at a point P if at least one of the partial derivatives F_x, F_y, F_z is nonzero at P.

For example, consider an elliptic curve defined by $F(x, y, z) = y^2 z - x^3 - Axz^2 - Bz^3 = 0$, and assume the characteristic of our field K is not 2 or 3.

We have

$$F_x = -3x^2 - Az^2, \quad F_y = 2yz, \quad F_z = y^2 - 2Axz - 3Bz^2.$$

Suppose $P = (x : y : z)$ is a singular point. If $z = 0$, then $F_x = 0$ implies $x = 0$ and $F_z = 0$ implies $y = 0$, so $P = (0 : 0 : 0)$, which is impossible. Therefore $z \neq 0$, so we may take $z = 1$ (and therefore ignore it). If $F_y = 0$, then $y = 0$. Since $(x : y : 1)$ lies on the curve, x must satisfy $x^3 + Ax + B = 0$. If $F_x = -(3x^2 + A) = 0$, then x is a root of a polynomial and a root of its derivative, hence a double root. Since we assumed that the cubic polynomial has no multiple roots, we have a contradiction. Therefore an elliptic curve has no singular points. Note that this is true even if we are considering points with coordinates in $\overline{K}$ (= algebraic closure of K). In general, by a **nonsingular curve** we mean a curve with no singular points in $\overline{K}$.

If we allow the cubic polynomial to have a multiple root x, then it is easy to see that the curve has a singularity at $(x : 0 : 1)$. This case will be discussed in Section 2.10.

If P is a nonsingular point of a curve $F(x, y, z) = 0$, then the tangent line at P is

$$F_x(P)x + F_y(P)y + F_z(P)z = 0.$$

For example, if $F(x, y, z) = y^2z - x^3 - Axz^2 - Bz^3 = 0$, then the **tangent line** at $(x_0 : y_0 : z_0)$ is

$$(-3x_0^2 - Az_0^2)x + 2y_0z_0y + (y_0^2 - 2Ax_0z_0 - 3Bz_0^2)z = 0.$$

If we set $z_0 = z = 1$, then we obtain

$$(-3x_0^2 - A)x + 2y_0y + (y_0^2 - 2Ax_0 - 3B) = 0.$$

Using the fact that $y_0^2 = x_0^3 + Ax_0 + B$, we can rewrite this as

$$(-3x_0^2 - A)(x - x_0) + 2y_0(y - y_0) = 0.$$

This is the tangent line in affine coordinates that we used in obtaining the formulas for adding a point to itself on an elliptic curve. Now let's look at the point at infinity on this curve. We have $(x_0 : y_0 : z_0) = (0 : 1 : 0)$. The tangent line is given by $0x + 0y + z = 0$, which is the "line at infinity" in $\mathbf{P}^2_K$. It intersects the elliptic curve only in the point $(0 : 1 : 0)$. This corresponds to the fact that $\infty + \infty = \infty$ on an elliptic curve.

LEMMA 2.5

Let $F(x, y, z) = 0$ define a curve C. If P is a nonsingular point of C, then there is exactly one line in $\mathbf{P}^2_K$ that intersects C to order at least 2, and it is the tangent to C at P.

PROOF Let L be a line intersecting C to order $k \geq 1$. Parameterize L by (2.2) and substitute into F. This yields $\tilde{F}(u, v)$. Let $(u_0 : v_0)$ correspond

to P. Then $\tilde{F} = (v_0u - u_0v)^k H(u,v)$ for some $H(u,v)$ with $H(u_0,v_0) \neq 0$. Therefore,

$$\tilde{F}_u(u,v) = kv_0(v_0u - u_0v)^{k-1}H(u,v) + (v_0u - u_0v)^k H_u(u,v)$$

and

$$\tilde{F}_v(u,v) = -ku_0(v_0u - u_0v)^{k-1}H(u,v) + (v_0u - u_0v)^k H_v(u,v).$$

It follows that $k \geq 2$ if and only if $\tilde{F}_u(u_0,v_0) = \tilde{F}_v(u_0,v_0) = 0$.

Suppose $k \geq 2$. The chain rule yields

$$\tilde{F}_u = a_1F_x + a_2F_y + a_3F_z = 0, \quad \tilde{F}_v = b_1F_x + b_2F_y + b_3F_z = 0 \tag{2.3}$$

at P. Recall that since the parameterization (2.2) yields a line, the vectors (a_1,a_2,a_3) and (b_1,b_2,b_3) must be linearly independent.

Suppose L' is another line that intersects C to order at least 2. Then we obtain another set of equations

$$a_1'F_x + a_2'F_y + a_3'F_z = 0, \quad b_1'F_x + b_2'F_y + b_3'F_z = 0$$

at P.

If the vectors $\mathbf{a}' = (a_1',a_2',a_3')$ and $\mathbf{b}' = (b_1',b_2',b_3')$ span the same plane in K^3 as $\mathbf{a} = (a_1,a_2,a_3)$ and $\mathbf{b} = (b_1,b_2,b_3)$, then

$$\mathbf{a}' = \alpha\mathbf{a} + \beta\mathbf{b}, \quad \mathbf{b}' = \gamma\mathbf{a} + \delta\mathbf{b}$$

for some invertible matrix $\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$. Therefore,

$$u\mathbf{a}' + v\mathbf{b}' = (u\alpha + v\gamma)\mathbf{a} + (u\beta + v\delta)\mathbf{b} = u_1\mathbf{a} + v_1\mathbf{b}$$

for a new choice of parameters u_1, v_1. This means that L and L' are the same line.

If L and L' are different lines, then $\mathbf{a}, \mathbf{b}$ and $\mathbf{a}', \mathbf{b}'$ span different planes, so the vectors $\mathbf{a}, \mathbf{b}, \mathbf{a}', \mathbf{b}'$ must span all of K^3. Since (F_x, F_y, F_z) has dot product 0 with these vectors, it must be the 0 vector. This means that P is a singular point, contrary to our assumption.

Finally, we need to show that the tangent line intersects the curve to order at least 2. Suppose, for example, that $F_x \neq 0$ at P. The cases where $F_y \neq 0$ and $F_z \neq 0$ are similar. The tangent line can be given the parameterization

$$x = -(F_y/F_x)u - (F_z/F_x)v, \quad y = u, \quad z = v,$$

so

$$a_1 = -F_y/F_x,\ b_1 = -F_z/F_x,\ a_2 = 1,\ b_2 = 0,\ a_3 = 0,\ b_3 = 1$$

in the notation of (2.2). Substitute into (2.3) to obtain

$$\tilde{F}_u = (-F_y/F_x)F_x + F_y = 0, \quad \tilde{F}_v = (-F_z/F_x)F_x + F_z = 0.$$

By the discussion at the beginning of the proof, this means that the tangent line intersects the curve to order $k \geq 2$. ∎

The associativity of elliptic curve addition will follow easily from the next result. The proof can be simplified if the points P_{ij} are assumed to be distinct. The cases where points are equal correspond to situations where tangent lines are used in the definition of the group law. Correspondingly, this is where it is more difficult to verify the associativity by direct calculation with the formulas for the group law.

THEOREM 2.6

Let $C(x, y, z)$ be a homogeneous cubic polynomial, and let C be the curve in $\mathbf{P}^2_K$ described by $C(x, y, z) = 0$. Let ℓ_1, ℓ_2, ℓ_3 and m_1, m_2, m_3 be lines in $\mathbf{P}^2_K$ such that $\ell_i \neq m_j$ for all i, j. Let P_{ij} be the point of intersection of ℓ_i and m_j. Suppose P_{ij} is a nonsingular point on the curve C for all $(i, j) \neq (3, 3)$. In addition, we require that if, for some i, there are $k \geq 2$ of the points P_{i1}, P_{i2}, P_{i3} equal to the same point, then ℓ_i intersects C to order at least k at this point. Also, if, for some j, there are $k \geq 2$ of the points P_{1j}, P_{2j}, P_{3j} equal to the same point, then m_j intersects C to order at least k at this point. Then P_{33} also lies on the curve C.

PROOF Express ℓ_1 in the parametric form (2.2). Then $C(x, y, z)$ becomes $\tilde{C}(u, v)$. The line ℓ_1 passes through P_{11}, P_{12}, P_{13}. Let $(u_1 : v_1), (u_2 : v_2), (u_3 : v_3)$ be the parameters on ℓ_1 for these points. Since these points lie on C, we have $\tilde{C}(u_i, v_i) = 0$ for $i = 1, 2, 3$.

Let m_j have equation $m_j(x, y, z) = a_j x + b_j y + c_j z = 0$. Substituting the parameterization for ℓ_1 yields $\tilde{m}_j(u, v)$. Since P_{ij} lies on m_j, we have $\tilde{m}_j(u_j, v_j) = 0$ for $j = 1, 2, 3$. Since $\ell_1 \neq m_j$ and since the zeros of $\tilde{m}_j$ yield the intersections of ℓ_1 and m_j, the function $\tilde{m}_j(u, v)$ vanishes only at P_{1j}, so the linear form $\tilde{m}_j$ is nonzero. Therefore, the product $\tilde{m}_1(u, v)\tilde{m}_2(u, v)\tilde{m}_3(u, v)$ is a nonzero cubic homogeneous polynomial. We need to relate this product to $\tilde{C}$.

LEMMA 2.7

Let $R(u, v)$ and $S(u, v)$ be homogeneous polynomials of degree 3, with $S(u, v)$ not identically 0, and suppose there are three points $(u_i : v_i)$, $i = 1, 2, 3$, at which R and S vanish. Moreover, if k of these points are equal to the same point, we require that R and S vanish to order at least k at this point (that is, $(v_i u - u_i v)^k$ divides R and S). Then there is a constant $\alpha \in K$ such that $R = \alpha S$.

PROOF First, observe that a nonzero cubic homogeneous polynomial $S(u, v)$ can have at most 3 zeros $(u : v)$ in $\mathbf{P}^1_K$ (counting multiplicities).

This can be proved as follows. Factor off the highest possible power of v, say v^k. Then $S(u,v)$ vanishes to order k at $(1:0)$, and $S(u,v) = v^k S_0(u,v)$ with $S_0(1,0) \neq 0$. Since $S_0(u,1)$ is a polynomial of degree $3-k$, the polynomial $S_0(u,1)$ can have at most $3-k$ zeros, counting multiplicities (it has exactly $3-k$ if K is algebraically closed). All points $(u:v) \neq (1:0)$ can be written in the form $(u:1)$, so $S_0(u,v)$ has at most $3-k$ zeros. Therefore, $S(u,v)$ has at most $k+(3-k)=3$ zeros in $\mathbf{P}^1_K$.

It follows easily that the condition that $S(u,v)$ vanish to order at least k could be replaced by the condition that $S(u,v)$ vanish to order exactly k. However, it is easier to check "at least" than "exactly." Since we are allowing the possibility that $R(u,v)$ is identically 0, this remark does not apply to R.

Let $(u_0,:v_0)$ be any point in $\mathbf{P}^1_K$ not equal to any of the $(u_i:v_i)$. (`Technical point`: If K has only two elements, then $\mathbf{P}^1_K$ has only three elements. In this case, enlarge K to $GF(4)$. The α we obtain is forced to be in K since it is the ratio of a coefficient of R and a coefficient of S, both of which are in K.) Since S can have at most three zeros, $S(u_0,v_0) \neq 0$. Let $\alpha = R(u_0,v_0)/S(u_0,v_0)$. Then $R(u,v) - \alpha S(u,v)$ is a cubic homogeneous polynomial that vanishes at the four points $(u_i:v_i)$, $i=0,1,2,3$. Therefore $R-\alpha S$ must be identically zero. ∎

Returning to the proof of the theorem, we note that $\tilde{C}$ and $\tilde{m}_1\tilde{m}_2\tilde{m}_3$ vanish at the points $(u_i:v_i)$, $i=1,2,3$. Moreover, if k of the points P_{1j} are the same point, then k of the linear functions vanish at this point, so the product $\tilde{m}_1(u,v)\tilde{m}_2(u,v)\tilde{m}_3(u,v)$ vanishes to order at least k. By assumption, $\tilde{C}$ vanishes to order at least k in this situation. By the lemma, there exists a constant α such that

$$\tilde{C} = \alpha \tilde{m}_1\tilde{m}_2\tilde{m}_3.$$

Let

$$C_1(x,y,z) = C(x,y,z) - \alpha m_1(x,y,z)m_2(x,y,z)m_3(x,y,z).$$

The line ℓ_1 can be described by a linear equation $\ell_1(x,y,z) = ax+by+cz = 0$. At least one coefficient is nonzero, so let's assume $a \neq 0$. The other cases are similar. The parameterization of the line ℓ_1 can be taken to be

$$x = -(b/a)u - (c/a)v, \quad y = u, \quad z = v. \tag{2.4}$$

Then $\tilde{C}_1(u,v) = C_1(-(b/a)u-(c/a)v,u,v)$. Write $C_1(x,y,z)$ as a polynomial in x with polynomials in y,z as coefficients. Writing

$$x^n = (1/a^n)\left((ax+by+cz)-(by+cz)\right)^n = (1/a^n)\left((ax+by+cz)^n + \cdots\right),$$

we can rearrange $C_1(x,y,z)$ to be a polynomial in $ax+by+cz$ whose coefficients are polynomials in y,z:

$$C_1(x,y,z) = a_3(y,z)(ax+by+cz)^3 + \cdots + a_0(y,z). \tag{2.5}$$

Substituting (2.4) into (2.5) yields

$$0 = \tilde{C}_1(u, v) = a_0(u, v),$$

since $ax + by + cz$ vanishes identically when x, y, z are written in terms of u, v. Therefore $a_0(y, z) = a_0(u, v)$ is the zero polynomial. It follows from (2.5) that $C_1(x, y, z)$ is a multiple of $\ell_1(x, y, z) = ax + by + cz$.

Similarly, there exists a constant β such that $C(x, y, z) - \beta\ell_1\ell_2\ell_3$ is a multiple of m_1.

Let

$$D(x, y, z) = C - \alpha m_1 m_2 m_3 - \beta\ell_1\ell_2\ell_3.$$

Then $D(x, y, z)$ is a multiple of ℓ_1 and a multiple of m_1.

LEMMA 2.8
$D(x, y, z)$ is a multiple of $\ell_1(x, y, z)m_1(x, y, z)$.

PROOF Write $D = m_1 D_1$. We need to show that ℓ_1 divides D_1. We could quote some result about unique factorization, but instead we proceed as follows. Parameterize the line ℓ_1 via (2.4) (again, we are considering the case $a \neq 0$). Substituting this into the relation $D = m_1 D_1$ yields $\tilde{D} = \tilde{m}_1\tilde{D}_1$. Since ℓ_1 divides D, we have $\tilde{D} = 0$. Since $m_1 \neq \ell_1$, we have $\tilde{m}_1 \neq 0$. Therefore $\tilde{D}_1(u, v)$ is the zero polynomial. As above, this implies that $D_1(x, y, z)$ is a multiple of ℓ_1, as desired. ∎

By the lemma,

$$D(x, y, z) = \ell_1 m_1 \ell,$$

where $\ell(x, y, z)$ is linear. By assumption, $C = 0$ at P_{22}, P_{23}, P_{32}. Also, $\ell_1\ell_2\ell_3$ and $m_1m_2m_3$ vanish at these points. Therefore, $D(x, y, z)$ vanishes at these points. Our goal is to show that D is identically 0.

LEMMA 2.9
$\ell(P_{22}) = \ell(P_{23}) = \ell(P_{32}) = 0$.

PROOF First suppose that $P_{13} \neq P_{23}$. If $\ell_1(P_{23}) = 0$, then P_{23} is on the line ℓ_1 and also on ℓ_2 and m_3 by definition. Therefore, P_{23} equals the intersection P_{13} of ℓ_1 and m_3. Since P_{23} and P_{13} are for the moment assumed to be distinct, this is a contradiction. Therefore $\ell_1(P_{23}) \neq 0$. Since $D(P_{23}) = 0$, it follows that $m_1(P_{23})\ell(P_{23}) = 0$.

Suppose now that $P_{13} = P_{23}$. Then, by the assumption in the theorem, m_3 is tangent to C at P_{23}, so $\text{ord}_{m_3,P_{23}}(C) \geq 2$. Since $P_{13} = P_{23}$ and P_{23} lies on m_3, we have $\text{ord}_{m_3,P_{23}}(\ell_1) = \text{ord}_{m_3,P_{23}}(\ell_2) = 1$. Therefore, $\text{ord}_{m_3,P_{23}}(\alpha\ell_1\ell_2\ell_3) \geq 2$. Also, $\text{ord}_{m_3,P_{23}}(\beta m_1m_2m_3) = \infty$. Therefore,

$\text{ord}_{m_3,P_{23}}(D) \geq 2$, since D is a sum of terms, each of which vanishes to order at least 2. But $\text{ord}_{m_3,P_{23}}(\ell_1) = 1$, so we have

$$\text{ord}_{m_3,P_{23}}(m_1\ell) = \text{ord}_{m_3,P_{23}}(D) - \text{ord}_{m_3,P_{23}}(\ell_1) \geq 1.$$

Therefore $m_1(P_{23})\ell(P_{23}) = 0$.

