# Introduction

The theory of Diophantine equations is that branch of number theory which deals with the solution of polynomial equations in either integers or rational numbers. The subject itself is named after one of the greatest of the ancient Greek algebraists, Diophantus of Alexandria,[1] who formulated and solved many such problems.

Most readers will undoubtedly be familiar with Fermat's Last Theorem.[2] This theorem says that if $n \geq 3$ is an integer, then the equation

$$X^n + Y^n = Z^n$$

has no solutions in non-zero integers $X$, $Y$, $Z$. Equivalently, the only solutions in rational numbers to the equation

$$x^n + y^n = 1$$

are those with either $x = 0$ or $y = 0$. Fermat's Theorem is now known to be true for all exponents $n \leq 125000$, so it is unlikely that anyone will find a counterexample by random guessing. On the other hand, there are still a lot of possible exponents left to check between 125000 and infinity!

As another example, we consider the problem of writing an integer as the difference of a square and a cube. In other words, we fix an integer $c \in \mathbb{Z}$ and look for solutions to the Diophantine equation[3]

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

[1] Diophantus lived sometime before the $3^{\text{rd}}$ century A.D. He wrote the *Arithmetica*, a treatise on algebra and number theory in 13 volumes, of which 6 volumes have survived.

[2] Fermat's Last "Theorem" is really a conjecture, because it is still unsolved after more than 350 years. Fermat stated his "Theorem" as a marginal note in his copy of Diophantus' *Arithmetica*; unfortunately, the margin was too small for him to write down his proof!

[3] This equation is often called Bachet's equation, after the $17^{\text{th}}$ century mathematician who originally discovered the duplication formula. It is also sometimes called Mordell's equation, in honor of the $20^{\text{th}}$ century mathematician L.J. Mordell, who made a fundamental contribution to the solution of this and many similar Diophantine equations. We will be proving a special case of Mordell's theorem in Chapter III.

Suppose we are interested in solutions in rational numbers $x, y \in \mathbb{Q}$. An amazing property of this equation is the existence of a *duplication formula*, discovered by Bachet in 1621. If $(x, y)$ is a solution with $x$ and $y$ rational, then it is easy to check that

$$\left(\frac{x^4 - 8cx}{4y^2}, \frac{-x^6 - 20cx^3 + 8c^2}{8y^3}\right)$$

is a solution in rational numbers to the same equation. Further, it is possible to prove (although Bachet was not able to) that if the original solution has $xy \neq 0$ and if $c \neq 1, -432$, then repeating this process leads to infinitely many distinct solutions. So if an integer can be expressed as the difference of a square and a cube of non-zero rational numbers, then it can be so expressed in infinitely many ways. For example, if we start with the solution $(3, 5)$ to the equation

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

and apply Bachet's duplication formula, we find a sequence of solutions that starts

$$(3,5), \quad \left(\frac{129}{10^2}, \frac{-383}{10^3}\right), \quad \left(\frac{2340922881}{7660^2}, \frac{113259286337292}{7660^3}\right), \ldots .$$

As you can see, the numbers rapidly get extremely large.

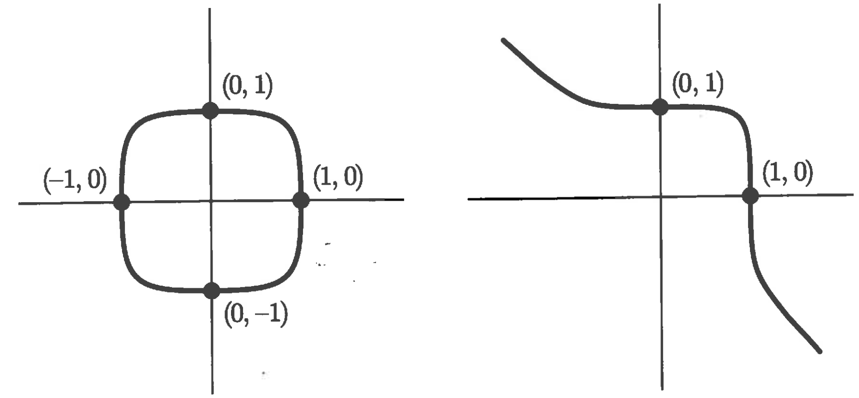
Next we'll take the same equation

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

and ask for solutions in integers $x, y \in \mathbb{Z}$. In the 1650's Fermat posed as a challenge to the English mathematical community the problem of showing that the equation $y^2 - x^3 = -2$ has only two solutions in integers, namely $(3, \pm 5)$. This is in marked contrast to the question of solutions in rational numbers, since we have just seen there are infinitely many of those. None of Fermat's contemporaries appears to have solved the problem, which was solved incorrectly by Euler in the 1730's, and given a correct proof 150 years later! Then in 1908, Axel Thue[4] made a tremendous breakthrough; he showed that for any non-zero integer $c$, the equation $y^2 - x^3 = c$ can have only a finite number of solutions in integers $x, y$. This is a tremendous (qualitative) generalization of Fermat's challenge problem; among the infinitely many solutions in rational numbers, there can be but finitely many integer solutions.

[4] Axel Thue made important contributions to the theory of Diophantine equations, especially to the problem of showing that certain equations have only finitely many solutions in integers. These theorems about integer solutions were generalized by C.L. Siegel during the 1920's and 1930's. We will prove a version of the Thue-Siegel theorem (actually a special case of Thue's original result) in Chapter V.





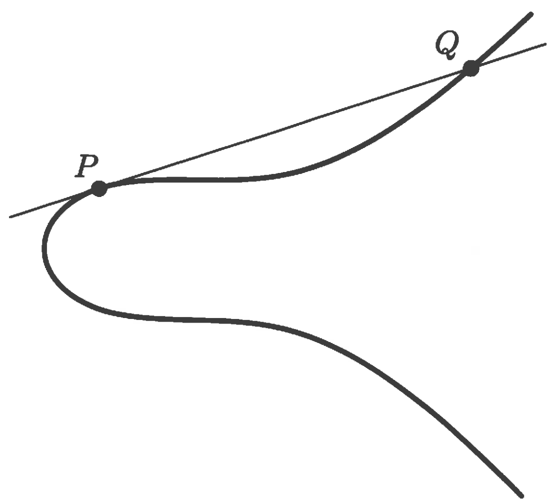

The Fermat Curves $x^4 + y^4 = 1$ and $x^5 + y^5 = 1$

**Figure 0.1**

The 17[th] century witnessed Descartes' introduction of coordinates into geometry, a revolutionary development which allowed geometric problems to be solved algebraically and algebraic problems to be studied geometrically. For example, if $n$ is even, then the real solutions to Fermat's equation $x^n + y^n = 1$ in the $xy$ plane form a geometric object that looks like a squashed circle. Fermat's Theorem is then equivalent to the assertion that the only points on that squashed circle having rational coordinates are the four points $(\pm 1, 0)$ and $(0, \pm 1)$. The Fermat equations with odd exponents look a bit different. We have illustrated the Fermat curves with exponents 4 and 5 in Figure 0.1.