In both cases, we have $m_1(P_{23})\ell(P_{23}) = 0$.

If $m_1(P_{23}) \neq 0$, then $\ell(P_{23}) = 0$, as desired.

If $m_1(P_{23}) = 0$, then P_{23} lies on m_1, and also on ℓ_2 and m_3, by definition. Therefore, $P_{23} = P_{21}$, since ℓ_2 and m_1 intersect in a unique point. By assumption, ℓ_2 is therefore tangent to C at P_{23}. Therefore, $\text{ord}_{\ell_2,P_{23}}(C) \geq 2$. As above, $\text{ord}_{\ell_2,P_{23}}(D) \geq 2$, so

$$\text{ord}_{\ell_2,P_{23}}(\ell_1\ell) \geq 1.$$

If in this case we have $\ell_1(P_{23}) = 0$, then P_{23} lies on ℓ_1, ℓ_2, m_3. Therefore $P_{13} = P_{23}$. By assumption, the line m_3 is tangent to C at P_{23}. Since P_{23} is a nonsingular point of C, Lemma 2.5 says that $\ell_2 = m_3$, contrary to hypothesis. Therefore, $\ell_1(P_{23}) \neq 0$, so $\ell(P_{23}) = 0$.

Similarly, $\ell(P_{22}) = \ell(P_{32}) = 0$. ∎

If $\ell(x, y, z)$ is identically 0, then D is identically 0. Therefore, assume that $\ell(x, y, z)$ is not zero and hence it defines a line ℓ.

First suppose that P_{23}, P_{22}, P_{32} are distinct. Then ℓ and ℓ_2 are lines through P_{23} and P_{22}. Therefore $\ell = \ell_2$. Similarly, $\ell = m_2$. Therefore $\ell_2 = m_2$, contradiction.

Now suppose that $P_{32} = P_{22}$. Then m_2 is tangent to C at P_{22}. As before,

$$\text{ord}_{m_2,P_{22}}(\ell_1 m_1 \ell) \geq 2.$$

We want to show that this forces ℓ to be the same line as m_2.

If $m_1(P_{22}) = 0$, then P_{22} lies on m_1, m_2, ℓ_2. Therefore, $P_{21} = P_{22}$. This means that ℓ_2 is tangent to C at P_{22}. By Lemma 2.5, $\ell_2 = m_2$, contradiction. Therefore, $m_1(P_{22}) \neq 0$.

If $\ell_1(P_{22}) \neq 0$, then $\text{ord}_{m_2,P_{22}}(\ell) \geq 2$. This means that ℓ is the same line as m_2.

If $\ell_1(P_{22}) = 0$, then $P_{22} = P_{32}$ lies on $\ell_1, \ell_2, \ell_3, m_2$, so $P_{12} = P_{22} = P_{32}$. Therefore $\text{ord}_{m_2,P_{22}}(C) \geq 3$. By the reasoning above, we now have $\text{ord}_{m_2,P_{22}}(\ell_1 m_1 \ell) \geq 3$. Since we have proved that $m_1(P_{22}) \neq 0$, we have $\text{ord}_{m_2,P_{22}}(\ell) \geq 2$. This means that ℓ is the same line as m_2.

So now we have proved, under the assumption that $P_{32} = P_{22}$, that ℓ is the same line as m_2. By Lemma 2.9, P_{23} lies on ℓ, and therefore on m_2. It also lies on ℓ_2 and m_3. Therefore, $P_{22} = P_{23}$. This means that ℓ_2 is tangent to C at P_{22}. Since $P_{32} = P_{22}$ means that m_2 is also tangent to C at P_{22}, we have $\ell_2 = m_2$, contradiction. Therefore, $P_{32} \neq P_{22}$ (under the assumption that $\ell \neq 0$).

Similarly, $P_{23} \neq P_{22}$.

Finally, suppose $P_{23} = P_{32}$. Then P_{23} lies on ℓ_2, ℓ_3, m_2, m_3. This forces $P_{22} = P_{32}$, which we have just shown is impossible.

Therefore, all possibilities lead to contradictions. It follows that $\ell(x, y, z)$ must be identically 0. Therefore $D = 0$, so

$$C = \alpha\ell_1\ell_2\ell_3 + \beta m_1 m_2 m_3.$$

Since ℓ_3 and m_3 vanish at P_{33}, we have $C(P_{33}) = 0$, as desired. This completes the proof of Theorem 2.6. ∎

REMARK 2.10 Note that we proved the stronger result that

$$C = \alpha\ell_1\ell_2\ell_3 + \beta m_1 m_2 m_3$$

for some constants α, β. Since there are 10 coefficients in an arbitrary homogeneous cubic polynomial in three variables and we have required that C vanish at eight points (when the P_{ij} are distinct), it is not surprising that the set of possible polynomials is a two-parameter family. When the P_{ij} are not distinct, the tangency conditions add enough restrictions that we still obtain a two-parameter family. ∎

We can now prove the associativity of addition for an elliptic curve. Let P, Q, R be points on E. Define the lines

$$\ell_1 = \overline{PQ}, \quad \ell_2 = \overline{\infty, Q+R}, \quad \ell_3 = \overline{R, P+Q}$$

$$m_1 = \overline{QR}, \quad m_2 = \overline{\infty, P+Q}, \quad m_3 = \overline{P, Q+R}.$$

We have the following intersections:

	ℓ_1	ℓ_2	ℓ_3
m_1	Q	$-(Q+R)$	R
m_2	$-(P+Q)$	∞	$P+Q$
m_3	P	$Q+R$	X

Assume for the moment that the hypotheses of the theorem are satisfied. Then all the points in the table, including X, lie on E. The line ℓ_3 has three points of intersection with E, namely $R, P+Q$, and X. By the definition of addition, $X = -((P+Q)+R)$. Similarly, m_3 intersects C in 3 points, which means that $X = -(P+(Q+R))$. Therefore, after reflecting across the x-axis, we obtain $(P+Q)+R = P+(Q+R)$, as desired.

It remains to verify the hypotheses of the theorem, namely that the orders of intersection are correct and that the lines ℓ_i are distinct from the lines m_j.

First we want to dispense with cases where ∞ occurs. The problem is that we treated ∞ as a special case in the definition of the group law. However,

as pointed out earlier, the tangent line at ∞ intersects the curve only at ∞ (and intersects to order 3 at ∞). It follows that if two of the entries in a row or column of the above table of intersections are equal to ∞, then so is the third, and the line intersects the curve to order 3. Therefore, this hypothesis is satisfied.

It is also possible to treat directly the cases where some of the intersection points $P, Q, R, \pm(P+Q), \pm(Q+R)$ are ∞. In the cases where at least one of P, Q, R is ∞, associativity is trivial.

If $P + Q = \infty$, then $(P + Q) + R = \infty + R = R$. On the other hand, the sum $Q + R$ is computed by first drawing the line L through Q and R, which intersects E in $-(Q + R)$. Since $P + Q = \infty$, the reflection of Q across the x-axis is P. Therefore, the reflection L' of L passes through P, $-R$, and $Q + R$. The sum $P + (Q + R)$ is found by drawing the line through P and $Q + R$, which is L'. We have just observed that the third point of intersection of L' with E is $-R$. Reflecting yields $P + (Q + R) = R$, so associativity holds in this case.

Similarly, associativity holds when $Q + R = \infty$.

Finally, we need to consider what happens if some line ℓ_i equals some line m_j, since then Theorem 2.6 does not apply.

First, observe that if P, Q, R are collinear, then associativity is easily verified directly.

Second, suppose that $P, Q, Q + R$ are collinear. Then $P + (Q + R) = -Q$. Also, $P + Q = -(Q + R)$, so $(P + Q) + R = -(Q + R) + R$. The second equation of the following shows that associativity holds in this case.

LEMMA 2.11

Let P_1, P_2 be points on an elliptic curve. Then $(P_1 + P_2) - P_2 = P_1$ and $-(P_1 + P_2) + P_2 = -P_1$.

PROOF The two relations are reflections of each other, so it suffices to prove the second one. The line L through P_1 and P_2 intersects the elliptic curve in $-(P_1 + P_2)$. Regarding L as the line through $-(P_1 + P_2)$ and P_2 yields $-(P_1 + P_2) + P_2 = -P_1$, as claimed. ∎

Suppose that $\ell_i = m_j$ for some i, j. We consider the various cases. By the above discussion, we may assume that all points in the table of intersections are finite, except for ∞ and possibly X. Note that each ℓ_i and each m_j meets E in three points (counting multiplicity), one of which is P_{ij}. If the two lines coincide, then the other two points must coincide in some order.

1. $\ell_1 = m_1$: Then P, Q, R are collinear, and associativity follows.

2. $\ell_1 = m_2$: In this case, P, Q, ∞ are collinear, so $P + Q = \infty$; associativity follows by the direct calculation made above.

3. $\ell_2 = m_1$: Similar to the previous case.

4. $\ell_1 = m_3$: Then $P, Q, Q+R$ are collinear; associativity was proved above.

5. $\ell_3 = m_1$: Similar to the previous case.

6. $\ell_2 = m_2$: Then $P + Q$ must be $\pm(Q + R)$. If $P + Q = Q + R$, then commutativity plus the above lemma yields

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

Therefore,

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

If $P + Q = -(Q + R)$, then

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

and

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

so associativity holds.

7. $\ell_2 = m_3$: In this case, the line m_3 through P and $(Q + R)$ intersects E in ∞, so $P = -(Q + R)$. Since $-(Q + R)$, Q, R are collinear, we have that P, Q, R are collinear and associativity holds.

8. $\ell_3 = m_2$: Similar to the previous case.

9. $\ell_3 = m_3$: Since ℓ_3 cannot intersect E in 4 points (counting multiplicities), it is easy to see that $P = R$ or $P = P + Q$ or $Q + R = P + Q$ or $Q + R = R$. The case $P = R$ was treated in the case $\ell_2 = m_2$. Assume $P = P + Q$. Adding $-P$ and applying Lemma 2.11 yields $\infty = Q$, in which case associativity immediately follows. The case $Q + R = R$ is similar. If $Q + R = P + Q$, then adding $-Q$ and applying Lemma 2.11 yields $P = R$, which we have already treated.

If $\ell_i \neq m_j$ for all i, j, then the hypotheses of the theorem are satisfied, so the addition is associative, as proved above. This completes the proof of the associativity of elliptic curve addition. ∎

REMARK 2.12 Note that for most of the proof, we did not use the Weierstrass equation for the elliptic curve. In fact, any nonsingular cubic curve would suffice. The identity O for the group law needs to be a point whose tangent line intersects to order 3. Three points sum to 0 if they lie on a straight line. Negation of a point P is accomplished by taking the line through O and P. The third point of intersection is then $-P$. Associativity of this group law follows just as in the Weierstrass case.

2.4.1 The Theorems of Pappus and Pascal

Theorem 2.6 has two other nice applications outside the realm of elliptic curves.

THEOREM 2.13 (Pascal's Theorem)
Let $ABCDEF$ be a hexagon inscribed in a conic section (ellipse, parabola, or hyperbola), where A, B, C, D, E, F are distinct points in the affine plane. Let X be the intersection of $\overline{AB}$ and $\overline{DE}$, let Y be the intersection of $\overline{BC}$ and $\overline{EF}$, and let Z be the intersection of $\overline{CD}$ and $\overline{FA}$. Then X, Y, Z are collinear (see Figure 2.4).

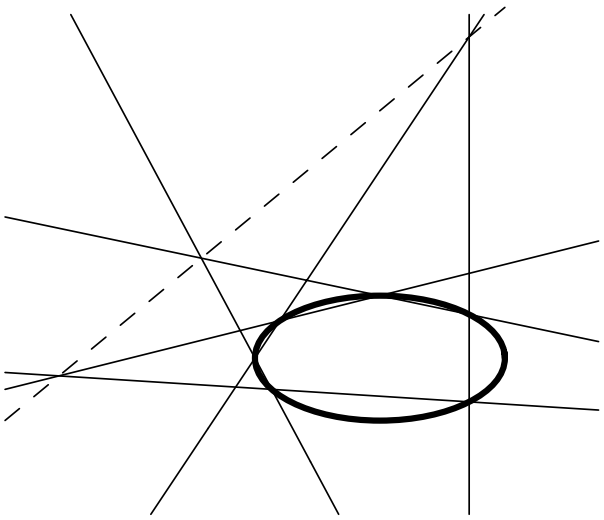


Figure 2.4
Pascal's Theorem

REMARK 2.14 (1) A conic is given by an equation $q(x, y) = ax^2 + bxy + cy^2 + dx + ey + f = 0$ with at least one of a, b, c nonzero. Usually, it is assumed that $b^2 - 4ac \neq 0$; otherwise, the conic degenerates into a product of two linear factors, and the graph is the union of two lines. The present theorem holds even in this case, as long as the points A, C, E lie on one of the lines, B, D, F lie on the other, and none is the intersection of the two lines.

(2) Possibly $\overline{AB}$ and $\overline{DE}$ are parallel, for example. Then X is an infinite point in $\mathbf{P}^2_K$.

(3) Note that X, Y, Z will always be distinct. This is easily seen as follows: First observe that X, Y, Z cannot lie on the conic since a line can intersect

the conic in at most two points; the points A, B, C, D, E, F are assumed to be distinct and therefore exhaust all possible intersections. If $X = Y$, then $\overline{AB}$ and $\overline{BC}$ meet in both B and Y, and therefore the lines are equal. But this means that $A = C$, contradiction. Similarly, $X \neq Z$ and $Y \neq Z$. ∎

PROOF Define the following lines:

$$\ell_1 = \overline{EF},\ \ell_2 = \overline{AB},\ \ell_3 = \overline{CD},\ m_1 = \overline{BC},\ m_2 = \overline{DE},\ m_3 = \overline{FA}.$$

We have the following table of intersections:

	ℓ_1	ℓ_2	ℓ_3
m_1	Y	B	C
m_2	E	X	D
m_3	F	A	Z

Let $q(x, y) = 0$ be the affine equation of the conic. In order to apply Theorem 2.6, we change $q(x, y)$ to its homogeneous form $Q(x, y, z)$. Let $\ell(x, y, z)$ be a linear form giving the line through X and Y. Then

$$C(x, y, z) = Q(x, y, z)\ell(x, y, z)$$

is a homogeneous cubic polynomial. The curve $C = 0$ contains all of the points in the table, with the possible exception of Z. It is easily checked that the only singular points of C are the points of intersection of $Q = 0$ and $\ell = 0$, and the intersection of the two lines comprising $Q = 0$ in the case of a degenerate conic. Since none of these points occur among the points we are considering, the hypotheses of Theorem 2.6 are satisfied. Therefore, $C(Z) = 0$. Since $Q(Z) \neq 0$, we must have $\ell(Z) = 0$, so Z lies on the line through X and Y. Therefore, X, Y, Z are collinear. This completes the proof of Pascal's theorem. ∎

COROLLARY 2.15 (Pappus's Theorem)
Let ℓ and m be two distinct lines in the plane. Let A, B, C be distinct points of ℓ and let A', B', C' be distinct points of m. Assume that none of these points is the intersection of ℓ and m. Let X be the intersection of $\overline{AB'}$ and $\overline{A'B}$, let Y be the intersection of $\overline{B'C}$ and $\overline{BC'}$, and let Z be the intersection of $\overline{CA'}$ and $\overline{C'A}$. Then X, Y, Z are collinear (see Figure 2.5).

PROOF This is the case of a degenerate conic in Theorem 2.13. The "hexagon" is $AB'CA'BC'$. ∎

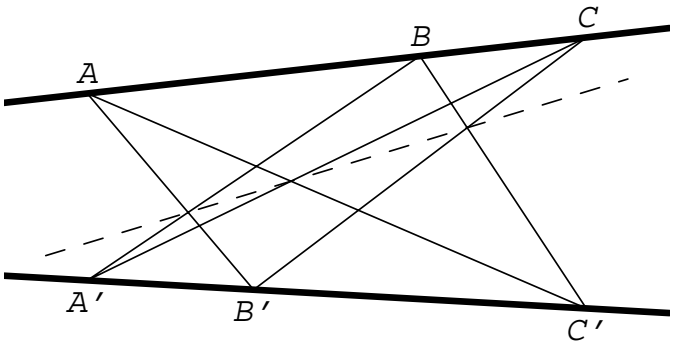


Figure 2.5
Pappus's Theorem

2.5 Other Equations for Elliptic Curves

In this book, we are mainly using the Weierstrass equation for an elliptic curve. However, elliptic curves arise in various other guises, and it is worthwhile to discuss these briefly.

2.5.1 Legendre Equation

This is a variant on the Weierstrass equation. Its advantage is that it allows us to express all elliptic curves over an algebraically closed field (of characteristic not 2) in terms of one parameter.

PROPOSITION 2.16
Let K be a field of characteristic not 2 and let

$$y^2 = x^3 + ax^2 + bx + c = (x - e_1)(x - e_2)(x - e_3)$$

be an elliptic curve E over K with $e_1, e_2, e_3 \in K$. Let

$$x_1 = (e_2 - e_1)^{-1}(x - e_1), \quad y_1 = (e_2 - e_1)^{-3/2}y, \quad \lambda = \frac{e_3 - e_1}{e_2 - e_1}.$$

Then $\lambda \neq 0, 1$ and

$$y_1^2 = x_1(x_1 - 1)(x_1 - \lambda).$$

PROOF This is a straightforward calculation. ∎

The parameter λ for E is not unique. In fact, each of

$$\{\lambda, \frac{1}{\lambda}, 1-\lambda, \frac{1}{1-\lambda}, \frac{\lambda}{\lambda-1}, \frac{\lambda-1}{\lambda}\}$$

yields a Legendre equation for E. They correspond to the six permutations of the roots e_1, e_2, e_3. It can be shown that these are the only values of λ corresponding to E, so the map $\lambda \mapsto E$ is six-to-one, except where $\lambda = -1$, $1/2$, 2, or $\lambda^2 - \lambda + 1 = 0$ (in these situations, the above set collapses; see Exercise 2.13).

2.5.2 Cubic Equations

It is possible to start with a cubic equation $C(x, y) = 0$, over a field K of characteristic not 2 or 3, that has a point with $x, y \in K$ and find an invertible change of variables that transforms the equation to Weierstrass form (although possibly $4A^3 + 27B^2 = 0$). The procedure is fairly complicated (see [25], [28], or [84]), so we restrict our attention to a specific example.

Consider the cubic Fermat equation

$$x^3 + y^3 + z^3 = 0.$$

The fact that this equation has no rational solutions with $xyz \neq 0$ was conjectured by the Arabs in the 900s and represents a special case of Fermat's Last Theorem, which asserts that the sum of two nonzero nth powers of integers cannot be a nonzero nth power when $n \geq 3$. The first proof in the case $n = 3$ was probably due to Fermat. We'll discuss some of the ideas for the proof in the general case in Chapter 15.

Suppose that $x^3 + y^3 + z^3 = 0$ and $xyz \neq 0$. Since $x^3 + y^3 = (x + y)(x^2 - xy + y^2)$, we must have $x + y \neq 0$. Write

$$\frac{x}{z} = u + v, \quad \frac{y}{z} = u - v.$$

Then $(u + v)^3 + (u - v)^3 + 1 = 0$, so $2u^3 + 6uv^2 + 1 = 0$. Divide by u^3 (since $x + y \neq 0$, we have $u \neq 0$) and rearrange to obtain

$$6(v/u)^2 = -(1/u)^3 - 2.$$

Let

$$x_1 = \frac{-6}{u} = -12\frac{z}{x+y}, \qquad y_1 = \frac{36v}{u} = 36\frac{x-y}{x+y}.$$

Then

$$y_1^2 = x_1^3 - 432.$$

It can be shown (this is somewhat nontrivial) that the only rational solutions to this equation are $(x_1, y_1) = (12, \pm 36)$, and ∞. The case $y_1 = 36$ yields

$x - y = x + y$, so $y = 0$. Similarly, $y_1 = -36$ yields $x = 0$. The point with $(x_1, y_1) = \infty$ corresponds to $x = -y$, which means that $z = 0$. Therefore, there are no solutions to $x^3 + y^3 + z^3 = 0$ when $xyz \neq 0$.

2.5.3 Quartic Equations

Occasionally, we will meet curves defined by equations of the form

$$v^2 = au^4 + bu^3 + cu^2 + du + e, \tag{2.6}$$

with $a \neq 0$. If we have a point (p, q) lying on the curve with $p, q \in K$, then the equation (when it is nonsingular) can be transformed into a Weierstrass equation by an invertible change of variables that uses rational functions with coefficients in the field K. Note that an elliptic curve E defined over a field K always has a point in $E(K)$, namely ∞ (whose projective coordinates $(0 : 1 : 0)$ certainly lie in K). Therefore, if we are going to transform a curve C into Weierstrass form in such a way that all coefficients of the rational functions describing the transformation lie in K, then we need to start with a point on C that has coordinates in K.

There are curves of the form (2.6) that do not have points with coordinates in K. This phenomenon will be discussed in more detail in Chapter 8.

Suppose we have a curve defined by an equation (2.6) and suppose we have a point (p, q) lying on the curve. By changing u to $u + p$, we may assume $p = 0$, so the point has the form $(0, q)$.

First, suppose $q = 0$. If $d = 0$, then the curve has a singularity at $(u, v) = (0, 0)$. Therefore, assume $d \neq 0$. Then

$$(\frac{v}{u^2})^2 = d(\frac{1}{u})^3 + c(\frac{1}{u})^2 + b(\frac{1}{u}) + a.$$

This can be easily transformed into a Weierstrass equation in d/u and dv/u^2.

The harder case is when $q \neq 0$. We have the following result.

THEOREM 2.17

Let K be a field of characteristic not 2. Consider the equation

$$v^2 = au^4 + bu^3 + cu^2 + du + q^2$$

with $a, b, c, d, q \in K$. Let

$$x = \frac{2q(v + q) + du}{u^2}, \quad y = \frac{4q^2(v + q) + 2q(du + cu^2) - (d^2u^2/2q)}{u^3}.$$

Define

$$a_1 = d/q, \quad a_2 = c - (d^2/4q^2), \quad a_3 = 2qb, \quad a_4 = -4q^2a, \quad a_6 = a_2a_4.$$

Then

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

The inverse transformation is

$$u = \frac{2q(x+c) - (d^2/2q)}{y}, \quad v = -q + \frac{u(ux-d)}{2q}.$$

The point $(u,v) = (0,q)$ corresponds to the point $(x,y) = \infty$ and $(u,v) = (0,-q)$ corresponds to $(x,y) = (-a_2,\, a_1a_2 - a_3)$.

PROOF Most of the proof is a "straightforward" calculation that we omit. For the image of the point $(0,-q)$, see [28]. ∎

Example 2.2

Consider the equation

$$v^2 = u^4 + 1. \tag{2.7}$$

Then $a = 1$, $b = c = d = 0$, and $q = 1$. If

$$x = \frac{2(v+1)}{u^2}, \quad y = \frac{4(v+1)}{u^3},$$

then we obtain the elliptic curve E given by

$$y^2 = x^3 - 4x.$$

The inverse transformation is

$$u = 2x/y, \quad v = -1 + (2x^3/y^2).$$

The point $(u,v) = (0,1)$ corresponds to ∞ on E, and $(u,v) = (0,-1)$ corresponds to $(0,0)$. We will show in Chapter 8 that

$$E(\mathbf{Q}) = \{\infty, (0,0), (2,0), (-2,0)\}.$$

These correspond to $(u,v) = (0,1), (0,-1)$, and points at infinity. Therefore, the only finite rational points on the quartic curve are $(u,v) = (0,\pm 1)$. It is easy to deduce from this that the only integer solutions to

$$a^4 + b^4 = c^2$$

satisfy $ab = 0$. This yields Fermat's Last Theorem for exponent 4. We will discuss this in more detail in Chapter 8.

It is worth considering briefly the situation at infinity in u, v. If we make the equation (2.7) homogeneous, we obtain

$$F(u,v,w) = v^2w^2 - u^4 - w^4 = 0.$$

The points at infinity have $w = 0$. To find them, we set $w = 0$ and get $0 = u^4$, which means $u = 0$. We thus find only the point $(u : v : w) = (0 : 1 : 0)$. But we have two points, namely $(2, 0)$ and $(-2, 0)$ in the corresponding Weierstrass model. The problem is that $(u : v : w) = (0 : 1 : 0)$ is a singular point in the quartic model. At this point we have

$$F_u = F_v = F_w = 0.$$

What is happening is that the curve intersects itself at the point $(u : v : w) = (0 : 1 : 0)$. One branch of the curve is $v = +u^2\sqrt{1 + (1/u)^4}$ and the other is $v = -u^2\sqrt{1 + (1/u)^4}$. For simplicity, let's work with real or complex numbers. If we substitute the second of these expressions into $x = 2(v+1)/u^2$ and take the limit as $u \to \infty$, we obtain

$$x = \frac{2(v+1)}{u^2} = \frac{2(1 - u^2\sqrt{1 + (1/u)^4})}{u^2} \to -2.$$

If we use the other branch, we find $x \to +2$. So the transformation that changes the quartic equation into the Weierstrass equation has pulled apart the two branches (the technical term is "resolved the singularities") at the singular point. □

2.5.4 Intersection of Two Quadratic Surfaces

The intersection of two quadratic surfaces in three-dimensional space, along with a point on this intersection, is usually an elliptic curve. Rather than work in full generality, we'll consider pairs of equations of the form

$$au^2 + bv^2 = e, \quad cu^2 + dw^2 = f,$$

where a, b, c, d, e, f are nonzero elements of a field K of characteristic not 2. Each separate equation may be regarded as a surface in uvw-space, and they intersect in a curve. We'll show that if we have a point P in the intersection, then we can transform this curve into an elliptic curve in Weierstrass form.

Before analyzing the intersection of these two surfaces, let's consider the first equation by itself. It can be regarded as giving a curve C in the uv-plane. Let $P = (u_0, v_0)$ be a point on C. Let L be the line through P with slope m:

$$u = u_0 + t, \quad v = v_0 + mt.$$

We want to find the other point where L intersects C. See Figure 2.6. Substitute into the equation for C and use the fact that $au_0^2 + bv_0^2 = e$ to obtain

$$a(2u_0t + t^2) + b(2v_0mt + m^2t^2) = 0.$$

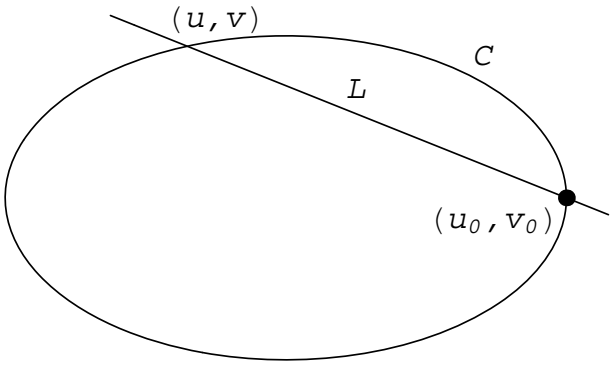


Figure 2.6

Since $t = 0$ corresponds to (u_0, v_0), we factor out t and obtain

$$t = -\frac{2au_0 + 2bv_0m}{a + bm^2}.$$

Therefore,

$$u = u_0 - \frac{2au_0 + 2bv_0m}{a + bm^2}, \quad v = v_0 - \frac{2amu_0 + 2bv_0m^2}{a + bm^2}.$$

We make the convention that $m = \infty$ yields $(u_0, -v_0)$, which is what we get if we are working with real numbers and let $m \to \infty$. Also, possibly the denominator $a + bm^2$ vanishes, in which case we get points "at infinity" in the uv-projective plane (see Exercise 2.14).

Note that if (u, v) is any point on C with coordinates in K, then the slope m of the line through (u, v) and P is in K (or is infinite). We have therefore obtained a bijection, modulo a few technicalities, between values of m (including ∞) and points on C (including points at infinity). The main point is that we have obtained a parameterization of the points on C. A similar procedure works for any conic section containing a point with coordinates in K.

Which value of m corresponds to the original point (u_0, v_0)? Let m be the slope of the tangent line at (u_0, v_0). The second point of intersection of the tangent line with the curve is again the point (u_0, v_0), so this slope is the desired value of m. The value $m = 0$ yields the point $(-u_0, v_0)$. This can be seen from the formulas, or from the fact that the line through $(-u_0, v_0)$ and (u_0, v_0) has slope 0.

We now want to intersect C, regarded as a "cylinder" in uvw-space, with the surface $cu^2 + dw^2 = f$. Substitute the expression just obtained for u to obtain

$$dw^2 = f - c\left(u_0 - \frac{2au_0 + 2bv_0m}{a + bm^2}\right)^2.$$

This may be rewritten as

$$\begin{aligned} d(w(a+bm^2))^2 &= (a+bm^2)^2 f - c(bu_0m^2 - 2bv_0m - au_0)^2 \\ &= (b^2f - cb^2u_0^2)m^4 + \cdots . \end{aligned}$$

This may now be changed to Weierstrass form by the procedure given earlier. Note that the leading coefficient $b^2f - cb^2u_0^2$ equals $b^2dw_0^2$. If $w_0 = 0$, then fourth degree polynomial becomes a cubic polynomial, so the equation just obtained is easily put into Weierstrass form. The leading term of this cubic polynomial vanishes if and only if $v_0 = 0$. But in this case, the point $(u_0, v_0, w_0) = (u_0, 0, 0)$ is a singular point of the uvw curve – a situation that we should avoid (see Exercise 2.15).

The procedure for changing "square = degree four polynomial" into Weierstrass form requires a point satisfying this equation. We could let m be the slope of the tangent line at (u_0, v_0), which corresponds to the point (u_0, v_0). The formula of Theorem 2.17 then requires that we shift the value of m to obtain $m = 0$. Instead, it's easier to use $m = 0$ directly, since this value corresponds to $(-u_0, v_0)$, as pointed out above.

Example 2.3

Consider the intersection

$$u^2 + v^2 = 2, \quad u^2 + 4w^2 = 5.$$

Let $(u_0, v_0, w_0) = (1, 1, 1)$. First, we parameterize the solutions to $u^2+v^2 = 2$. Let $u = 1 + t, v = 1 + mt$. This yields

$$(1+t)^2 + (1+mt)^2 = 2,$$

which yields $t(2 + 2m) + t^2(1 + m^2) = 0$. Discarding the solution $t = 0$, we obtain $t = -(2 + 2m)/(1 + m^2)$, hence

$$u = 1 - \frac{2+2m}{1+m^2} = \frac{m^2-2m-1}{1+m^2}, \quad v = 1 - m\frac{2+2m}{1+m^2} = \frac{1-2m-m^2}{1+m^2}.$$

Note that $m = -1$ corresponds to $(u, v) = (1, 1)$ (this is because the tangent at this point has slope $m = -1$). Substituting into $u^2 + 4w^2 = 5$ yields

$$4(w(1+m^2))^2 = 5(1+m^2)^2 - (m^2-2m-1)^2 = 4m^4 + 4m^3 + 8m^2 - 4m + 4.$$

Letting $r = w(1 + m^2)$ yields

$$r^2 = m^4 + m^3 + 2m^2 - m + 1.$$

In Theorem 2.17, we use $q = 1$. The formulas then change this curve to the generalized Weierstrass equation

$$y^2 - xy + 2y = x^3 + \frac{7}{4}x^2 - 4x - 7.$$

Completing the square yields

$$y_1^2 = x^3 + 2x^2 - 5x - 6,$$

where $y_1 = y + 1 - \frac{1}{2}x$. □

2.6 Other Coordinate Systems

The formulas for adding two points on an elliptic curve in Weierstrass form require 2 multiplications, 1 squaring, and 1 inversion in the field. Although finding inverses is fast, it is much slower than multiplication. In [27, p. 282], it is estimated that inversion takes between 9 and 40 times as long as multiplication. Moreover, squaring takes about 0.8 the time of multiplication. In many situations, this distinction makes no difference. However, if a central computer needs to verify many signatures in a second, such distinctions can become relevant. Therefore, it is sometimes advantageous to avoid inversion in the formulas for point addition. In this section, we discuss a few alternative formulas where this can be done.

2.6.1 Projective Coordinates

A natural method is to write all the points as points $(x : y : z)$ in projective space. By clearing denominators in the standard formulas for addition, we obtain the following:

Let $P_i = (x_i : y_i : z_i)$, $i = 1, 2$, be points on the elliptic curve $y^2z = x^3 + Axz^2 + Bz^3$. Then

$$(x_1 : y_1 : z_1) + (x_2 : y_2 : z_2) = (x_3 : y_3 : z_3),$$

where x_3, y_3, z_3 are computed as follows: When $P_1 \neq \pm P_2$,

$$u = y_2z_1 - y_1z_2, \quad v = x_2z_1 - x_1z_2, \quad w = u^2z_1z_2 - v^3 - 2v^2x_1z_2,$$
$$x_3 = vw, \quad y_3 = u(v^2x_1z_2 - w) - v^3y_1z_2, \quad z_3 = v^3z_1z_2.$$

When $P_1 = P_2$,

$$t = Az_1^2 + 3x_1^2, \quad u = y_1z_1, \quad v = ux_1y_1, \quad w = t^2 - 8v,$$
$$x_3 = 2uw, \quad y_3 = t(4v - w) - 8y_1^2u^2, \quad z_3 = 8u^3.$$

When $P_1 = -P_2$, we have $P_1 + P_2 = \infty$.