Similarly, we can look at Bachet's equation $y^2 - x^3 = c$, which we have graphed in Figure 0.2. Recall that Bachet discovered a duplication formula which allows us to take a given rational solution and produce a new rational solution. Bachet's formula is rather complicated, and one might wonder where it comes from. The answer is, it comes from geometry! Thus, suppose we let $P = (x, y)$ be our original solution, so $P$ is a point on the curve (as illustrated in Figure 0.2). Next we draw the tangent line to the curve at the point $P$, an easy exercise suitable for a first semester calculus course.[5] This tangent line will intersect the curve at one further point, which we have labeled $Q$. Then, if you work out the algebra to calculate the coordinates of $Q$, you will find Bachet's duplication formula. So Bachet's complicated algebraic formula has a simple geometric interpretation in terms of the intersection of a tangent line with a curve. This is our first intimation of the fruitful interplay that is possible among algebra, number theory, and geometry.

[5] Of course, Bachet had neither calculus nor analytic geometry; so he probably discovered his formula by clever algebraic manipulation.




Bachet's Equation $y^2 - x^3 = c$

**Figure 0.2**

The simplest sort of Diophantine equation is a polynomial equation in one variable:

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0.$$

Assuming that $a_0, \ldots, a_n$ are integers, how can we find all integer and all rational solutions? Gauss' lemma provides the simple answer. If $p/q$ is a rational solution written in lowest terms, then Gauss' lemma tells us that $q$ divides $a_n$ and $p$ divides $a_0$. This gives us a small list of possible rational solutions, and we can substitute each of them into the equation to determine the actual solutions. So Diophantine equations in one variable are easy.

When we move to Diophantine equations in two variables, the situation changes dramatically. Suppose we take a polynomial $f(x, y)$ with integer coefficients and look at the equation

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

For example, Fermat's and Bachet's equations are equations of this sort. Here are some natural questions we might ask:

(a) Are there any solutions in integers?
(b) Are there any solutions in rational numbers?
(c) Are there infinitely many solutions in integers?
(d) Are there infinitely many solutions in rational numbers?

In this generality, only question (c) has been fully answered, although much progress has recently been made on (d).[6]

[6] For polynomials $f(x_1,\ldots,x_n)$ with more than two variables, our four questions have only been answered for some very special sorts of equations. Even worse, work of Davis, Matijasevič, and Robinson has shown that in general it is not possible to find a solution to question (a). That is, there does not exist an algorithm which takes as input the polynomial $f$ and produces as output either "YES" or "NO" as an answer to question (a).



The set of real solutions to an equation $f(x, y) = 0$ forms a curve in the $xy$ plane. Such curves are often called *algebraic curves* to indicate that they are the solutions of a polynomial equation. In trying to answer questions (a)–(d), we might begin by looking at simple polynomials, such as polynomials of degree 1 (also called linear polynomials, because their graphs are straight lines.) For a linear equation

$$ax + by = c$$

with integer coefficients, it is easy to answer our questions. There are always infinitely many rational solutions, there are no integer solutions if $\gcd(a, b)$ does not divide $c$, and otherwise there are infinitely many integer solutions. So linear equations are even easier than equations in one variable.

Next we might turn to polynomials of degree 2 (also called quadratic polynomials). Their graphs are conic sections. It turns out that if such an equation has one rational solution, then there are infinitely many. The complete set of solutions can be described very easily using geometry. We will explain how this is done in the first section of Chapter I. We will also briefly indicate how to answer question (b) for quadratic polynomials. So although it would be untrue to say that quadratic polynomials are easy, it is fair to say that their solutions are completely understood.

This brings us to the main topic of this book, namely, the solution of degree 3 polynomial equations in rational numbers and in integers. One example of such an equation is Bachet's equation $y^2 - x^3 = c$ which we looked at earlier; some other examples which will appear during our studies are

$$y^2 = x^3 + ax^2 + bx + c \qquad \text{and} \qquad ax^3 + by^3 = c.$$

The real solutions to these equations are called *cubic curves* or *elliptic curves*. (However, they are not ellipses, since ellipses are conic sections, and conic sections are given by quadratic equations! The curious chain of events that led to elliptic curves being so named will be recounted in Chapter I, Section 3.) In contrast to linear and quadratic equations, the rational and integer solutions to cubic equations are still not completely understood; and even in those cases where the complete answers are known, the proofs involve a subtle blend of techniques from algebra, number theory, and geometry. Our main goal in this book is to introduce you to the beautiful subject of Diophantine equations by studying in depth the first case of such equations which are still imperfectly understood, namely cubic equations in two variables. To give you an idea of the sorts of results we will be studying, we briefly indicate what is known about questions (a)–(d).

First, a cubic equation has only finitely many integer solutions[7] (Siegel, 1920's); and there is an explicit upper bound for the largest solution in terms of the coefficients of the polynomial (Baker-Coates, 1970). This provides a satisfactory answer to (a) and (c), although the actual bounds for the largest solution are generally too large to be practical. We will prove a special case of Siegel's theorem (for equations of the form $ax^3 + by^3 = c$) in Chapter V.

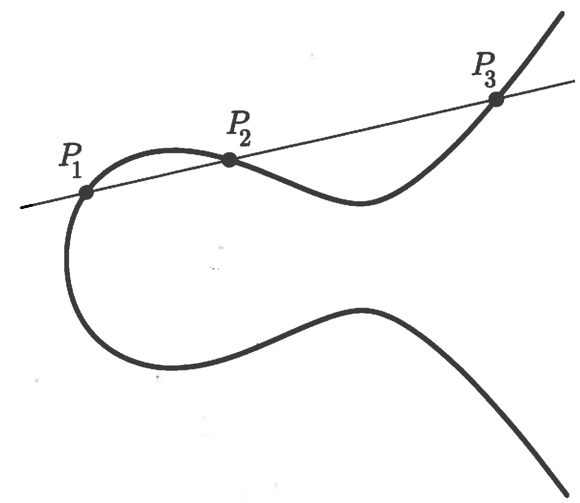
Second, all of the (possibly infinitely many) rational solutions to a cubic equation may be found by starting with a finite set of solutions and repeatedly applying a geometric procedure similar to Bachet's duplication formula. The fact that there exists such a finite generating set was suggested by Poincaré in 1901 and proven by L.J. Mordell in 1923. We will prove (a special case of) Mordell's theorem in Chapter III. However, we must in truth point out that Mordell's theorem does not really answer questions (b) and (d). As we will see, the proof of Mordell's theorem gives a procedure which *often* allows one to find a finite generating set for the set of rational solutions. But it is only conjectured, and not yet proven, that Mordell's method always yields a generating set. So even for special sorts of cubic equations, such as $y^2 - x^3 = c$ and $ax^3 + by^3 = c$, there is no general method (i.e., algorithm) currently known which is guaranteed to answer question (b) or (d).