Point addition takes 12 multiplications and 2 squarings, while point doubling takes 7 multiplications and 5 squarings. No inversions are needed. Since

addition and subtraction are much faster than multiplication, we do not consider them in our analysis. Similarly, multiplication by a constant is not included.

2.6.2 Jacobian Coordinates

A modification of projective coordinates leads to a faster doubling procedure. Let $(x : y : z)$ represent the affine point $(x/z^2, y/z^3)$. This is somewhat natural since, as we'll see in Chapter 11, the function x has a double pole at ∞ and the function y has a triple pole at ∞. The elliptic curve $y^2 = x^3 + Ax + B$ becomes

$$y^2 = x^3 + Axz^4 + Bz^6.$$

The point at infinity now has the coordinates $\infty = (1 : 1 : 0)$.

Let $P_i = (x_i : y_i : z_i)$, $i = 1, 2$, be points on the elliptic curve $y^2 = x^3 + Axz^4 + Bz^6$. Then

$$(x_1 : y_1 : z_1) + (x_2 : y_2 : z_2) = (x_3 : y_3 : z_3),$$

where x_3, y_3, z_3 are computed as follows: When $P_1 \neq \pm P_2$,

$$r = x_1 z_2^2, \quad s = x_2 z_1^2, \quad t = y_1 z_2^3, \quad u = y_2 z_1^3, \quad v = s - r, \quad w = u - t,$$
$$x_3 = -v^3 - 2rv^2 + w^2, \quad y_3 = -tv^3 + (rv^2 - x_3)w, \quad z_3 = vz_1 z_2.$$

When $P_1 = P_2$,

$$v = 4x_1 y_1^2, \quad w = 3x_1^2 + Az_1^4,$$
$$x_3 = -2v + w^2, \quad y_3 = -8y_1^4 + (v - x_3)w, \quad z_3 = 2y_1 z_1.$$

When $P_1 = -P_2$, we have $P_1 + P_2 = \infty$.

Addition of points takes 12 multiplications and 4 squarings. Doubling takes 3 multiplications and 6 squarings. There are no inversions.

When $A = -3$, a further speed-up is possible in doubling: we have $w = 3(x_1^2 - z_1^4) = 3(x_1 + z_1^2)(x_1 - z_1^2)$, which can be computed in one squaring and one multiplication, rather than in 3 squarings. Therefore, doubling takes only 4 multiplications and 4 squarings in this case. The elliptic curves in NIST's list of curves over fields $\mathbf{F}_p$ ([86], [48, p. 262]) have $A = -3$ for this reason.

There are also situations where a point in one coordinate system can be efficiently added to a point in another coordinate system. For example, it takes only 8 multiplications and 3 squarings to add a point in Jacobian coordinates to one in affine coordinates. For much more on other choices for coordinates and on efficient point addition, see [48, Sections 3.2, 3.3] and [27, Sections 13.2, 13.3].

2.6.3 Edwards Coordinates

In [36], Harold Edwards describes a form for elliptic curves that has certain computational advantages. The case with $c = 1, d = -1$ occurs in work of Euler and Gauss. Edwards restricts to the case $d = 1$. The more general form has subsequently been discussed by Bernstein and Lange [11].

PROPOSITION 2.18
Let K be a field of characteristic not 2. Let $c, d \in K$ with $c, d \neq 0$ and d not a square in K. The curve

$$C: \quad u^2 + v^2 = c^2(1 + du^2v^2)$$

is isomorphic to the elliptic curve

$$E: \quad y^2 = (x - c^4d - 1)(x^2 - 4c^4d)$$

via the change of variables

$$x = \frac{-2c(w - c)}{u^2}, \quad y = \frac{4c^2(w - c) + 2c(c^4d + 1)u^2}{u^3},$$

where $w = (c^2du^2 - 1)v$.

The point $(0, c)$ is the identity for the group law on C, and the addition law is

$$(u_1, v_1) + (u_2, v_2) = \left(\frac{u_1v_2 + u_2v_1}{c(1 + du_1u_2v_1v_2)}, \frac{v_1v_2 - u_1u_2}{c(1 - du_1u_2v_1v_2)} \right)$$

for all points $(u_i, v_i) \in C(K)$. The negative of a point is $-(u, v) = (-u, v)$.

PROOF Write the equation of the curve as

$$u^2 - c^2 = \left(c^2du^2 - 1\right) v^2 = \frac{w^2}{c^2du^2 - 1}.$$

This yields the curve

$$w^2 = c^2du^4 - (c^4d + 1)u^2 + c^2.$$

The formulas in Section 2.5.3 then change this curve to Weierstrass form. The formula for the addition law can be obtained by a straightforward computation.

It remains to show that the addition law is defined for all points in $C(K)$. In other words, we need to show that the denominators are nonzero. Suppose

$du_1v_1u_2v_2 = -1$. Then $u_i, v_i \neq 0$ and $u_1v_1 = -1/du_2v_2$. Substituting into the formula for C yields

$$u_1^2 + v_1^2 = c^2\left(1 + \frac{1}{du_2^2v_2^2}\right) = \frac{u_2^2 + v_2^2}{du_2^2v_2^2}.$$

Therefore,

$$(u_1 + v_1)^2 = u_1^2 + v_1^2 + 2u_1v_1$$
$$= \frac{1}{d}\left(\frac{u_2^2 + v_2^2 - 2u_2v_2}{u_2^2v_2^2}\right) = \frac{1}{d}\frac{(u_2 - v_2)^2}{(u_2v_2)^2}.$$

Since d is not a square, this must reduce to $0 = 0$, so $u_1 + v_1 = 0$.

Similarly,

$$(u_1 - v_1)^2 = \frac{1}{d}\frac{(u_2 + v_2)^2}{(u_2v_2)^2},$$

which implies that $u_1 - v_1 = 0$. Therefore, $u_1 = v_1 = 0$, which is a contradiction.

The case where $du_1v_1u_2v_2 = 1$ similarly produces a contradiction. Therefore, the addition formula is always defined for points in $C(K)$. ∎

An interesting feature is that there are not separate formulas for $2P$ and $P_1 + P_2$ when $P_1 \neq P_2$.

The formula for adding points can be written in projective coordinates. The resulting computation takes 10 multiplications and 1 squaring for both point addition and point doubling.

Although any elliptic curve can be put into the form of the proposition over an algebraically closed field, this often cannot be done over the base field. An easy way to see this is that there is a point of order 2. In fact, the point $(c, 0)$ on C has order 4 (Exercise 2.7), so a curve that can be put into Edwards form over a field must have a point of order 4 defined over that field.

2.7 The j-invariant

Let E be the elliptic curve given by $y^2 = x^3 + Ax + B$, where A, B are elements of a field K of characteristic not 2 or 3. If we let

$$x_1 = \mu^2 x, \quad y_1 = \mu^3 y, \tag{2.8}$$

with $\mu \in \overline{K}^\times$, then we obtain

$$y_1^2 = x_1^3 + A_1x_1 + B_1,$$

with

$$A_1 = \mu^4 A, B_1 = \mu^6 B.$$

(In the generalized Weierstrass equation $y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$, this change of variables yields new coefficients $\mu^i a_i$. This explains the numbering of the coefficients.)

Define the **j-invariant** of E to be

$$j = j(E) = 1728\frac{4A^3}{4A^3 + 27B^2}.$$

Note that the denominator is the negative of the discriminant of the cubic, hence is nonzero by assumption. The change of variables (2.8) leaves j unchanged. The converse is true, too.

THEOREM 2.19

Let $y_1^2 = x_1^3 + A_1x_1 + B_1$ and $y_2^2 = x_2^3 + A_2x_2 + B_2$ be two elliptic curves with j-invariants j_1 and j_2, respectively. If $j_1 = j_2$, then there exists $\mu \neq 0$ in $\overline{K}$ (= algebraic closure of K) such that

$$A_2 = \mu^4 A_1, \quad B_2 = \mu^6 B_1.$$

The transformation

$$x_2 = \mu^2 x_1, \quad y_2 = \mu^3 y_1$$

takes one equation to the other.

PROOF First, assume that $A_1 \neq 0$. Since this is equivalent to $j_1 \neq 0$, we also have $A_2 \neq 0$. Choose μ such that $A_2 = \mu^4 A_1$. Then

$$\frac{4A_2^3}{4A_2^3 + 27B_2^2} = \frac{4A_1^3}{4A_1^3 + 27B_1^2} = \frac{4\mu^{-12}A_2^3}{4\mu^{-12}A_2^3 + 27B_1^2} = \frac{4A_2^3}{4A_2^3 + 27\mu^{12}B_1^2},$$

which implies that

$$B_2^2 = (\mu^6 B_1)^2.$$

Therefore $B_2 = \pm\mu^6 B_1$. If $B_2 = \mu^6 B_1$, we're done. If $B_2 = -\mu^6 B_1$, then change μ to $i\mu$ (where $i^2 = -1$). This preserves the relation $A_2 = \mu^4 A_1$ and also yields $B_2 = \mu^6 B_1$.

If $A_1 = 0$, then $A_2 = 0$. Since $4A_i^3 + 27B_i^2 \neq 0$, we have $B_1, B_2 \neq 0$. Choose μ such that $B_2 = \mu^6 B_1$. ∎

There are two special values of j that arise quite often:

1. $j = 0$: In this case, the elliptic curve E has the form $y^2 = x^3 + B$.
2. $j = 1728$: In this case, the elliptic curve has the form $y^2 = x^3 + Ax$.

The first one, with $B = -432$, was obtained in Section 2.5.2 from the Fermat equation $x^3 + y^3 + z^3 = 0$. The second curve, once with $A = -25$ and once with $A = -4$, appeared in Chapter 1.

The curves with $j = 0$ and with $j = 1728$ have automorphisms (bijective group homomorphisms from the curve to itself) other than the one defined by $(x, y) \mapsto (x, -y)$, which is an automorphism for any elliptic curve in Weierstrass form.

1. $y^2 = x^3 + B$ has the automorphism $(x, y) \mapsto (\zeta x, -y)$, where ζ is a nontrivial cube root of 1.
2. $y^2 = x^3 + Ax$ has the automorphism $(x, y) \mapsto (-x, iy)$, where $i^2 = -1$.

(See Exercise 2.17.)

Note that the j-invariant tells us when two curves are isomorphic over an algebraically closed field. However, if we are working with a nonalgebraically closed field K, then it is possible to have two curves with the same j-invariant that cannot be transformed into each other using rational functions with coefficients in K. For example, both $y^2 = x^3 - 25x$ and $y^2 = x^3 - 4x$ have $j = 1728$. The first curve has infinitely points with coordinates in $\mathbf{Q}$, for example, all integer multiples of $(-4, 6)$ (see Section 8.4). The only rational points on the second curve are ∞, $(2, 0)$, $(-2, 0)$, and $(0, 0)$ (see Section 8.4). Therefore, we cannot change one curve into the other using only rational functions defined over $\mathbf{Q}$. Of course, we can use the field $\mathbf{Q}(\sqrt{10})$ to change one curve to the other via $(x, y) \mapsto (\mu^2 x, \mu^3 y)$, where $\mu = \sqrt{10}/2$.

If two different elliptic curves defined over a field K have the same j-invariant, then we say that the two curves are **twists** of each other.

Finally, we note that j is the j-invariant of

$$y^2 = x^3 + \frac{3j}{1728 - j}x + \frac{2j}{1728 - j} \tag{2.9}$$

when $j \neq 0, 1728$. Since $y^2 = x^3 + 1$ and $y^2 = x^3 + x$ have j-invariants 0 and 1728, we find the j-invariant gives a bijection between elements of K and $\overline{K}$-isomorphism classes of elliptic curves defined over K (that is, each $j \in K$ corresponds to an elliptic curve defined over K, and any two elliptic curves defined over K and with the same j-invariant can be transformed into each other by a change of variables (2.8) defined over $\overline{K}$).

If the characteristic of K is 2 or 3, the j-invariant can also be defined, and results similar to the above one hold. See Section 2.8 and Exercise 2.18.

2.8 Elliptic Curves in Characteristic 2

Since we have been using the Weierstrass equation rather than the generalized Weierstrass equation in most of the preceding sections, the formulas

given do not apply when the field K has characteristic 2. In this section, we sketch what happens in this case.

Note that the Weierstrass equation is singular. Let $f(x, y) = y^2 - x^3 - Ax - B$. Then $f_y = 2y = 0$, since $2 = 0$ in characteristic 2. Let x_0 be a root (possibly in some extension of K) of $f_x = -3x^2 - A = 0$ and let y_0 be the square root of $x_0^3 + Ax_0 + B$. Then (x_0, y_0) lies on the curve and $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$.

Therefore, we work with the generalized Weierstrass equation for an elliptic curve E:

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6.$$

If $a_1 \neq 0$, then the change of variables

$$x = a_1^2x_1 + (a_3/a_1), \quad y = a_1^3y_1 + a_1^{-3}(a_1^2a_4 + a_3^2)$$

changes the equation to the form

$$y_1^2 + x_1y_1 = x_1^3 + a_2'x_1^2 + a_6'.$$

This curve is nonsingular if and only if $a_6' \neq 0$. The j-invariant in this case is defined to be $1/a_6'$ (more precisely, there are formulas for the j-invariant of the generalized Weierstrass form, and these yield $1/a_6'$ in this case).

If $a_1 = 0$, we let $x = x_1 + a_2$, $y = y_1$ to obtain an equation of the form

$$y_1^2 + a_3'y_1 = x_1^3 + a_4'x_1 + a_6'.$$

This curve is nonsingular if and only if $a_3' \neq 0$. The j-invariant is defined to be 0.

Let's return to the generalized Weierstrass equation and look for points at infinity. Make the equation homogeneous:

$$y^2z + a_1xyz + a_3yz^2 = x^3 + a_2x^2z + a_4xz^2 + a_6z^3.$$

Now set $z = 0$ to obtain $0 = x^3$. Therefore, $\infty = (0 : 1 : 0)$ is the only point at infinity on E, just as with the standard Weierstrass equation. A line L through (x_0, y_0) and ∞ is a vertical line $x = x_0$. If (x_0, y_0) lies on E then the other point of intersection of L and E is $(x_0, -a_1x_0 - a_3 - y_0)$. See Exercise 2.9.

We can now describe addition of points. Of course, $P + \infty = P$, for all points P. Three points P, Q, R add to ∞ if and only if they are collinear. The negation of a point is given by

$$-(x, y) = (x, -a_1x - a_3 - y).$$

To add two points P_1 and P_2, we therefore proceed as follows. Draw the line L through P_1 and P_2 (take the tangent if $P_1 = P_2$). It will intersect E in a third point P_3'. Now compute $P_3 = -P_3'$ by the formula just given (do not simply reflect across the x-axis). Then $P_1 + P_2 = P_3$.

The proof that this addition law is associative is the same as that given in Section 2.4. The points on E, including ∞, therefore form an abelian group.

Since we will need it later, let's look at the formula for doubling a point in characteristic 2. To keep the formulas from becoming too lengthy, we'll treat separately the two cases obtained above.

1. $y^2 + xy = x^3 + a_2x^2 + a_6$. Rewrite this as $y^2 + xy + x^3 + a_2x^2 + a_6 = 0$ (remember, we are in characteristic 2). Implicit differentiation yields

$$xy' + (y + x^2) = 0$$

(since $2 = 0$ and $3 = 1$). Therefore the slope of the line L through $P = (x_0, y_0)$ is $m = (y_0 + x_0^2)/x_0$. The line is

$$y = m(x - x_0) + y_0 = mx + b$$

for some b. Substitute to find the intersection (x_1, y_1) of L and E:

$$0 = (mx+b)^2 + x(mx+b) + x^3 + a_2x^2 + a_6 = x^3 + (m^2 + m + a_2)x^2 + \cdots .$$

The sum $x_0 + x_0 + x_1$ of the roots is $(m^2 + m + a_2)$, so we obtain

$$x_1 = m^2 + m + a_2 = \frac{y_0^2 + x_0^4 + x_0y_0 + x_0^3 + a_2x_0^2}{x_0^2} = \frac{x_0^4 + a_6}{x_0^2}$$

(since $y_0^2 = x_0y_0 + x_0^3 + a_2x_0^2 + a_6$). The y-coordinate of the intersection is $y_1 = m(x_1 - x_0) + y_0$. The point (x_1, y_1) equals $-2P$. Therefore $2P = (x_2, y_2)$, with

$$x_2 = (x_0^4 + a_6)/x_0^2, \quad y_2 = -x_1 - y_1 = x_1 + y_1.$$

2. $y^2 + a_3y = x^3 + a_4x + a_6$. Rewrite this as $y^2 + a_3y + x^3 + a_4x + a_6 = 0$. Implicit differentiation yields

$$a_3y' + (x^2 + a_4) = 0.$$

Therefore the tangent line L is

$$y = m(x - x_0) + y_0, \quad \text{with} \quad m = \frac{x_0^2 + a_4}{a_3}.$$

Substituting and solving, as before, finds the point of intersection (x_1, y_1) of L and E, where

$$x_1 = m^2 = \frac{x_0^4 + a_4^2}{a_3^2}$$

and $y_1 = m(x_1 - x_0) + y_0$. Therefore, $2P = (x_2, y_2)$ with

$$x_2 = (x_0^4 + a_4^2)/a_3^2, \quad y_2 = a_3 + y_1.$$

2.9 Endomorphisms

The main purpose of this section is to prove Proposition 2.21, which will be used in the proof of Hasse's theorem in Chapter 4. We'll also prove a few technical results on separable endomorphisms. The reader willing to believe that every endomorphism used in this book is separable, except for powers of the Frobenius map and multiplication by multiples of p in characteristic p, can safely omit the technical parts of this section.

By an **endomorphism** of E, we mean a homomorphism $\alpha : E(\overline{K}) \to E(\overline{K})$ that is given by rational functions. In other words, $\alpha(P_1+P_2) = \alpha(P_1)+\alpha(P_2)$, and there are rational functions (quotients of polynomials) $R_1(x,y), R_2(x,y)$ with coefficients in $\overline{K}$ such that

$$\alpha(x,y) = (R_1(x,y),\, R_2(x,y))$$

for all $(x,y) \in E(\overline{K})$. There are a few technicalities when the rational functions are not defined at a point. These will be dealt with below. Of course, since α is a homomorphism, we have $\alpha(\infty) = \infty$. We will also assume that α is nontrivial; that is, there exists some (x,y) such that $\alpha(x,y) \neq \infty$. The trivial endomorphism that maps every point to ∞ will be denoted by 0.

Example 2.4

Let E be given by $y^2 = x^3 + Ax + B$ and let $\alpha(P) = 2P$. Then α is a homomorphism and

$$\alpha(x,y) = (R_1(x,y),\, R_2(x,y))\,,$$

where

$$R_1(x,y) = \left(\frac{3x^2+A}{2y}\right)^2 - 2x$$
$$R_2(x,y) = \left(\frac{3x^2+A}{2y}\right)\left(3x - \left(\frac{3x^2+A}{2y}\right)^2\right) - y.$$

Since α is a homomorphism given by rational functions it is an endomorphism of E. □

It will be useful to have a standard form for the rational functions describing an endomorphism. For simplicity, we assume that our elliptic curve is given in Weierstrass form. Let $R(x,y)$ be any rational function. Since $y^2 = x^3+Ax+B$ for all $(x,y) \in E(\overline{K})$, we can replace any even power of y by a polynomial in x and replace any odd power of y by y times a polynomial in x and obtain a

rational function that gives the same function as $R(x, y)$ on points in $E(\overline{K})$. Therefore, we may assume that

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

Moreover, we can rationalize the denominator by multiplying the numerator and denominator by $p_3 - p_4 y$ and then replacing y^2 by $x^3 + Ax + B$. This yields

$$R(x, y) = \frac{q_1(x) + q_2(x)y}{q_3(x)}. \tag{2.10}$$

Consider an endomorphism given by

$$\alpha(x, y) = (R_1(x, y),\ R_2(x, y)),$$

as above. Since α is a homomorphism,

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

This means that

$$R_1(x, -y) = R_1(x, y) \quad \text{and} \quad R_2(x, -y) = -R_2(x, y).$$

Therefore, if R_1 is written in the form (2.10), then $q_2(x) = 0$, and if R_2 is written in the form (2.10), then the corresponding $q_1(x) = 0$. Therefore, we may assume that

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

with rational functions $r_1(x), r_2(x)$.

We can now say what happens when one of the rational functions is not defined at a point. Write

$$r_1(x) = p(x)/q(x)$$

with polynomials $p(x)$ and $q(x)$ that do not have a common factor. If $q(x) = 0$ for some point (x, y), then we assume that $\alpha(x, y) = \infty$. If $q(x) \neq 0$, then Exercise 2.19 shows that $r_2(x)$ is defined; hence the rational functions defining α are defined.

We define the **degree** of α to be

$$\deg(\alpha) = \text{Max}\{\deg p(x),\ \deg q(x)\}$$

if α is nontrivial. When $\alpha = 0$, let $\deg(0) = 0$. Define $\alpha \neq 0$ to be a **separable** endomorphism if the derivative $r_1'(x)$ is not identically zero. This is equivalent to saying that at least one of $p'(x)$ and $q'(x)$ is not identically zero. See Exercise 2.22. (In characteristic 0, a nonconstant polynomial will

have nonzero derivative. In characteristic $p > 0$, the polynomials with zero derivative are exactly those of the form $g(x^p)$.)

Example 2.5

We continue with the previous example, where $\alpha(P) = 2P$. We have

$$R_1(x, y) = \left(\frac{3x^2 + A}{2y}\right)^2 - 2x.$$

The fact that $y^2 = x^3 + Ax + B$, plus a little algebraic manipulation, yields

$$r_1(x) = \frac{x^4 - 2Ax^2 - 8Bx + A^2}{4(x^3 + Ax + B)}.$$

(This is the same as the expression in terms of division polynomials that will be given in Section 3.2.) Therefore, $\deg(\alpha) = 4$. The polynomial $q'(x) = 4(3x^2 + A)$ is not zero (including in characteristic 3, since if $A = 0$ then $x^3 + B$ has multiple roots, contrary to assumption). Therefore α is separable. □

Example 2.6

Let's repeat the previous example, but in characteristic 2. We'll use the formulas from Section 2.8 for doubling a point. First, let's look at $y^2 + xy = x^3 + a_2x^2 + a_6$. We have

$$\alpha(x, y) = (r_1(x), R_2(x, y))$$

with $r_1(x) = (x^4 + a_6)/x^2$. Therefore $\deg(\alpha) = 4$. Since $p'(x) = 4x^3 = 0$ and $q'(x) = 2x = 0$, the endomorphism α is not separable.

Similarly, in the case $y^2 + a_3y = x^3 + a_4x + a_6$, we have $r_1(x) = (x^4 + a_4^2)/a_3^2$. Therefore, $\deg(\alpha) = 4$, but α is not separable. □

In general, in characteristic p, the map $\alpha(Q) = pQ$ has degree p^2 and is not separable. The statement about the degree is Corollary 3.7. The fact that α is not separable is proved in Proposition 2.28.

An important example of an endomorphism is the **Frobenius map**. Suppose E is defined over the finite field $\mathbf{F}_q$. Let

$$\phi_q(x, y) = (x^q, y^q).$$

The Frobenius map ϕ_q plays a crucial role in the theory of elliptic curves over $\mathbf{F}_q$.

LEMMA 2.20

Let E be defined over $\mathbf{F}_q$. Then ϕ_q is an endomorphism of E of degree q, and ϕ_q is not separable.

PROOF Since $\phi_q(x,y) = (x^q, y^q)$, the map is given by rational functions (in fact, by polynomials) and the degree is q. The main point is that $\phi_q : E(\overline{\mathbf{F}}_q) \to E(\overline{\mathbf{F}}_q)$ is a homomorphism. Let $(x_1,y_1),(x_2,y_2) \in E(\overline{\mathbf{F}}_q)$ with $x_1 \neq x_2$. The sum is (x_3,y_3), with

$$x_3 = m^2 - x_1 - x_2, \qquad y_3 = m(x_1 - x_3) - y_1, \qquad \text{where } m = \frac{y_2 - y_1}{x_2 - x_1}$$

(we are working with the Weierstrass form here; the proof for the generalized Weierstrass form is essentially the same). Raise everything to the qth power to obtain

$$x_3^q = m'^2 - x_1^q - x_2^q, \qquad y_3^q = m'(x_1^q - x_3^q) - y_1^q, \qquad \text{where } m' = \frac{y_2^q - y_1^q}{x_2^q - x_1^q}.$$

This says that

$$\phi_q(x_3,y_3) = \phi_q(x_1,y_1) + \phi_q(x_2,y_2).$$

The cases where $x_1 = x_2$ or where one of the points is ∞ are checked similarly. However, there is one subtlety that arises when adding a point to itself. The formula says that $2(x_1,y_1) = (x_3,y_3)$, with

$$x_3 = m^2 - 2x_1, \qquad y_3 = m(x_1 - x_3) - y_1, \qquad \text{where } m = \frac{3x_1^2 + A}{2y_1}.$$

When this is raised to the qth power, we obtain

$$x_3^q = m'^2 - 2x_1^q, \qquad y_3^q = m'(x_1^q - x_3^q) - y_1^q, \qquad \text{where } m' = \frac{3^q (x_1^q)^2 + A^q}{2^q y_1^q}.$$

Since $2, 3, A \in \mathbf{F}_q$, we have $2^q = 2, 3^q = 3, A^q = A$. This means that we obtain the formula for doubling the point (x_1^q, y_1^q) on E (if A^q didn't equal A, we would be working on a new elliptic curve with A^q in place of A).

Since ϕ_q is a homomorphism given by rational functions, it is an endomorphism of E. Since $q = 0$ in $\mathbf{F}_q$, the derivative of x^q is identically zero. Therefore, ϕ_q is not separable. ∎

The following result will be crucial in the proof of Hasse's theorem in Chapter 4 and in the proof of Theorem 3.2.

PROPOSITION 2.21

Let $\alpha \neq 0$ be a separable endomorphism of an elliptic curve E. Then

$$\deg \alpha = \#\mathrm{Ker}(\alpha),$$

where $\mathrm{Ker}(\alpha)$ is the kernel of the homomorphism $\alpha : E(\overline{K}) \to E(\overline{K})$.

If $\alpha \neq 0$ is not separable, then

$$\deg \alpha > \#\mathrm{Ker}(\alpha).$$

PROOF Write $\alpha(x,y) = (r_1(x), yr_2(x))$ with $r_1(x) = p(x)/q(x)$, as above. Then $r_1' \neq 0$, so $p'q - pq'$ is not the zero polynomial.

Let S be the set of $x \in \overline{K}$ such that $(pq'-p'q)(x)\, q(x) = 0$. Let $(a,b) \in E(\overline{K})$ be such that

1. $a \neq 0,\ b \neq 0,\ (a,b) \neq \infty$,
2. $\deg(p(x) - aq(x)) = \text{Max}\{\deg(p), \deg(q)\} = \deg(\alpha)$,
3. $a \notin r_1(S)$, and
4. $(a,b) \in \alpha(E(\overline{K}))$.

Since $pq'-p'q$ is not the zero polynomial, S is a finite set, hence its image under α is finite. The function $r_1(x)$ is easily seen to take on infinitely many distinct values as x runs through $\overline{K}$. Since, for each x, there is a point $(x,y) \in E(\overline{K})$, we see that $\alpha(E(\overline{K}))$ is an infinite set. Therefore, such an (a,b) exists.

We claim that there are exactly $\deg(\alpha)$ points $(x_1,y_1) \in E(\overline{K})$ such that $\alpha(x_1,y_1) = (a,b)$. For such a point, we have

$$\frac{p(x_1)}{q(x_1)} = a, \quad y_1 r_2(x_1) = b.$$

Since $(a,b) \neq \infty$, we must have $q(x_1) \neq 0$. By Exercise 2.19, $r_2(x_1)$ is defined. Since $b \neq 0$ and $y_1 r_2(x_1) = b$, we must have $y_1 = b/r_2(x_1)$. Therefore, x_1 determines y_1 in this case, so we only need to count values of x_1.

By assumption (2), $p(x) - aq(x) = 0$ has $\deg(\alpha)$ roots, counting multiplicities. We therefore must show that $p - aq$ has no multiple roots. Suppose that x_0 is a multiple root. Then

$$p(x_0) - aq(x_0) = 0 \quad \text{and} \quad p'(x_0) - aq'(x_0) = 0.$$

Multiplying the equations $p = aq$ and $aq' = p'$ yields

$$ap(x_0)q'(x_0) = ap'(x_0)q(x_0).$$

Since $a \neq 0$, this implies that x_0 is a root of $pq' - p'q$, so $x_0 \in S$. Therefore, $a = r_1(x_0) \in r_1(S)$, contrary to assumption. It follows that $p - aq$ has no multiple roots, and therefore has $\deg(\alpha)$ distinct roots.

Since there are exactly $\deg(\alpha)$ points (x_1,y_1) with $\alpha(x_1,y_1) = (a,b)$, the kernel of α has $\deg(\alpha)$ elements.

Of course, since α is a homomorphism, for each $(a,b) \in \alpha(E(\overline{K}))$, there are exactly $\deg(\alpha)$ points (x_1,y_1) with $\alpha(x_1,y_1) = (a,b)$. The assumptions on (a,b) were made during the proof to obtain this result for at least one point, which suffices.

If α is not separable, then the steps of the above proof hold, except that $p' - aq'$ is always the zero polynomial, so $p(x) - aq(x) = 0$ always has multiple roots and therefore has fewer than $\deg(\alpha)$ solutions. ∎

THEOREM 2.22

Let E be an elliptic curve defined over a field K. Let $\alpha \neq 0$ be an endomorphism of E. Then $\alpha : E(\overline{K}) \to E(\overline{K})$ is surjective.

REMARK 2.23 We definitely need to be working with $\overline{K}$ instead of K in the theorem. For example, the Mordell-Weil theorem (Theorem 8.17) implies that multiplication by 2 cannot be surjective on $E(\mathbf{Q})$ if there is a point in $E(\mathbf{Q})$ of infinite order. Intuitively, working with an algebraically closed field allows us to solve the equations defining α in order to find the inverse image of a point. ∎

PROOF Let $(a, b) \in E(\overline{K})$. Since $\alpha(\infty) = \infty$, we may assume that $(a, b) \neq \infty$. Let $r_1(x) = p(x)/q(x)$ be as above. If $p(x) - aq(x)$ is not a constant polynomial, then it has a root x_0. Since p and q have no common roots, $q(x_0) \neq 0$. Choose $y_0 \in \overline{K}$ to be either square root of $x_0^3 + Ax_0 + B$. Then $\alpha(x_0, y_0)$ is defined (Exercise 2.19) and equals (a, b') for some b'. Since $b'^2 = a^3 + Aa + B = b^2$, we have $b = \pm b'$. If $b' = b$, we're done. If $b' = -b$, then $\alpha(x_0, -y_0) = (a, -b') = (a, b)$.

We now need to consider the case when $p - aq$ is constant. Since $E(\overline{K})$ is infinite and the kernel of α is finite, only finitely many points of $E(\overline{K})$ can map to a point with a given x-coordinate. Therefore, either $p(x)$ or $q(x)$ is not constant. If p and q are two nonconstant polynomials, then there is at most one constant a such that $p - aq$ is constant (if a' is another such number, then $(a' - a)q = (p - aq) - (p - a'q)$ is constant and $(a' - a)p = a'(p - aq) - a(p - a'q)$ is constant, which implies that p and q are constant). Therefore, there are at most two points, (a, b) and $(a, -b)$ for some b, that are not in the image of α. Let (a_1, b_1) be any other point. Then $\alpha(P_1) = (a_1, b_1)$ for some P_1. We can choose (a_1, b_1) such that $(a_1, b_1) + (a, b) \neq (a, \pm b)$, so there exists P_2 with $\alpha(P_2) = (a_1, b_1) + (a, b)$. Then $\alpha(P_2 - P_1) = (a, b)$, and $\alpha(P_1 - P_2) = (a, -b)$. Therefore, α is surjective. ∎

For later applications, we need a convenient criterion for separability. If (x, y) is a variable point on $y^2 = x^3 + Ax + B$, then we can differentiate y with respect to x:

$$2yy' = 3x^2 + A.$$

Similarly, we can differentiate a rational function $f(x, y)$ with respect to x:

$$\frac{d}{dx} f(x, y) = f_x(x, y) + f_y(x, y)y',$$

where f_x and f_y denote the partial derivatives.