We have mentioned several times the idea that the study of Diophantine equations involves an interplay among algebra, number theory, and geometry. The geometric component is clear, because the equation itself defines (in the case of two variables) a curve in the plane; and we have already seen how it may be useful to consider the intersection of that curve with various lines. The number theory is also clearly present, because we are searching for solutions in either integers or rational numbers, and what else is number theory other than a study of the relations between integers and/or rational numbers. But what of the algebra? We could point out that polynomials are essentially algebraic objects. However, algebra plays a much more important role than this.

Recall that Bachet's duplication formula can be described as follows: start with a point $P$ on a cubic curve, draw the tangent line at $P$, and take the third point of intersection of the line with the curve. Similarly, if we start with two points $P_1$ and $P_2$ on the curve, we can draw the line through $P_1$ and $P_2$ and look at the third intersection point $P_3$. (This will work for most points, because the intersection of a line and a cubic curve will usually consist of exactly three points.) We might describe this procedure, which we illustrate in Figure 0.3, as a way to "add" two points on the curve and get a third point on the curve. Amazingly enough, we will show that with a slight modification this geometric operation takes the set of rational

[7] Actually, Siegel's theorem applies only to "non-singular" cubic equations. However, most cubic equations are non-singular; and in practice it is quite easy to check whether or not a given equation is non-singular.






"Adding" Two Points on a Cubic Curve

**Figure 0.3**

solutions to a cubic equation and turns it into an abelian group! And Mordell's theorem alluded to earlier can be rephrased by saying that this group has a finite number of generators. So here is algebra, number theory, and geometry all packaged together in one of the greatest theorems of this century.

We hope that the foregoing introduction has convinced you of some of the beauty and elegance to be found in the theory of Diophantine equations. But the study of Diophantine equations, in particular the theory of elliptic curves, also has its practical applications. We will study one such application in this book. Everyone is familiar with the Fundamental Theorem of Arithmetic, which asserts that every positive integer factors uniquely into a product of primes. It is less well known that if the integer is fairly large, say on the order of $10^{100}$ or $10^{200}$, it may be virtually impossible to perform that factorization. This is true even though there are very quick ways to check that an integer of this size is not itself a prime. In other words, if one is presented with an integer $N$ with (say) 150 digits, then one can easily check that $N$ is not prime, even though one cannot in general find any of the prime factors of $N$.

This curious state of affairs has been used by Rivest, Shamir, and Adleman to construct what is known as a public key cipher based on a trapdoor function. These are ciphers in which one can publish, for all to see, the method of enciphering a message; but even with the encipherment method on hand, a would-be spy will not be able to decipher any messages. Needless to say, such ciphers have numerous applications, ranging from espionage to ensuring secure telecommunications between banks and other financial institutions. To describe the relation with elliptic curves, we will need to briefly indicate how such a "trapdoor cipher" works.

First one chooses two large primes, say $p$ and $q$, each with around 100 digits. Next one publishes the product $N = pq$. In order to encipher a message, your correspondent only needs to know the value of $N$. But in order to decipher a message, the factors $p$ and $q$ are needed. So your messages will be safe as long as no one is able to factor $N$. This means that in order to ensure the safety of your messages, you need to know the largest integers that your enemies are able to factor in a reasonable amount of time.

So how does one factor a large number which is known to be composite? One can start trying possible divisors $2, 3, \ldots$, but this is hopelessly inefficient. Using techniques from number theory, various algorithms have been devised, with exotic sounding names like the continued fraction method, the ideal class group method, the $p-1$ method, and the quadratic sieve method. But one of the best methods currently available is Lenstra's Elliptic Curve Algorithm, which as the name indicates relies on the theory of elliptic curves. So it is essential to understand the strength of Lenstra's algorithm if one is to ensure that one's public key cipher will not be broken. We will describe how Lenstra's algorithm works in Chapter IV.

# CHAPTER I

# Geometry and Arithmetic

## 1. Rational Points on Conics

Everyone knows what a rational number is, a quotient of two integers. We call a point in the $(x, y)$ plane a *rational point* if both its coordinates are rational numbers. We call a line a *rational line* if the equation of the line can be written with rational numbers; that is, if its equation is

$$ax + by + c = 0$$

with $a, b, c$, rational. Now it is pretty obvious that if you have two rational points, the line through them is a rational line. And it is neither hard to guess nor hard to prove that if you have two rational lines, then the point where they intersect is a rational point. If you have two linear equations with rational numbers as coefficients and you solve them, you get rational numbers as answers.

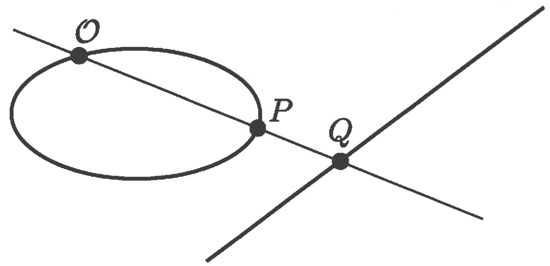
The general subject of these notes is rational points on curves, especially on cubic curves. But as an introduction, we will start with conics. Let

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

be a conic. We will say that the conic is *rational* if we can write its equation with rational numbers.

Now what about the intersection of a rational line with a rational conic? Will it be true that the points of intersection are rational. By writing down some examples, it is easy to see that the answer is, in general, no. If you use analytic geometry to find the coordinates of these points, you will come out with a quadratic equation for the $x$ coordinate of the intersection. And if the conic is rational and the line is rational, the quadratic equation you come out with will have rational coefficients. So the two points of intersection will be rational if and only if the roots of that quadratic equation are rational. In general, they might be conjugate quadratic irrationalities.

However, if one of those points is rational, then so is the other. This is true because if a quadratic equation with rational coefficients has one




Projecting a Conic onto a Line

**Figure 1.1**

rational root, then the other root is rational, because the sum of the roots is the middle coefficient. This very simple idea enables one to describe the rational points on a conic completely. Given a rational conic, the first question is whether or not there are any rational points on it. (We will return to this question later.) But let us suppose that we know of one rational point $\mathcal{O}$ on our rational conic. Then we can get all of them very simply. We just draw some rational line and we project the conic onto the line from this point $\mathcal{O}$. (To project $\mathcal{O}$ itself onto the line, we use the tangent line to the conic at $\mathcal{O}$.)

A line meets a conic in two points, so for every point $P$ on the conic we get a point $Q$ on the line; and conversely, for every point $Q$ on the line, by joining it to the point $\mathcal{O}$, we get a point $P$ on the conic. (See Figure 1.1.) We get a one-to-one correspondence between the points on the conic and points on the line.[†] But now you see by the remarks we have made that if the point $P$ on the conic has rational coordinates, then the point $Q$ on the line will have rational coordinates. And conversely, if $Q$ is rational, then because $\mathcal{O}$ is assumed to be rational, the line through $P$ and $Q$ meets the conic in two points, one of which is rational. So the other point is rational, too. Thus the rational points on the conic are in one-to-one correspondence with the rational points on the line. Of course, the rational points on the line are easily described in terms of rational values of some parameter.