LEMMA 2.24

Let E be the elliptic curve $y^2 = x^3 + Ax + B$. Fix a point (u, v) on E. Write

$$(x, y) + (u, v) = (f(x, y),\ g(x, y)),$$

where $f(x, y)$ and $g(x, y)$ are rational functions of x, y (the coefficients depend on (u, v)) and y is regarded as a function of x satisfying $dy/dx = (3x^2 + A)/(2y)$. Then

$$\frac{\frac{d}{dx} f(x, y)}{g(x, y)} = \frac{1}{y}.$$

PROOF The addition formulas give

$$\begin{aligned} f(x, y) &= \left(\frac{y - v}{x - u}\right)^2 - x - u \\ g(x, y) &= \frac{-(y - v)^3 + x(y - v)(x - u)^2 + 2u(y - v)(x - u)^2 - v(x - u)^3}{(x - u)^3} \\ \frac{d}{dx} f(x, y) &= \frac{2y'(y - v)(x - u) - 2(y - v)^2 - (x - u)^3}{(x - u)^3}. \end{aligned}$$

A straightforward but lengthy calculation, using the fact that $2yy' = 3x^2 + A$, yields

$$\begin{aligned} &(x - u)^3 (y \frac{d}{dx} f(x, y) - g(x, y)) \\ &= v(Au + u^3 - v^2 - Ax - x^3 + y^2) + y(-Au - u^3 + v^2 + Ax + x^3 - y^2). \end{aligned}$$

Since (u, v) and (x, y) are on E, we have $v^2 = u^3 + Au + B$ and $y^2 = x^3 + Ax + B$. Therefore, the above expression becomes

$$v(-B + B) + y(B - B) = 0.$$

Therefore, $y \frac{d}{dx} f(x, y) = g(x, y)$. ∎

REMARK 2.25 Lemma 2.24 is perhaps better stated in terms of differentials. It says that the differential dx/y is translation invariant. In fact, it is the unique translation invariant differential, up to scalar multiples, for E. See [109]. ∎

LEMMA 2.26

Let $\alpha_1, \alpha_2, \alpha_3$ be nonzero endomorphisms of an elliptic curve E with $\alpha_1 + \alpha_2 = \alpha_3$. Write

$$\alpha_j(x, y) = (R_{\alpha_j}(x),\ yS_{\alpha_j}(x)).$$

Suppose there are constants $c_{\alpha_1}, c_{\alpha_2}$ such that

$$\frac{R'_{\alpha_1}(x)}{S_{\alpha_1}(x)} = c_{\alpha_1}, \quad \frac{R'_{\alpha_2}(x)}{S_{\alpha_2}(x)} = c_{\alpha_2}.$$

Then

$$\frac{R'_{\alpha_3}(x)}{S_{\alpha_3}(x)} = c_{\alpha_1} + c_{\alpha_2}.$$

PROOF Let (x_1, y_1) and (x_2, y_2) be variable points on E. Write

$$(x_3, y_3) = (x_1, y_1) + (x_2, y_2),$$

where

$$(x_1, y_1) = \alpha_1(x, y), \quad (x_2, y_2) = \alpha_2(x, y).$$

Then x_3 and y_3 are rational functions of x_1, y_1, x_2, y_2, which in turn are rational functions of x, y. By Lemma 2.24, with $(u, v) = (x_2, y_2)$,

$$\frac{\partial x_3}{\partial x_1} + \frac{\partial x_3}{\partial y_1}\frac{dy_1}{dx_1} = \frac{y_3}{y_1}.$$

Similarly,

$$\frac{\partial x_3}{\partial x_2} + \frac{\partial x_3}{\partial y_2}\frac{dy_2}{dx_2} = \frac{y_3}{y_2}.$$

By assumption,

$$\frac{dx_j}{dx} = c_{\alpha_j}\frac{y_j}{y}$$

for $j = 1, 2$. By the chain rule,

$$\begin{aligned}\frac{dx_3}{dx} &= \frac{\partial x_3}{\partial x_1}\frac{dx_1}{dx} + \frac{\partial x_3}{\partial y_1}\frac{dy_1}{dx_1}\frac{dx_1}{dx} + \frac{\partial x_3}{\partial x_2}\frac{dx_2}{dx} + \frac{\partial x_3}{\partial y_2}\frac{dy_2}{dx_2}\frac{dx_2}{dx} \\ &= \frac{y_3}{y_1}\frac{y_1}{y}c_{\alpha_1} + \frac{y_3}{y_2}\frac{y_2}{y}c_{\alpha_2} \\ &= (c_{\alpha_1} + c_{\alpha_2})\frac{y_3}{y}.\end{aligned}$$

Dividing by y_3/y yields the result. ∎

REMARK 2.27 In terms of differentials (see the previous Remark), we have $(dx/y)\circ\alpha$ is a translation-invariant differential on E. Therefore it must be a scalar multiple $c_\alpha dx/y$ of dx/y. It follows that every nonzero endomorphism α satisfies the hypotheses of Lemma 2.26. ∎

PROPOSITION 2.28

Let E be an elliptic curve defined over a field K, and let n be a nonzero integer. Suppose that multiplication by n on E is given by

$$n(x, y) = (R_n(x), yS_n(x))$$

for all $(x, y) \in E(\overline{K})$, where R_n and S_n are rational functions. Then

$$\frac{R_n'(x)}{S_n(x)} = n.$$

Therefore, multiplication by n is separable if and only if n is not a multiple of the characteristic p of the field.

PROOF Since $R_{-n} = R_n$ and $S_{-n} = -S_n$, we have $R'_{-n}/S_{-n} = -R_n'/S_n$. Therefore, the result for positive n implies the result for negative n.

Note that the first part of the proposition is trivially true for $n = 1$. If it is true for n, then Lemma 2.26 implies that it is true for $n + 1$, which is the sum of n and 1. Therefore, $\frac{R_n'(x)}{S(x)} = n$ for all n.

We have $R_n'(x) \neq 0$ if and only if $n = R_n'(x)/S_n(x) \neq 0$, which is equivalent to p not dividing n. Since the definition of separability is that $R_n' \neq 0$, this proves the second part of the proposition. ∎

Finally, we use Lemma 2.26 to prove a result that will be needed in Sections 3.2 and 4.2. Let E be an elliptic curve defined over a finite field $\mathbf{F}_q$. The Frobenius endomorphism ϕ_q is defined by $\phi_q(x, y) = (x^q, y^q)$. It is an endomorphism of E by Lemma 2.20.

PROPOSITION 2.29

Let E be an elliptic curve defined over $\mathbf{F}_q$, where q is a power of the prime p. Let r and s be integers, not both 0. The endomorphism $r\phi_q + s$ is separable if and only if $p \nmid s$.

PROOF Write the multiplication by r endomorphism as

$$r(x, y) = (R_r(x), yS_r(x)).$$

Then

$$\begin{aligned}(R_{r\phi_q}(x), yS_{r\phi_q}(x)) &= (\phi_q r)(x, y) = (R_r^q(x),\ y^q S_r^q(x)) \\ &= \left(R_r^q(x),\ y(x^3 + Ax + B)^{(q-1)/2} S_r^q(x)\right).\end{aligned}$$

Therefore,

$$c_{r\phi_q} = R'_{r\phi_q}/S_{r\phi_q} = qR_r^{q-1}R_r'/S_{r\phi_q} = 0.$$

Also, $c_s = R'_s/S_s = s$ by Proposition 2.28. By Lemma 2.26,

$$R'_{r\phi_q+s}/S_{r\phi_q+s} = c_{r\phi_q+s} = c_{r\phi_q} + c_s = 0 + s = s.$$

Therefore, $R'_{r\phi_q+s} \neq 0$ if and only if $p \nmid s$. ∎

2.10 Singular Curves

We have been working with $y^2 = x^3 + Ax + B$ under the assumption that $x^3 + Ax + B$ has distinct roots. However, it is interesting to see what happens when there are multiple roots. It will turn out that elliptic curve addition becomes either addition of elements in K or multiplication of elements in $K^\times$ or in a quadratic extension of K. This means that an algorithm for a group $E(K)$ arising from elliptic curves, such as one to solve a discrete logarithm problem (see Chapter 5), will probably also apply to these more familiar situations. See also Chapter 7. Moreover, as we'll discuss briefly at the end of this section, singular curves arise naturally when elliptic curves defined over the integers are reduced modulo various primes.

We first consider the case where $x^3 + Ax + B$ has a triple root at $x = 0$, so the curve has the equation

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

The point $(0, 0)$ is the only singular point on the curve (see Figure 2.7). Since

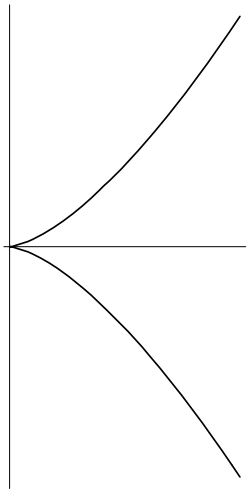


Figure 2.7
$y^2 = x^3$

any line through this point intersects the curve in at most one other point,

$(0,0)$ causes problems if we try to include it in our group. So we leave it out. The remaining points, which we denote $E_{ns}(K)$, form a group, with the group law defined in the same manner as when the cubic has distinct roots. The only thing that needs to be checked is that the sum of two points cannot be $(0,0)$. But since a line through $(0,0)$ has at most one other intersection point with the curve, a line through two nonsingular points cannot pass through $(0,0)$ (this will also follow from the proof of the theorem below).

THEOREM 2.30

Let E be the curve $y^2 = x^3$ and let $E_{ns}(K)$ be the nonsingular points on this curve with coordinates in K, including the point $\infty = (0:1:0)$. The map

$$E_{ns}(K) \to K, \quad (x,y) \mapsto \frac{x}{y}, \quad \infty \mapsto 0$$

is a group isomorphism between $E_{ns}(K)$ and K, regarded as an additive group.

PROOF Let $t = x/y$. Then $x = (y/x)^2 = 1/t^2$ and $y = x/t = 1/t^3$. Therefore we can express all of the points in $E_{ns}(K)$ in terms of the parameter t. Let $t = 0$ correspond to $(x,y) = \infty$. It follows that the map of the theorem is a bijection. (Note that $1/t$ is the slope of the line through $(0,0)$ and (x,y), so this parameterization is obtained similarly to the one obtained for quadratic curves in Section 2.5.4.)

Suppose $(x_1,y_1) + (x_2,y_2) = (x_3,y_3)$. We must show that $t_1 + t_2 = t_3$, where $t_i = x_i/y_i$. If $(x_1,y_1) \neq (x_2,y_2)$, the addition formulas say that

$$x_3 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)^2 - x_1 - x_2.$$

Substituting $x_i = 1/t_i^2$ and $y_i = 1/t_i^3$ yields

$$t_3^{-2} = \left(\frac{t_2^{-3} - t_1^{-3}}{t_2^{-2} - t_1^{-2}}\right)^2 - t_1^{-2} - t_2^{-2}.$$

A straightforward calculation simplifies this to

$$t_3^{-2} = (t_1 + t_2)^{-2}.$$

Similarly,

$$-y_3 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x_3 - x_1) + y_1$$

may be rewritten in terms of the t_i to yield

$$t_3^{-3} = (t_1 + t_2)^{-3}.$$

Taking the ratio of the expressions for t_3^{-2} and t_3^{-3} gives

$$t_3 = t_1 + t_2,$$

as desired.

If $(x_1, y_1) = (x_2, y_2)$, the proof is similar. Finally, the cases where one or more of the points $(x_i, y_i) = \infty$ are easily checked. ∎

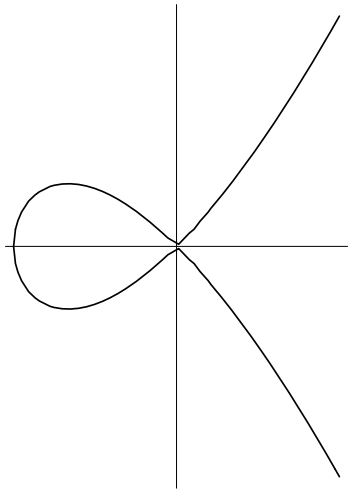


Figure 2.8
$y^2 = x^3 + x^2$

We now consider the case where $x^3 + Ax + B$ has a double root. By translating x, we may assume that this root is 0 and the curve E has the equation

$$y^2 = x^2(x + a)$$

for some $a \neq 0$. The point $(0, 0)$ is the only singularity (see Figure 2.8). Let $E_{ns}(K)$ be the nonsingular points on E with coordinates in K, including the point ∞. Let $\alpha^2 = a$ (so α might lie in an extension of K). The equation for E may be rewritten as

$$\left(\frac{y}{x}\right)^2 = a + x.$$

When x is near 0, the right side of this equation is approximately a. Therefore, E is approximated by $(y/x)^2 = a$, or $y/x = \pm\alpha$ near $x = 0$. This means that the two "tangents" to E at $(0, 0)$ are

$$y = \alpha x \quad \text{and} \quad y = -\alpha x$$

(for a different way to obtain these tangents, see Exercise 2.20).

THEOREM 2.31

Let E be the curve $y^2 = x^2(x+a)$ with $0 \neq a \in K$. Let $E_{ns}(K)$ be the nonsingular points on E with coordinates in K. Let $\alpha^2 = a$. Consider the map

$$\psi : (x, y) \mapsto \frac{y + \alpha x}{y - \alpha x}, \quad \infty \mapsto 1.$$

1. If $\alpha \in K$, then ψ gives an isomorphism from $E_{ns}(K)$ to $K^\times$, considered as a multiplicative group.

2. If $\alpha \notin K$, then ψ gives an isomorphism

$$E_{ns}(K) \simeq \{u + \alpha v \,|\, u, v \in K,\ u^2 - av^2 = 1\},$$

where the right hand side is a group under multiplication.

PROOF Let

$$t = \frac{y + \alpha x}{y - \alpha x}.$$

This may be solved for y/x to obtain

$$\frac{y}{x} = \alpha \frac{t+1}{t-1}.$$

Since $x + a = (y/x)^2$, we obtain

$$x = \frac{4\alpha^2 t}{(t-1)^2} \quad \text{and} \quad y = \frac{4\alpha^3 t(t+1)}{(t-1)^3}$$

(the second is obtained from the first using $y = x(y/x)$). Therefore, (x, y) determines t and t determines (x, y), so the map ψ is injective, and is a bijection in case (1).

In case (2), rationalize the denominator by multiplying the numerator and denominator of $(y + \alpha x)/(y - \alpha x)$ by $y + \alpha x$ to obtain an expression of the form $u + \alpha v$:

$$\frac{(y + \alpha x)}{(y - \alpha x)} = u + \alpha v.$$

We can change the sign of α throughout this equation and preserve the equality. Now multiply the resulting expression by the original to obtain

$$u^2 - av^2 = (u + \alpha v)(u - \alpha v) = \frac{(y + \alpha x)}{(y - \alpha x)} \frac{(y - \alpha x)}{(y + \alpha x)} = 1.$$

Conversely, suppose $u^2 - av^2 = 1$. Let

$$x = \left(\frac{u+1}{v}\right)^2 - a, \quad y = \left(\frac{u+1}{v}\right) x.$$

Then (x, y) is on the curve E and

$$\psi(x,y) = \frac{(y/x)+\alpha}{(y/x)-\alpha} = \frac{u+1+\alpha v}{u+1-\alpha v} = u + \alpha v$$

(the last equality uses the fact that $u^2 - av^2 = 1$). Therefore, ψ is surjective, hence is a bijection in case (2), too.

It remains to show that ψ is a homomorphism. Suppose $(x_1, y_1)+(x_2, y_2) = (x_3, y_3)$. Let

$$t_i = \frac{y_i + \alpha x_i}{y_i - \alpha x_i}.$$

We must show that $t_1 t_2 = t_3$.

When $(x_1, y_1) \neq (x_2, y_2)$, we have

$$x_3 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)^2 - a - x_1 - x_2.$$

Substituting $x_i = \dfrac{4\alpha^2 t_i}{(t_i - 1)^2}$ and $y_i = \dfrac{4\alpha^3 t_i(t_i+1)}{(t_i-1)^3}$ and simplifying yields

$$\frac{4t_3}{(t_3-1)^2} = \frac{4t_1t_2}{(t_1t_2-1)^2}. \tag{2.11}$$

Similarly,

$$-y_3 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x_3 - x_1) + y_1$$

yields

$$\frac{4\alpha^3 t_3(t_3+1)}{(t_3-1)^3} = \frac{4\alpha^3 t_1 t_2(t_1t_2+1)}{(t_1t_2-1)^3}.$$

The ratio of this equation and (2.11) yields

$$\frac{t_3 - 1}{t_3 + 1} = \frac{t_1t_2 - 1}{t_1t_2 + 1}.$$

This simplifies to yield

$$t_1 t_2 = t_3,$$

as desired.