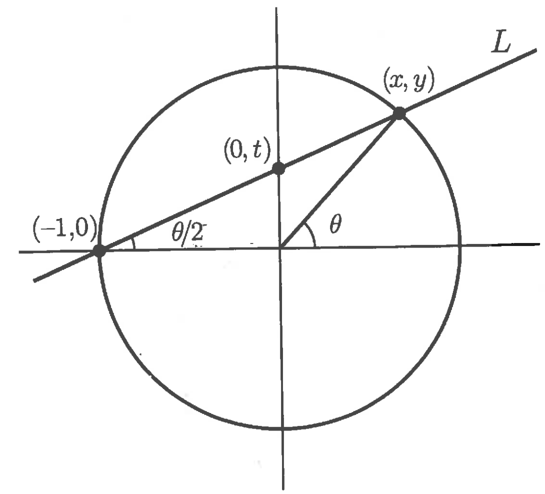
Let's carry out this procedure for the circle

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

We will project from the point $(-1, 0)$ onto the $y$ axis. Let's call the point of intersection $(0, t)$. (See Figure 1.2.) If we know $x$ and $y$, then we can

[†] More precisely, there is a one-to-one correspondence between the points of the line and all but one of the points of the conic. The missing point of the conic is the unique point $\mathcal{O}'$ on the conic such that the line connecting $\mathcal{O}$ and $\mathcal{O}'$ is parallel to the line we are projecting onto. However, if we work in the projective plane and use homogeneous coordinates, then this problem disappears and we get a perfect one-to-one correspondence. See Appendix A for details.






A Rational Parametrization of the Circle

**Figure 1.2**

get $t$ very easily. The equation of the line $L$ connecting $(-1, 0)$ to $(0, t)$ is $y = t(1 + x)$. The point $(x, y)$ is assumed to be on the line $L$ and also on the circle, so we get the relation

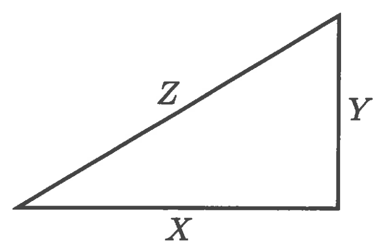
$$1 - x^2 = y^2 = t^2(1 + x)^2.$$

For a fixed value of $t$, this is a quadratic equation whose roots are the $x$ coordinates of the two intersections of the line $L$ with the circle. Clearly, $x = -1$ is a root, because the point $(-1, 0)$ is on both $L$ and the circle. To find the other root, we cancel a factor of $1 + x$ from both sides of the equation. This gives the linear equation $1 - x = t^2(1 + x)$. Solving this for $x$ in terms of $t$, and then using the relation $y = t(1 + x)$ to find $y$, we obtain

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

This is the familiar rational parametrization of the circle . And now the assertion made above is clear from these formulas. That is, if $x$ and $y$ are rational numbers, then $t$ will be a rational number. Conversely, if $t$ is a rational number, then from these formulas it is obvious that the coordinates $x$ and $y$ will be rational numbers. So this is the way you get rational points on the circle: simply plug in an arbitrary rational number for $t$. That will give you all points except $(-1, 0)$. [If you want to get $(-1, 0)$, you must "substitute" infinity for $t$!]

These formulas can be used to solve the elementary problem of describing all right triangles with integer sides. Let us consider the problem of finding some other triangles, besides 3,4,5, which have whole number




A Right Triangle

**Figure 1.3**

sides. Let us call the lengths of the sides $X, Y, Z$. (See Figure 1.3.) That means we want to find integers such that

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

Now if we have such integers where $X$, $Y$, and $Z$ have a common factor, then we can take the common factor out. So we may as well assume that the three of them do not have any common factors. Right triangles whose integer sides have no common factor are called *primitive*. But then it follows that no two of the sides have a common factor, either. For example, if there is some prime dividing both $Y$ and $Z$, then it would divide $X^2 = Z^2 - Y^2$, hence it would divide $X$, contrary to our assumption that $X, Y, Z$ have no common factor. So if we make the trivial reduction to the case of primitive triangles, then no two of the sides have a common factor.

In particular, the point $(x, y)$ defined by

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

is a rational point on the circle $x^2+y^2 = 1$. Further, the rational numbers $x$ and $y$ are in lowest terms.

Since $X$ and $Y$ have no common factor, they cannot both be even. We claim that neither can they both be odd. The point is that the square of an odd number is congruent to 1 modulo 4. If $X$ and $Y$ were both odd, then $X^2 + Y^2$ would be congruent to 2 modulo 4. But $X^2 + Y^2 = Z^2$, and $Z^2$ is congruent to either 0 or 1 modulo 4. So $X$ and $Y$ are not both odd. Say $X$ is odd and $Y$ is even.

Since $(x, y)$ is a rational point on the circle, there is some rational number $t$ so that $x$ and $y$ are given by the formulas we derived above. Write $t = m/n$ in lowest terms. Then

$$\frac{X}{Z} = x = \frac{n^2 - m^2}{n^2 + m^2}, \qquad \frac{Y}{Z} = y = \frac{2mn}{n^2 + m^2}.$$



Since $X/Z$ and $Y/Z$ are in lowest terms, this means that there is some integer $\lambda$ so that

$$\lambda Z = n^2 + m^2, \qquad \lambda Y = 2mn, \qquad \lambda X = n^2 - m^2.$$

We want to show that $\lambda = 1$. Because $\lambda$ divides both $n^2+m^2$ and $n^2 - m^2$, it divides their sum $2n^2$ and their difference $2m^2$. But $m$ and $n$ have no common divisors. Hence, $\lambda$ divides 2, and so $\lambda = 1$ or $\lambda = 2$. If $\lambda = 2$, then $n^2 - m^2 = \lambda X$ is divisible by 2, but not by 4, because we are assuming that $X$ is odd. In other words, $n^2 - m^2$ is congruent to 2 modulo 4. But $n^2$ and $m^2$ are each congruent to either 0 or 1 modulo 4, so this is not possible. Hence, $\lambda = 1$.

This proves that to get all primitive triangles, you take two relatively prime integers $m$ and $n$ and let

$$X = n^2 - m^2, \qquad Y = 2mn, \qquad Z = n^2 + m^2,$$

be the sides of the triangle. These are the ones with $X$ odd and $Y$ even. The others are obtained by interchanging $X$ and $Y$.

These formulas have other uses; you may have met them in calculus . In Figure 1.2, we have

$$x = \cos\theta, \quad y = \sin\theta; \qquad \text{and so} \quad t = \tan\tfrac{1}{2}\theta = \frac{\sin\theta}{1+\cos\theta}.$$

So the formulas given above allow us to express cosine and sine rationally in terms of the tangent of the half angle:

$$x = \cos\theta = \frac{1-t^2}{1+t^2}, \qquad y = \sin\theta = \frac{2t}{1+t^2}.$$

If you have some complicated identity in sine and cosine that you want to test, all you have to do is substitute these formulas, collect powers of $t$, and see if you get zero. (If they had told you this in high school, the whole business of trigonometric identities would have become a trivial exercise in algebra!)

Another use comes from the observation that these formulas let us express all trigonometric functions of an angle $\theta$ as rational expressions in $t = \tan(\theta/2)$. Note that

$$\theta = 2\arctan(t), \qquad d\theta = \frac{2dt}{1+t^2}.$$

So if you have an integral which involves $\cos\theta$ and $\sin\theta$ and $d\theta$, and you make the appropriate substitutions, then you transform it into an integral in $t$ and $dt$. If the integral is a rational function of $\sin\theta$ and $\cos\theta$, you obviously come out with the integral of a rational function of $t$. Since rational functions can be integrated in terms of elementary functions, it follows that any rational function of $\sin\theta$ and $\cos\theta$ can be integrated in terms of elementary functions.

What if we take the circle

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

and ask to find the rational points on it? That is the easiest problem of all, because the answer is that there are none. It is impossible for the sum of the squares of two rational numbers to equal 3. How can we see that it is impossible?

If there is a rational point, we can write it as

$$x = \frac{X}{Z} \quad \text{and} \quad y = \frac{Y}{Z}$$

for some integers $X, Y, Z$; and then

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

If $X, Y, Z$ have a common factor, then we can remove it; so we may assume that they have no common factor. It follows that both $X$ and $Y$ are not divisible by 3. This is true because if 3 divides $X$, then 3 divides $Y^2 = 3Z^2 - X^2$, so 3 divides $Y$. But then 9 divides $X^2 + Y^2 = 3Z^2$, so 3 divides $Z$, contradicting the fact that $X, Y, Z$ have no common factors. Hence 3 does not divide $X$, and similarly for $Y$.

Since $X$ and $Y$ are not divisible by 3, we have

$$X \equiv \pm 1 \pmod 3), \quad Y \equiv \pm 1 \pmod 3, \quad \text{and so} \quad X^2 \equiv Y^2 \equiv 1 \pmod 3.$$

But then

$$0 \equiv 3Z^2 = X^2 + Y^2 \equiv 1 + 1 \equiv 2 \pmod 3.$$

This contradiction shows that no two rational numbers have squares which add up to 3.

We have seen by the projection argument that if you have one rational point on a rational conic, then all of the rational points can be described in terms of a rational parameter $t$. But how do you check whether or not there is one rational point? The argument we gave for $x^2 + y^2 = 3$ gives the clue. We showed that there were no rational points by checking that a certain equation had no solutions modulo 3.

There is a general method to test, in a finite number of steps, whether or not a given rational conic has a rational point. The method consists in seeing whether a certain congruence can be satisfied. The theorem goes back to Legendre. Let us take the simple case

$$aX^2 + bY^2 = cZ^2,$$



which is to be solved in integers. Legendre's theorem states that there is an integer $m$, depending in a simple fashion on $a$, $b$, and $c$, so that the above equation has a solution in integers, not all zero, if and only if the congruence

$$aX^2 + bY^2 \equiv cZ^2 \pmod m$$

has a solution in integers relatively prime to $m$.

There is a much more elegant way to state this theorem, due to Hasse: "A homogeneous quadratic equation in several variables is solvable by integers, not all zero, if and only if it is solvable in real numbers and in $p$-adic numbers for each prime $p$." Once one has Hasse's result, then one gets Legendre's theorem in a fairly elementary way. Legendre's theorem combined with the work we did earlier provides a very satisfactory answer to the question of rational points on rational conics. So now we move on to cubics.

## 2. The Geometry of Cubic Curves

Now we are ready to begin our study of cubics. Let

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

be the equation for a general cubic. We will say that a cubic is *rational* if the coefficients of its equation are rational numbers. A famous example is

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

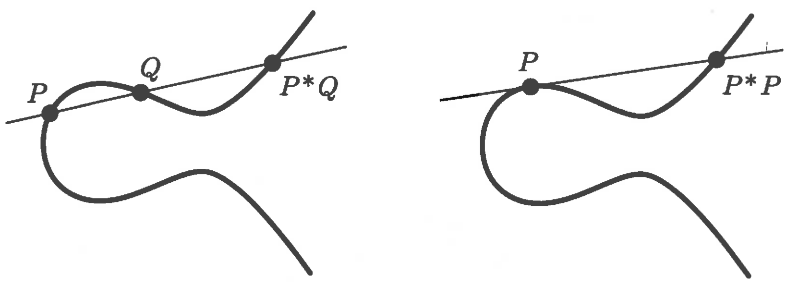
or, in homogeneous form,

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

To find rational solutions of $x^3 + y^3 = 1$ amounts to finding integer solutions of $X^3 + Y^3 = Z^3$, the first non-trivial case of Fermat's Last "Theorem."

We cannot use the geometric principle that worked so well for conics because a line generally meets a cubic in three points. And if we have one rational point, we cannot project the cubic onto a line, because each point on the line would then correspond to two points on the curve.

But there is a geometric principle we can use. If we can find two rational points on the curve, then we can generally find a third one. Namely, draw the line connecting the two points you have found. This will be a rational line, and it meets the cubic in one more point. If we look and see what happens when we try to find the three intersections of a rational line with a rational cubic, we find that we come out with a cubic equation with rational coefficients. If two of the roots are rational, then the third must be also. We will work out some explicit examples below, but the principle is clear. So this gives some kind of composition law: Starting with two points $P$ and $Q$, we draw the line through $P$ and $Q$ and let $P * Q$ denote the third point of intersection of the line with the cubic. (See Figure 1.4.)




The Composition of Points on a Cubic
**Figure 1.4**

Even if we only have one rational point $P$, we can still generally get another. By drawing the tangent line to the cubic at $P$, we are essentially drawing the line through $P$ and $P$. The tangent line meets the cubic twice at $P$, and the same argument will show that the third intersection point is rational. Then we can join these new points up and get more points. So if we start with a few rational points, then by drawing lines, we generally get lots of others.

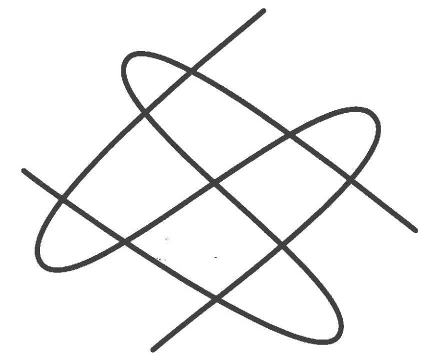
One of the main theorems we want to prove in this book is the theorem of Mordell (1921) which states that if $C$ is a non-singular rational cubic curve, then there is a *finite* set of rational points such that all other rational points can be obtained in the way we described. We will prove Mordell's theorem for a wide class of cubic curves, using only elementary number theory of the ordinary integers. The principle of the proof in the general case is exactly the same, but requires some tools from the theory of algebraic numbers.