The case where $(x_1, y_1) = (x_2, y_2)$ is similar, and the cases where one or more of the points is ∞ are trivial. This completes the proof. ∎

One situation where the above singular curves arise naturally is when we are working with curves with integral coefficients and reduce modulo various primes. For example, let E be $y^2 = x(x + 35)(x - 55)$. Then we have

$$\begin{aligned} E \mod 5 &: \ y^2 \equiv x^3, \\ E \mod 7 &: \ y^2 \equiv x^2(x+1), \\ E \mod 11 &: \ y^2 \equiv x^2(x+2). \end{aligned}$$

The first case is treated in Theorem 2.30 and is called **additive reduction**. The second case is **split multiplicative reduction** and is covered by Theorem 2.31(1). In the third case, $\alpha \notin \mathbf{F}_{11}$, so we are in the situation of Theorem 2.31(2). This is called **nonsplit multiplicative reduction**. For all primes $p \geq 13$, the cubic polynomial has distinct roots mod p, so $E \mod p$ is nonsingular. This situation is called **good reduction**.

2.11 Elliptic Curves mod n

In a few situations, we'll need to work with elliptic curves mod n, where n is composite. We'll also need to take elliptic curves over $\mathbf{Q}$ and reduce them mod n, where n is an integer. Both situations are somewhat subtle, as the following three examples show.

Example 2.7

Let E be given by

$$y^2 = x^3 - x + 1 \pmod{5^2}.$$

Suppose we want to compute $(1,1) + (21,4)$. The slope of the line through the two points is $3/20$. The denominator is not zero mod 25, but it is also not invertible. Therefore the slope is neither infinite nor finite mod 25. If we compute the sum using the formulas for the group law, the x-coordinate of the sum is

$$\left(\frac{3}{20}\right)^2 - 1 - 21 \equiv \infty \pmod{25}.$$

But $(1,1) + (1,24) = \infty$, so we cannot also have $(1,1) + (21,4) = \infty$. □

Example 2.8

Let E be given by

$$y^2 = x^3 - x + 1 \pmod{35}.$$

Suppose we want to compute $(1,1) + (26,24)$. The slope is $23/25$, which is infinite mod 5 but finite mod 7. Therefore, the formulas for the sum yield a point that is ∞ mod 5 but is finite mod 7. In a sense, the point is partially at ∞. We cannot express it in affine coordinates mod 35. One remedy is to use the Chinese Remainder Theorem to write

$$E(\mathbf{Z}_{35}) = E(\mathbf{Z}_5) \oplus E(\mathbf{Z}_7)$$

and then work mod 5 and mod 7 separately. This strategy works well in the present case, but it doesn't help in the previous example. □

Example 2.9

Let E be given by

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

over $\mathbf{Q}$. Suppose we want to compute

$$(1,1) + (\frac{571}{361}, \frac{16379}{6859}).$$

Since the points are distinct, we compute the slope of the line through them in the usual way. This allows us to find the sum. Now consider E mod 7. The two points are seen to be congruent mod 7, so the line through them mod 7 is the tangent line. Therefore, the formula we use to add the points mod 7 is different from the one used in $\mathbf{Q}$. Suppose we want to show that the reduction map from $E(\mathbf{Q})$ to $E(\mathbf{F}_7)$ is a homomorphism. At first, it would seem that this is obvious, since we just take the formulas for the group law over $\mathbf{Q}$ and reduce them mod 7. But the present example says that sometimes we are using different formulas over $\mathbf{Q}$ and mod 7. A careful analysis shows that this does not cause problems, but it should be clear that the reduction map is more subtle than one might guess. □

The remedy for the above problems is to develop a theory of elliptic curves over rings. We follow [74]. The reader willing to believe Corollaries 2.32, 2.33, and 2.34 can safely skip the details in this section.

Let R be a ring (always assumed to be commutative with 1). A tuple of elements $(x_1, x_2, \dots)$ from R is said to be **primitive** if there exist elements $r_1, r_2, \dots \in R$ such that

$$r_1x_1 + r_2x_2 + \cdots = 1.$$

When $R = \mathbf{Z}$, this means that $\gcd(x_1, x_2, \dots) = 1$. When $R = \mathbf{Z}_n$, primitivity means that $\gcd(n, x_1, x_2, \dots) = 1$. When R is a field, primitivity means that at least one of the x_i is nonzero. In general, primitivity means that the ideal generated by $x_1, x_2, \dots$ is R. We say that two primitive triples (x, y, z) and (x', y', z') are equivalent if there exists a unit $u \in R^\times$ such that

$$(x', y', z') = (ux, uy, uz)$$

(in fact, it follows easily from the existence of r, s, t with $rx' + sy' + tz' = 1$ that any u satisfying this equation must be a unit). Define 2-dimensional **projective space** over R to be

$$\mathbf{P}^2(R) = \{(x, y, z) \in R^3 \mid (x, y, z) \text{ is primitive}\} \text{ mod equivalence.}$$

The equivalence class of (x, y, z) is denoted by $(x : y : z)$.

If R is a field, $\mathbf{P}^2(R)$ is the same as that defined in Section 2.3. If $(x : y : z) \in \mathbf{P}^2(\mathbf{Q})$, we can multiply by a suitable rational number to clear

denominators and remove common factors from the numerators and therefore obtain a triple of integers with gcd=1. Therefore, $\mathbf{P}^2(\mathbf{Q})$ and $\mathbf{P}^2(\mathbf{Z})$ will be regarded as equal. Similarly, if R is a ring with

$$\mathbf{Z} \subseteq R \subseteq \mathbf{Q},$$

then $\mathbf{P}^2(R) = \mathbf{P}^2(\mathbf{Z})$.

In order to work with elliptic curves over R, we need to impose two conditions on R.

1. $2 \in R^\times$

2. If (a_{ij}) is an $m \times n$ matrix such that $(a_{11}, a_{12}, \ldots, a_{mn})$ is primitive and such that all 2×2 subdeterminants vanish (that is, $a_{ij}a_{k\ell} - a_{i\ell}a_{kj} = 0$ for all i, j, k, ℓ), then some R-linear combination of the rows is a primitive n-tuple.

The first condition is needed since we'll be working with the Weierstrass equation. In fact, we should add the condition that $3 \in R^\times$ if we want to change an arbitrary elliptic curve into Weierstrass form. Note that $\mathbf{Z}$ does not satisfy the first condition. This can be remedied by working with

$$\mathbf{Z}_{(2)} = \{\frac{x}{2^k} \,|\, x \in \mathbf{Z},\, k \geq 0\}.$$

This is a ring. As pointed out above, $\mathbf{P}^2(\mathbf{Z}_{(2)})$ equals $\mathbf{P}^2(\mathbf{Z})$, so the introduction of $\mathbf{Z}_{(2)}$ is a minor technicality.

The second condition is perhaps best understood when R is a field. In this case, the primitivity of the matrix simply means that at least one entry is nonzero. The vanishing of the 2×2 subdeterminants says that the rows are proportional to each other. The conclusion is that some linear combination of the rows (in this case, some row itself) is a nonzero vector.

When $R = \mathbf{Z}$, the primitivity of the matrix means that the gcd of the elements in the matrix is 1. Since the rows are assumed to be proportional, there is a vector $\mathbf{v}$ and integers $a_1, \ldots, a_m$ such that the ith row is $a_i\mathbf{v}$. The m-tuple $(a_1, \ldots, a_m)$ must be primitive since the gcd of its entries divides the gcd of the entries of the matrix. Therefore, there is a linear combination of the a_i's that equals 1. This means that some linear combination of the rows of the matrix is $\mathbf{v}$. The vector $\mathbf{v}$ is primitive since the gcd of its entries divides the gcd of the entries of the matrix. Therefore, we have obtained a primitive vector as a linear combination of the rows of the matrix. This shows that $\mathbf{Z}$ satisfies the second condition. The same argument, slightly modified to handle powers of 2, shows that $\mathbf{Z}_{(2)}$ also satisfies the second condition.

In general, condition 2 says that projective modules over R of rank 1 are free (see [74]). In particular, this holds for local rings, for finite rings, and for $\mathbf{Z}_{(2)}$. These suffice for our purposes.

For the rest of this section, assume R is a ring satisfying 1 and 2. An **elliptic curve** E over R is given by a homogeneous equation

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

with $A, B \in R$ such that $4A^3 + 27B^2 \in R^\times$. Define

$$E(R) = \{(x : y : z) \in \mathbf{P}^2(R) \mid y^2z = x^3 + Axz^2 + Bz^3\}.$$

The addition law is defined in essentially the same manner as in Section 2.2, but the formulas needed are significantly more complicated. To make a long story short (maybe not so short), the answer is the following.

GROUP LAW

Let $(x_i : y_i : z_i) \in E(R)$ for $i = 1, 2$. Consider the following three sets of equations:

I.

$$\begin{aligned}
x_3' &= (x_1y_2 - x_2y_1)(y_1z_2 + y_2z_1) + (x_1z_2 - x_2z_1)y_1y_2 \\
&\quad -A(x_1z_2 + x_2z_1)(x_1z_2 - x_2z_1) - 3B(x_1z_2 - x_2z_1)z_1z_2 \\
y_3' &= -3x_1x_2(x_1y_2 - x_2y_1) - y_1y_2(y_1z_2 - y_2z_1) - A(x_1y_2 - x_2y_1)z_1z_2 \\
&\quad +A(x_1z_2 + x_2z_1)(y_1z_2 - y_2z_1) + 3B(y_1z_2 - y_2z_1)z_1z_2 \\
z_3' &= 3x_1x_2(x_1z_2 - x_2z_1) - (y_1z_2 + y_2z_1)(y_1z_2 - y_2z_1) \\
&\quad +A(x_1z_2 - x_2z_1)z_1z_2
\end{aligned}$$

II.

$$\begin{aligned}
x_3'' &= y_1y_2(x_1y_2 + x_2y_1) - Ax_1x_2(y_1z_2 + y_2z_1) \\
&\quad -A(x_1y_2 + x_2y_1)(x_1z_2 + x_2z_1) - 3B(x_1y_2 + x_2y_1)z_1z_2 \\
&\quad -3B(x_1z_2 + x_2z_1)(y_1z_2 + y_2z_1) + A^2(y_1z_2 + y_2z_1)z_1z_2 \\
y_3'' &= y_1^2y_2^2 + 3Ax_1^2x_2^2 + 9Bx_1x_2(x_1z_2 + x_2z_1) \\
&\quad -A^2x_1z_2(x_1z_2 + 2x_2z_1) - A^2x_2z_1(2x_1z_2 + x_2z_1) \\
&\quad -3ABz_1z_2(x_1z_2 + x_2z_1) - (A^3 + 9B^2)z_1^2z_2^2 \\
z_3'' &= 3x_1x_2(x_1y_2 + x_2y_1) + y_1y_2(y_1z_2 + y_2z_1) + A(x_1y_2 + x_2y_1)z_1z_2 \\
&\quad +A(x_1z_2 + x_2z_1)(y_1z_2 + y_2z_1) + 3B(y_1z_2 + y_2z_1)z_1z_2
\end{aligned}$$

III.

$$\begin{aligned}
x_3''' &= (x_1y_2 + x_2y_1)(x_1y_2 - x_2y_1) + Ax_1x_2(x_1z_2 - x_2z_1) \\
&\quad +3B(x_1z_2 + x_2z_1)(x_1z_2 - x_2z_1) - A^2(x_1z_2 - x_2z_1)z_1z_2 \\
y_3''' &= (x_1y_2 - x_2y_1)y_1y_2 - 3Ax_1x_2(y_1z_2 - y_2z_1) \\
&\quad +A(x_1y_2 + x_2y_1)(x_1z_2 - x_2z_1) + 3B(x_1y_2 - x_2y_1)z_1z_2 \\
&\quad -3B(x_1z_2 + x_2z_1)(y_1z_2 - y_2z_1) + A^2(y_1z_2 - y_2z_1)z_1z_2
\end{aligned}$$

$$z_3''' = -(x_1y_2 + x_2y_1)(y_1z_2 - y_2z_1) - (x_1z_2 - x_2z_1)y_1y_2$$
$$-A(x_1z_2 + x_2z_1)(x_1z_2 - x_2z_1) - 3B(x_1z_2 - x_2z_1)z_1z_2$$

Then the matrix

$$\begin{pmatrix} x_3' & y_3' & z_3' \\ x_3'' & y_3'' & z_3'' \\ x_3''' & y_3''' & z_3''' \end{pmatrix}$$

is primitive and all 2×2 subdeterminants vanish. Take a primitive R-linear combination (x_3, y_3, z_3) of the rows. Define

$$(x_1 : y_1 : z_1) + (x_2 : y_2 : z_2) = (x_3 : y_3 : z_3).$$

Also, define

$$-(x_1 : y_1 : z_1) = (x_1 : -y_1 : z_1).$$

Then $E(R)$ is an abelian group under this definition of point addition. The identity element is $(0 : 1 : 0)$.

For some of the details concerning this definition, see [74]. The equations are deduced (with a slight correction) from those in [18]. A similar set of equations is given in [72].

When R is a field, each of these equations can be shown to give the usual group law when the output is a point in $\mathbf{P}^2(R)$ (that is, not all three coordinates vanish). If two or three of the equations yield points in $\mathbf{P}^2(R)$, then these points are equal (since the 2×2 subdeterminants vanish). If R is a ring, then it is possible that each of the equations yields a nonprimitive output (for example, perhaps 5 divides the output of I, 7 divides the output of II, and 11 divides the output of III). If we are working with $\mathbf{Z}$ or $\mathbf{Z}_{(2)}$, this is no problem. Simply divide by the gcd of the entries in an output. But in an arbitrary ring, gcd's might not exist, so we must take a linear combination to obtain a primitive vector, and hence an element in $\mathbf{P}^2(R)$.

Example 2.10

Let $R = \mathbf{Z}_{25}$ and let E be given by

$$y^2 = x^3 - x + 1 \pmod{5^2}.$$

Suppose we want to compute $(1, 1) + (21, 4)$, as in Example 2.7 above. Write the points in homogeneous coordinates as

$$(x_1 : y_1 : z_1) = (1 : 1 : 1), \quad (x_2 : y_2 : z_2) = (21 : 4 : 1).$$

Formulas I, II, III yield

$$(x_3', y_3', z_3') = (5, 23, 0)$$
$$(x_3'', y_3'', z_3'') = (5, 8, 0)$$
$$(x_3''', y_3''', z_3''') = (20, 12, 0),$$

respectively. Note that these are all the same point in $\mathbf{P}^2(\mathbf{Z}_{25})$ since

$$(5, 23, 0) = 6(5, 8, 0) = 4(20, 12, 0).$$

If we reduce the point $(5 : 8 : 0)$ mod 5, we obtain $(0 : 3 : 0) = (0 : 1 : 0)$, which is the point ∞. The fact that the point is at infinity mod 5 but not mod 25 is what caused the difficulties in our calculations in Example 2.7. □

Example 2.11

Let E be an elliptic curve. Suppose we use the formulas to calculate

$$(0 : 1 : 0) + (0 : 1 : 0).$$

Formulas I, II, III yield

$$(0, 0, 0), \quad (0, 1, 0), \quad (0, 0, 0),$$

respectively. The first and third outputs do not yield points in projective space. The second says that

$$(0 : 1 : 0) + (0 : 1 : 0) = (0 : 1 : 0).$$

This is of course the rule $\infty + \infty = \infty$ from the usual group law on elliptic curves. □

The present version of the group law allows us to work with elliptic curves over rings in theoretical settings. We give three examples.

COROLLARY 2.32

Let n_1 and n_2 be odd integers with $\gcd(n_1, n_2) = 1$. Let E be an elliptic curve defined over $\mathbf{Z}_{n_1 n_2}$. Then there is a group isomorphism

$$E(\mathbf{Z}_{n_1 n_2}) \simeq E(\mathbf{Z}_{n_1}) \oplus E(\mathbf{Z}_{n_2}).$$

PROOF Suppose that E is given by $y^2 z = x^3 + Axz^2 + Bz^3$ with $A, B \in \mathbf{Z}_{n_1 n_2}$ and $4A^3 + 27B^2 \in \mathbf{Z}_{n_1 n_2}^{\times}$. Then we can regard A and B as elements of $\mathbf{Z}_{n_i}$ and we have $4A^3 + 27B^2 \in \mathbf{Z}_{n_i}^{\times}$. Therefore, we can regard E as an elliptic curve over $\mathbf{Z}_{n_i}$, so the statement of the corollary makes sense.