The theorem can be reformulated to be more enlightening. To do that, we first give an elementary geometric property of cubics. We will not prove it completely, but we will make it very plausible, which should suffice. (For more details, see Appendix A.) In general, two cubic curves meet in nine points. To make that statement correct, one should first of all use the projective plane, which has extra points at infinity. Secondly, one should introduce multiplicities of intersections, counting points of tangency for example as intersections of multiplicity greater than one. And finally, one must allow complex numbers for coordinates. We will ignore these technicalities. Then a curve of degree $m$ and a curve of degree $n$ meet in $mn$ points. This is Bezout's theorem, one of the basic theorems in the theory of plane curves. (See Appendix A, Section 4, for a proof of Bezout's theorem.) So two cubics meet in nine points. (See Figure 1.5.)

The theorem that we want to use is the following:

> Let $C$, $C_1$, and $C_2$ be three cubic curves. Suppose $C$ goes through eight of the nine intersection points of $C_1$ and $C_2$. Then $C$ goes through the ninth intersection point.





The Intersection of Two Cubics
**Figure 1.5**

Why is this true, at least in general? The trick is to consider the problem of constructing a cubic curve which goes through a certain set of points. To define a cubic curve, we have to give ten coefficients $a, b, c, d, e, f, g, h, i, j$. If we multiply all the coefficients by a constant, then we get the same curve. So really the set of all possible cubics is, so to speak, nine dimensional. And if we want the cubic to go through a point whose $x$ and $y$ coordinates are given, that imposes one linear condition on those coefficients. The set of cubics which go through one given point is, so to speak, eight dimensional. Each time you impose a condition that the cubic should go through a given point, that imposes an extra linear condition on the coefficients. Thus, the family of all cubics which go through the eight points of intersection of the two given cubics $C_1$ and $C_2$ forms a one dimensional family.

Let $F_1(x,y) = 0$ and $F_2(x,y) = 0$ be the cubic equations giving $C_1$ and $C_2$. We can then find cubics going through the eight points by taking linear combinations $\lambda_1 F_1 + \lambda_2 F_2$. Because the cubics going through the eight points form a one dimensional family, and because the set of cubics $\lambda_1 F_1 + \lambda_2 F_2$ is a one dimensional family, we see that the cubic $C$ has an equation $\lambda_1 F_1 + \lambda_2 F_2 = 0$ for a suitable choice of $\lambda_1, \lambda_2$.

Now how about the ninth point? Since that ninth point is on both $C_1$ and $C_2$, we know that $F_1(x,y)$ and $F_2(x,y)$ both vanish at that point. It follows that $\lambda_1 F_1 + \lambda_2 F_2$ also vanishes there, and this means that $C$ also contains that point.

In passing, we will mention that there is no known method to determine in a finite number of steps whether a given rational cubic has a rational point. There is no analogue of Hasse's theorem for cubics. That question is still open, and it is a very important question. The idea of looking modulo $m$ for all integers $m$ is not sufficient. Selmer gave the example

$$3X^3 + 4Y^3 + 5Z^3 = 0.$$

This is a cubic, and Selmer shows by an ingenious argument that it has no integer solutions other than $(0,0,0)$. However, one can check that for every integer $m$, the congruence

$$3X^3 + 4Y^3 + 5Z^3 \equiv 0 \pmod{m}$$

has a solution in integers with no common factor. So for general cubics, the existence of a solution modulo $m$ for all $m$ does not ensure that a rational solution exists. We will leave this difficult problem aside, and assume that we have a cubic which has a rational point $\mathcal{O}$.

We want to reformulate Mordell's theorem in a way which has great aesthetic and technical advantages. If we have any two rational points on a rational cubic, say $P$ and $Q$, then we can draw the line joining $P$ to $Q$, obtaining the third point which we denoted $P * Q$. This has the flavor of many of the constructions you have studied in modern algebra. If we consider the set of all rational points on the cubic, we can say that set has a law of composition. Given any two points $P$, $Q$, we have defined a third point $P * Q$. We might ask about the algebraic structure of this set and this composition law; for example, is it a group? Unfortunately, it is not a group; to start with, it is fairly clear that there is no identity element.

However, by playing around with it a bit, we can make it into a group in such a way that the given rational point $\mathcal{O}$ becomes the zero element of the group. We will denote the group law by $+$ because it is going to be a commutative group. The rule is as follows:

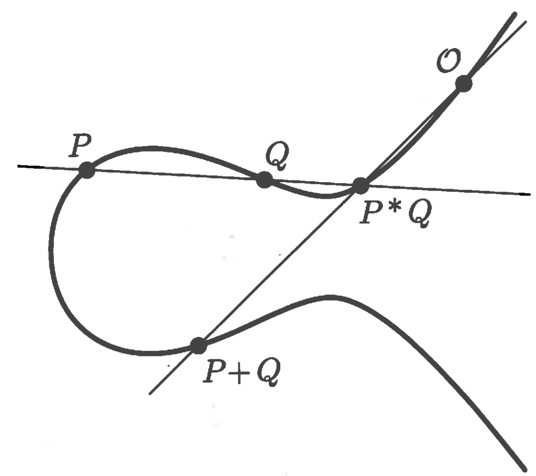
> To add $P$ and $Q$, take the third intersection point $P*Q$, join it to $\mathcal{O}$, and then take the third intersection point to be $P+Q$. Thus by definition, $P+Q = \mathcal{O} * (P * Q)$.

The group law is illustrated in Figure 1.6, and the fact that $\mathcal{O}$ acts as the zero element is shown in Figure 1.7.

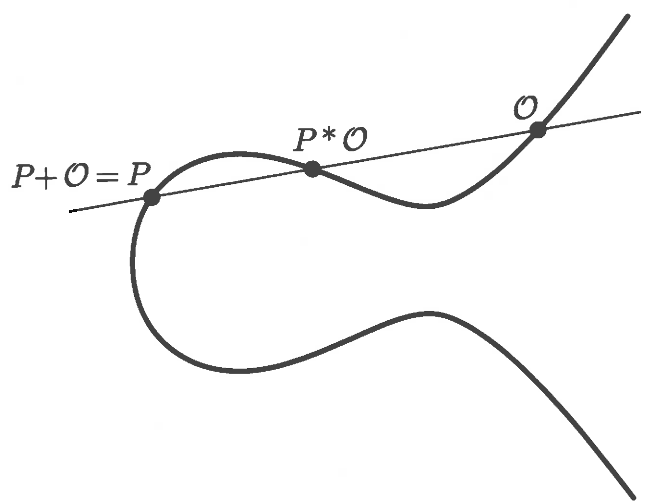
It is clear that this operation is commutative, that is, $P+Q = Q+P$. We claim first that $P + \mathcal{O} = P$, so $\mathcal{O}$ acts as the zero element. Why is that? Well, if we join $P$ to $\mathcal{O}$, then we get the point $P * \mathcal{O}$ as the third intersection. Next we must join $P*\mathcal{O}$ to $\mathcal{O}$ and take the third intersection point. That third intersection point is clearly $P$. So $P + \mathcal{O} = P$.