The Chinese remainder theorem says that there is an isomorphism of rings

$$\mathbf{Z}_{n_1 n_2} \simeq \mathbf{Z}_{n_1} \oplus \mathbf{Z}_{n_2}$$

given by

$$x \mod n_1 n_2 \longleftrightarrow (x \mod n_1,\ x \mod n_2).$$

This yields a bijection between triples in $\mathbf{Z}_{n_1 n_2}$ and pairs of triples, one in $\mathbf{Z}_{n_1}$ and one in $\mathbf{Z}_{n_2}$. It is not hard to see that primitive triples for $\mathbf{Z}_{n_1 n_2}$ correspond to pairs of primitive triples in $\mathbf{Z}_{n_1}$ and $\mathbf{Z}_{n_2}$. Moreover,

$$y^2 z \equiv x^3 + Axz^2 + Bz^3 \pmod{n_1 n_2}$$

$$\iff \begin{cases} y^2 z \equiv x^3 + Axz^2 + Bz^3 \pmod{n_1} \\ y^2 z \equiv x^3 + Axz^2 + Bz^3 \pmod{n_2} \end{cases}$$

Therefore, there is a bijection

$$\psi : E(\mathbf{Z}_{n_1 n_2}) \longrightarrow E(\mathbf{Z}_{n_1}) \oplus E(\mathbf{Z}_{n_2}).$$

It remains to show that ψ is a homomorphism. Let $P_1, P_2 \in E(\mathbf{Z}_{n_1 n_2})$ and let $P_3 = P_1 + P_2$. This means that there is a linear combination of the outputs of formulas I, II, III that is primitive and yields P_3. Reducing all of these calculations mod n_i (for $i = 1, 2$) yields exactly the same result, namely the primitive point $P_3 \pmod{n_i}$ is the sum of $P_1 \pmod{n_i}$ and $P_2 \pmod{n_i}$. This means that $\psi(P_3) = \psi(P_1) + \psi(P_2)$, so ψ is a homomorphism. ∎

COROLLARY 2.33

Let E be an elliptic curve over $\mathbf{Q}$ given by

$$y^2 = x^3 + Ax + B$$

with $A, B \in \mathbf{Z}$. Let n be a positive odd integer such that $\gcd(n, 4A^3 + 27B^2) = 1$. Represent the elements of $E(\mathbf{Q})$ as primitive triples $(x : y : z) \in \mathbf{P}^2(\mathbf{Z})$. The map

$$\begin{aligned} \mathrm{red}_n : E(\mathbf{Q}) &\longrightarrow E(\mathbf{Z}_n) \\ (x : y : z) &\mapsto (x : y : z) \pmod{n} \end{aligned}$$

is a group homomorphism.

PROOF If $P_1, P_2 \in E(\mathbf{Q})$ and $P_1 + P_2 = P_3$, then P_3 is a primitive point that can be expressed as a linear combination of the outputs of formulas I, II, III. Reducing all of the calculations mod n yields the result. ∎

Corollary 2.33 can be generalized as follows.

COROLLARY 2.34

Let R be a ring and let I be an ideal of R. Assume that both R and R/I satisfy conditions (1) and (2) on page 66. Let E be given by

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

with $A, B \in R$ and assume there exists $r \in R$ such that

$$(4A^3 + 27B^2)r - 1 \in I.$$

Then the map

$$\begin{aligned} \text{red}_I : \ E(R) &\longrightarrow E(R/I) \\ (x : y : z) &\mapsto (x : y : z) \mod I \end{aligned}$$

is a group homomorphism.

PROOF The proof is the same as for Corollary 2.33, with R in place of $\mathbf{Z}$ and mod I in place of mod n. The condition that $(4A^3 + 27B^2)r - 1 \in I$ for some r is the requirement that $4A^3 + 27B^2$ is a unit in R/I, which was required in the definition of an elliptic curve over the ring R/I. ∎

Exercises

2.1 (a) Show that the constant term of a monic cubic polynomial is the negative of the product of the roots.

(b) Use (a) to derive the formula for the sum of two distinct points P_1, P_2 in the case that the x-coordinates x_1 and x_2 are nonzero, as in Section 2.2. Note that when one of these coordinates is 0, you need to divide by zero to obtain the usual formula.

2.2 The point $(3, 5)$ lies on the elliptic curve $E : y^2 = x^3 - 2$, defined over $\mathbf{Q}$. Find a point (not ∞) with rational, nonintegral coordinates in $(\mathbf{Q})$.

2.3 The points $P = (2, 9)$, $Q = (3, 10)$, and $R = (-4, -3)$ lie on the elliptic curve $E : y^2 = x^3 + 73$.

(a) Compute $P + Q$ and $(P + Q) + R$.

(b) Compute $Q + R$ and $P + (Q + R)$. Your answer for $P + (Q + R)$ should agree with the result of part (a). However, note that one computation used the doubling formula while the other did not use it.

2.4 Let E be the elliptic curve $y^2 = x^3 - 34x + 37$ defined over $\mathbf{Q}$. Let $P = (1, 2)$ and $Q = (6, 7)$.

(a) Compute $P + Q$.

(b) Note that $P \equiv Q \pmod 5$. Compute $2P$ on E mod 5. Show that the answer is the same as $(P+Q)$ mod 5. Observe that since $P \equiv Q$, the formula for adding the points mod 5 is not the reduction of the formula for adding $P+Q$. However, the answers are the same. This shows that the fact that reduction mod a prime is a homomorphism is subtle, and this is the reason for the complicated formulas in Section 2.11.

2.5 Let (x, y) be a point on the elliptic curve E given by $y^2 = x^3 + Ax + B$. Show that if $y = 0$ then $3x^2 + A \neq 0$. (H int: What is the condition for a polynomial to have x as a multiple root?)

2.6 Show that three points on an elliptic curve add to ∞ if and only if they are collinear.

2.7 Let C be the curve $u^2 + v^2 = c^2\left(1 + du^2v^2\right)$, as in Section 2.6.3. Show that the point $(c, 0)$ has order 4.

2.8 Show that the method at the end of Section 2.2 actually computes kP. (H int Use induction on the length of the binary expansion of k. If $k = k_0 + 2k_1 + 4k_2 + \cdots + 2^\ell a_\ell$, assume the result holds for $k' = k_0 + 2k_1 + 4k_2 + \cdots + 2^{\ell-1}a_{\ell-1}$.)

2.9 If $P = (x, y) \neq \infty$ is on the curve described by (2.1), then $-P$ is the other finite point of intersection of the curve and the vertical line through P. Show that $-P = (x, -a_1x - a_3 - y)$. (H int: This involves solving a quadratic in y. Note that the sum of the roots of a monic quadratic polynomial equals the negative of the coefficient of the linear term.)

2.10 Let $\mathbf{R}$ be the real numbers. Show that the map $(x, y, z) \mapsto (x : y : z)$ gives a two-to-one map from the sphere $x^2 + y^2 + z^2 = 1$ in $\mathbf{R}^3$ to $\mathbf{P}^2_{\mathbf{R}}$. Since the sphere is compact, this shows that $\mathbf{P}^2_{\mathbf{R}}$ is compact under the topology inherited from the sphere (a set is open in $\mathbf{P}^2_{\mathbf{R}}$ if and only if its inverse image is open in the sphere).

2.11 (a) Show that two lines $a_1x + b_1y + c_1z = 0$ and $a_2x + b_2y + c_2z = 0$ in two-dimensional projective space have a point of intersection.

(b) Show that there is exactly one line through two distinct given points in $\mathbf{P}^2_K$.

2.12 Suppose that the matrix

$$M = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{pmatrix}$$

has rank 2. Let (a, b, c) be a nonzero vector in the left nullspace of M, so $(a, b, c)M = 0$. Show that the parametric equations

$$x = a_1u + b_1v, \quad y = a_2u + b_2v, \quad z = a_3u + b_3v,$$

describe the line $ax + by + cz = 0$ in $\mathbf{P}^2_K$. (It is easy to see that the points $(x : y : z)$ lie on the line. The main point is that each point on the line corresponds to a pair (u, v).)

2.13 (a) Put the Legendre equation $y^2 = x(x-1)(x-\lambda)$ into Weierstrass form and use this to show that the j-invariant is

$$j = 2^8 \frac{(\lambda^2 - \lambda + 1)^3}{\lambda^2(\lambda - 1)^2}.$$

(b) Show that if $j \neq 0, 1728$ then there are six distinct values of λ giving this j, and that if λ is one such value then the full set is

$$\{\lambda, \frac{1}{\lambda}, 1-\lambda, \frac{1}{1-\lambda}, \frac{\lambda}{\lambda-1}, \frac{\lambda-1}{\lambda}\}.$$

(c) Show that if $j = 1728$ then $\lambda = -1, 2, 1/2$, and if $j = 0$ then $\lambda^2 - \lambda + 1 = 0$.

2.14 Consider the equation $u^2 - v^2 = 1$, and the point $(u_0, v_0) = (1, 0)$.

(a) Use the method of Section 2.5.4 to obtain the parameterization

$$u = \frac{m^2+1}{m^2-1}, \quad v = \frac{2m}{m^2-1}.$$

(b) Show that the projective curve $u^2 - v^2 = w^2$ has two points at infinity, $(1 : 1 : 0)$ and $(1 : -1 : 0)$.

(c) The parameterization obtained in (a) can be written in projective coordinates as $(u : v : w) = (m^2 + 1 : 2m : m^2 - 1)$ (or $(m^2 + n^2 : 2mn : m^2 - n^2)$ in a homogeneous form). Show that the values $m = \pm 1$ correspond to the two points at infinity. Explain why this is to be expected from the graph (using real numbers) of $u^2 - v^2 = 1$. (H int: Where does an asymptote intersect a hyperbola?)

2.15 Suppose $(u_0, v_0, w_0) = (u_0, 0, 0)$ lies in the intersection

$$au^2 + bv^2 = e, \quad cu^2 + dw^2 = f.$$

(a) Show that the procedure of Section 2.5.4 leads to an equation of the form "square = degree 2 polynomial in m."

(b) Let $F = au^2 + bv^2 = e$ and $G = cu^2 + dw^2 = f$. Show that the Jacobian matrix $\begin{pmatrix} F_u & F_v & F_w \\ G_u & G_v & G_w \end{pmatrix}$ at $(u_0, 0, 0)$ has rank 1. Since the rank is less than 2, this means that the point is a singular point.

2.16 Show that the cubic equation $x^3 + y^3 = d$ can be transformed to the elliptic curve $y_1^2 = x_1^3 - 432d^2$.

2.17 (a) Show that $(x, y) \mapsto (x, -y)$ is a group homomorphism from E to itself, for any elliptic curve in Weierstrass form.

(b) Show that $(x, y) \mapsto (\zeta x, -y)$, where ζ is a nontrivial cube root of 1, is an automorphism of the elliptic curve $y^2 = x^3 + B$.

(c) Show that $(x, y) \mapsto (-x, iy)$, where $i^2 = -1$, is an automorphism of the elliptic curve $y^2 = x^3 + Ax$.

2.18 Let K have characteristic 3 and let E be defined by $y^2 = x^3 + a_2x^2 + a_4x + a_6$. The j-invariant in this case is defined to be

$$j = \frac{a_2^6}{a_2^2a_4^2 - a_2^3a_6 - a_4^3}$$

(this formula is false if the characteristic is not 3).

(a) Show that either $a_2 \neq 0$ or $a_4 \neq 0$ (otherwise, the cubic has a triple root, which is not allowed).

(b) Show that if $a_2 \neq 0$, then the change of variables $x_1 = x - (a_4/a_2)$ yields an equation of the form $y_1^2 = x_1^3 + a_2'x_1^2 + a_6'$. This means that we may always assume that exactly one of a_2 and a_4 is 0.

(c) Show that if two elliptic curves $y^2 = x^3 + a_2x^2 + a_6$ and $y^2 = x^3 + a_2'x^2 + a_6'$ have the same j-invariant, then there exists $\mu \in \overline{K}^\times$ such that $a_2' = \mu^2a_2$ and $a_6' = \mu^6a_6$.

(d) Show that if $y^2 = x^3 + a_4x + a_6$ and $y^2 = x^3 + a_4'x^2 + a_6'$ are two elliptic curves (in characteristic 3), then there is a change of variables $y \mapsto ay$, $x \mapsto bx + c$, with $a, b \in \overline{K}^\times$ and $c \in \overline{K}$, that changes one equation into the other.

(e) Observe that if $a_2 = 0$ then $j = 0$ and if $a_4 = 0$ then $j = -a_2^3/a_6$. Show that every element of K appears as the j-invariant of a curve defined over K.

(f) Show that if two curves have the same j-invariant then there is a change of variables over $\overline{K}$ that changes one into the other.

2.19 Let $\alpha(x, y) = (p(x)/q(x),\ y \cdot s(x)/t(x))$ be an endomorphism of the elliptic curve E given by $y^2 = x^3 + Ax + B$, where p, q, s, t are polynomials such that p and q have no common root and s and t have no common root.

(a) Using the fact that (x, y) and $\alpha(x, y)$ lie on E, show that

$$\frac{(x^3 + Ax + B)\, s(x)^2}{t(x)^2} = \frac{u(x)}{q(x)^3}$$

for some polynomial $u(x)$ such that q and u have no common root. (Hint: Show that a common root of u and q must also be a root of p.)

(b) Suppose $t(x_0) = 0$. Use the facts that $x^3 + Ax + B$ has no multiple roots and all roots of t^2 are multiple roots to show that $q(x_0) = 0$. This shows that if $q(x_0) \neq 0$ then $\alpha(x_0, y_0)$ is defined.

2.20 Consider the singular curve $y^2 = x^3 + ax^2$ with $a \neq 0$. Let $y = mx$ be a line through $(0, 0)$. Show that the line always intersects the curve to order at least 2, and show that the order is 3 exactly when $m^2 = a$. This may be interpreted as saying that the lines $y = \pm\sqrt{a}x$ are the two tangents to the curve at $(0, 0)$.

2.21 (a) Apply the method of Section 2.5.4 to the circle $u^2 + v^2 = 1$ and the point $(-1, 0)$ to obtain the parameterization

$$u = \frac{1 - t^2}{1 + t^2}, \quad v = \frac{2t}{1 + t^2}.$$

(b) Suppose x, y, z are integers such that $x^2 + y^2 = z^2$, $\gcd(x, y, z) = 1$, and x is even. Use (a) to show that there are integers m, n such that

$$x = 2mn, \quad y = m^2 - n^2, \quad z = m^2 + n^2.$$

Also, show that $\gcd(x, y, z) = 1$ implies that $\gcd(m, n) = 1$ and that $m \not\equiv n \pmod 2$.

2.22 Let $p(x)$ and $q(x)$ be polynomials with no common roots. Show that

$$\frac{d}{dx}\left(\frac{p(x)}{q(x)}\right) = 0$$

(that is, the identically 0 rational function) if and only if both $p'(x) = 0$ and $q'(x) = 0$. (If p or q is nonconstant, then this can happen only in positive characteristic.)

2.23 Let E be given by $y^2 = x^3 + Ax + B$ over a field K and let $d \in K^\times$. The **twist** of E by d is the elliptic curve $E^{(d)}$ given by $y^2 = x^3 + Ad^2x + Bd^3$.

(a) Show that $j(E^{(d)}) = j(E)$.

(b) Show that $E^{(d)}$ can be transformed into E over $K(\sqrt{d})$.

(c) Show that $E^{(d)}$ can be transformed over K to the form $dy_1^2 = x_1^3 + Ax_1 + B$.

2.24 Let $\alpha, \beta \in \mathbf{Z}$ be such that $\gcd(\alpha, \beta) = 1$. Assume that $\alpha \equiv -1 \pmod 4$ and $\beta \equiv 0 \pmod{32}$. Let E be given by $y^2 = x(x - \alpha)(x - \beta)$.

(a) Let p be prime. Show that the cubic polynomial $x(x - \alpha)(x - \beta)$ cannot have a triple root mod p.

(b) Show that the substitution

$$x = 4x_1, \quad y = 8y_1 + 4x_1$$

changes E into E_1, given by

$$y_1^2 + x_1 y_1 = x_1^3 + \frac{-\beta - \alpha - 1}{4} x_1^2 + \frac{\alpha\beta}{16} x_1.$$

(c) Show that the reduction mod 2 of the equation for E_1 is

$$y_1^2 + x_1 y_1 = x_1^3 + e x_1^2$$

for some $e \in \mathbf{F}_2$. This curve is singular at $(0, 0)$.

(d) Let γ be a constant and consider the line $y_1 = \gamma x_1$. Show that if $\gamma^2 + \gamma = e$, then the line intersects the curve in part (c) to order 3, and if $\gamma^2 + \gamma \neq e$ then this line intersects the curve to order 2.

(e) Show that there are two distinct values of $\gamma \in \overline{\mathbf{F}}_2$ such that $\gamma^2 + \gamma = e$. This implies that there are two distinct tangent lines to the curve E_1 mod 2 at (0,0), as in Exercise 2.20.

We take the property of part (e) to be the definition of multiplicative reduction in characteristic 2. Therefore, parts (a) and (e) show that the curve E_1 has good or multiplicative reduction at all primes. A **semistable** elliptic curve over $\mathbf{Q}$ is one that has good or multiplicative reduction at all primes, possibly after a change of variables (over $\mathbf{Q}$) such as the one in part (b). Therefore, E is semistable. See Section 15.1 for a situation where this fact is used.