It is a little harder to get negatives, but not very hard. Draw the tangent line to the cubic at $\mathcal{O}$, and let the tangent meet the cubic at the additional point $S$. (We are assuming that the cubic is non-singular, so there is a tangent line at every point.) Then given a point $Q$, we join $Q$ to $S$, and the third intersection point $Q * S$ will be $-Q$. (See Figure 1.8.) To check that this is so, let us add $Q$ and $-Q$. To do this we take the third intersection of the line through $Q$ and $-Q$, which is $S$; and then join $S$ to $\mathcal{O}$ and take the third intersection point $S * \mathcal{O}$. But the line through $S$ and $\mathcal{O}$, because it is tangent to the cubic at $\mathcal{O}$, meets the cubic once at $S$ and twice at $\mathcal{O}$. (You must interpret these things properly.) So the third intersection is the second time it meets $\mathcal{O}$. Therefore, $Q + (-Q) = \mathcal{O}$.



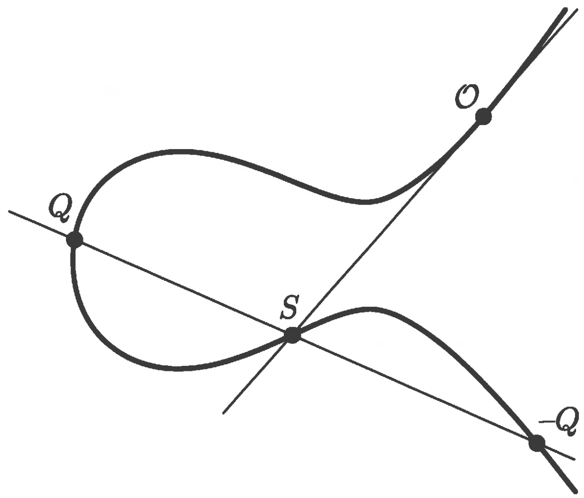



The Group Law on a Cubic
**Figure 1.6**




Verifying $\mathcal{O}$ Is the Zero Element
**Figure 1.7**

If we only knew that $+$ was associative, then we would have a group. Let us try to prove the associative law. Let $P, Q, R$ be three points on the curve. We want to prove that $(P+Q)+R = P+(Q+R)$. To get $P+Q$, we form $P*Q$ and take the third intersection of the line connecting it to $\mathcal{O}$. To add $P+Q$ to $R$, we draw the line from $R$ through $P+Q$, and that meets the curve at $(P+Q)*R$. To get $(P+Q)+R$, we would have to join $(P+Q)*R$




The Negative of a Point

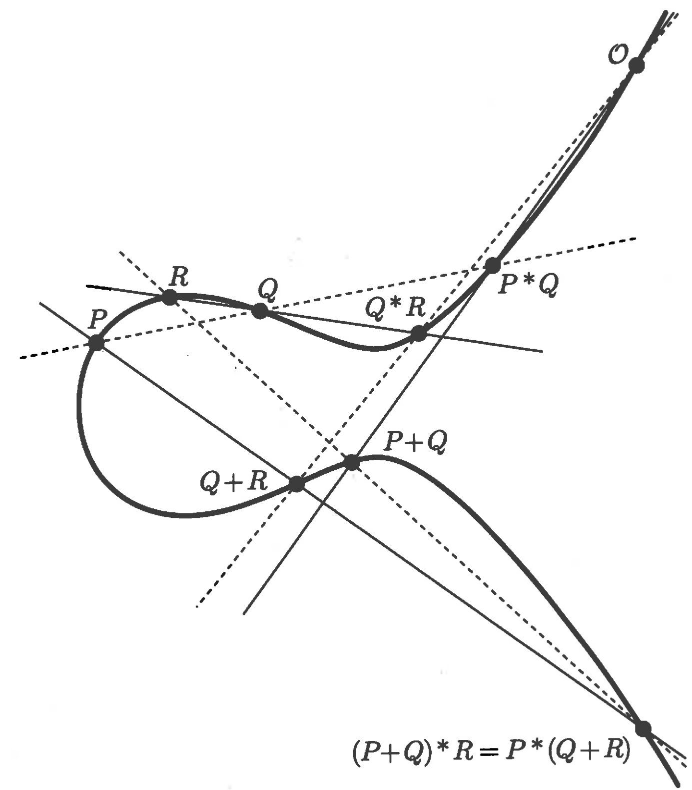
**Figure 1.8**

to $\mathcal{O}$ and take the third intersection. Now that does not show up too well in the picture, but to show $(P+Q)+R = P+(Q+R)$, it will be enough to show that $(P+Q)*R = P*(Q+R)$. To form $P*(Q+R)$, we first have to find $Q*R$, join that to $\mathcal{O}$, and take the third intersection which is $Q+R$. Then we must join $Q+R$ to $P$, which gives the point $P*(Q+R)$; and that should be the same as $(P+Q)*R$. In Figure 1.9, each of the points $\mathcal{O}, P, Q, R, P*Q, P+Q, Q*R, Q+R$ lies on one of the dotted lines and one of the solid lines. Let us consider the dotted line through $P+Q$ and $R$ and the solid line through $P$ and $Q+R$. Does their intersection lie on the cubic? If so, the we will have proven that $P*(Q+R) = (P+Q)*R$.

We have nine points: $\mathcal{O}, P, Q, R, P*Q, P+Q, Q*R, Q+R$ and the intersection of the two lines. So we have two (degenerate) cubics that go through the nine points. Namely, a line has a linear equation, and if you have three linear equations and multiply them together, you get a cubic equation. The set of solutions to that cubic equation is just the union of the three lines. Now we apply our theorem, taking for $C_1$ the union of the three dotted lines and for $C_2$ the union of the three solid lines. By construction, these two cubics go through the nine points. But the original cubic curve $C$ goes through eight of the points, and therefore it goes through the ninth. Thus, the intersection of the two lines lies on $C$, which proves that $(P+Q)*R = P*(Q+R)$.

We will not do any more toward proving that the operation $+$ makes the points of $C$ into a group. Later, when we have a normal form, we will give explicit formulas for adding points. So if our use of unproven assertions bothers you, then you can spend a day or two computing with those explicit






Verifying the Associative Law

**Figure 1.9**

formulas and verify directly that associativity holds.

We also want to mention that there is nothing special about our choice of $\mathcal{O}$; if we choose a different point $\mathcal{O}'$ to be the zero element of our group, then we get a group with exactly the same structure. In fact, the map

$$P \longmapsto P + (\mathcal{O}' - \mathcal{O})$$

is an isomorphism from the group "$C$ with zero element $\mathcal{O}$" to the group "$C$ with zero element $\mathcal{O}'$."

Maybe we should explain that we have dodged some of the subtleties. If the line through $P$ and $Q$ is tangent to the curve at $P$, then the third point of intersection must be interpreted as $P$. And if you think of that tangent line as the line through $P$ and $P$, the third intersection point is $Q$. Further, if you have a point of inflection $P$ on $C$, then the tangent line $P$ meets the curve three times at $P$. So in this case the third point of intersection for the line through $P$ and $P$ is again $P$. [That is, if $P$ is an inflection point, then $P*P = P$.] You just have to count intersections in the proper way, and it is clear why if you think of the points varying a little bit. But to put everything on solid ground is a big task. If you are going into this business, it is important to start with better foundations and from a more general point of view. Then all these questions would be taken care of.

How does this allow us to reformulate Mordell's theorem? Mordell's theorem says that we can get all of the rational points by starting with a finite set, drawing lines through those points to get new points, then drawing lines through the new points to get more points, and so on. In terms of the group law, this says that the group of rational points is finitely generated. So we have the following statement of Mordell's theorem.

> **Mordell's Theorem.** If a non-singular plane cubic curve has a rational point, then the group of rational points is finitely generated.

This version is obviously technically a much better form because we can use a little elementary group theory, nothing very deep, but a convenient device in the proof.

## 3. Weierstrass Normal Form

We are going to prove Mordell's theorem as Mordell did, using explicit formulas for the addition law. To make these formulas as simple as possible, it is important to know that any cubic with a rational point can be transformed into a certain special form called *Weierstrass normal form*. We will not completely prove this, but we will give enough of an indication of the proof so that anyone who is familiar with projective geometry can carry out the details. (See Appendix A for an introduction to projective geometry.) Also, we will work out a specific example to illustrate the general theory. After that, we will restrict attention to cubics which are given in the Weierstrass form, which classically consists of equations that look like

$$y^2 = 4x^3 - g_2x - g_3.$$

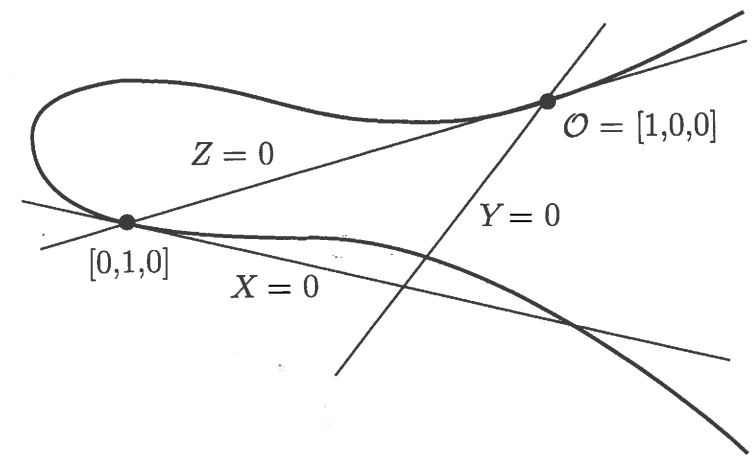
We will also use the slightly more general equation

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

and will call either of them Weierstrass form. What we need to show is that any cubic is, as one says, birationally equivalent to a cubic of this type. We will now explain what this means, assuming that the reader knows a (very) little bit of projective geometry.

We start with a cubic curve, which we will think of as being in the projective plane. The idea is to choose axes in the projective plane so that the equation for the curve has a simple form. We assume we are given






Choosing Axes to Put $C$ into Weierstrass Form

**Figure 1.10**

a rational point $\mathcal{O}$ on $C$, so we begin by taking $Z = 0$ to be the tangent line to $C$ at $\mathcal{O}$. This tangent line intersects $C$ at one other point, and we take the $X = 0$ axis to be tangent to $C$ at this new point. Finally, we choose $Y = 0$ to be any line (other than $Z = 0$) which goes through $\mathcal{O}$. (See Figure 1.10. We are assuming that $\mathcal{O}$ is not a point of inflection; otherwise we can take $X = 0$ to be any line not containing $\mathcal{O}$.)

If we choose axes in this fashion and let $x = \dfrac{X}{Z}$ and $y = \dfrac{Y}{Z}$, then we get some linear conditions on the form the equation will take in these coordinates. This is called a projective transformation. We will not work out the algebra, but will just tell you that at the end the equation for $C$ takes the form

$$xy^2 + (ax + b)y = cx^2 + dx + e.$$

Next we multiply through by $x$,

$$(xy)^2 + (ax + b)xy = cx^3 + dx^2 + ex.$$

Now if we give a new name to $xy$, we will just call it $y$ again, then we obtain

$$y^2 + (ax + b)y = \text{cubic in } x.$$

Replacing $y$ by $y - \frac{1}{2}(ax + b)$, another linear transformation which amounts to completing the square on the left-hand side of the equation, we obtain

$$y^2 = \text{cubic in } x.$$

The cubic in $x$ might not have leading coefficient 1, but we can adjust that by replacing $x$ and $y$ by $\lambda x$ and $\lambda^2 y$, where $\lambda$ is the leading coefficient of the cubic. So we do finally get an equation in Weierstrass form. And if we want to get rid of the $x^2$ term in the cubic, replace $x$ by $x - \alpha$ for an appropriate choice of $\alpha$.

Tracing through all of the transformations from the original coordinates to the new coordinates, we see that the transformation is not linear, but it is rational. In other words, the new coordinates are given as ratios of polynomials in the old coordinates. Hence, rational points on the original curve correspond to rational points on the new curve.

An example should make all of this clear. Suppose we start with a cubic of the form

$$u^3 + v^3 = \alpha,$$

where $\alpha$ is a given rational number. The homogeneous form of this equation is $U^3+V^3 = \alpha W^3$, so in the projective plane this curve contains the rational point $[1,-1,0]$. Applying the above procedure (note that $[1,-1,0]$ is an inflection point) leads to new coordinates $x$ and $y$ which are given in terms of $u$ and $v$ by the rational functions

$$x = \frac{12\alpha}{u+v} \qquad \text{and} \qquad y = 36\alpha\frac{u-v}{u+v}.$$

If you work everything out, you will see that $x$ and $y$ satisfy the Weierstrass equation

$$y^2 = x^3 - 432\alpha^2.$$

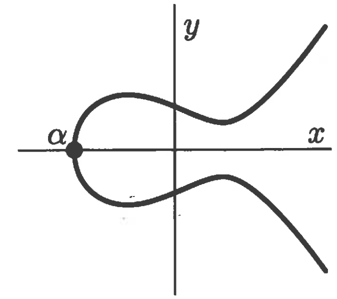
Further, the process can be inverted, and one finds that $u$ and $v$ can be expressed in terms of $x$ and $y$ by

$$u = \frac{36\alpha + y}{6x} \qquad \text{and} \qquad v = \frac{36\alpha - y}{6x}.$$

Thus if we have a rational solution to $u^3+v^3 = \alpha$, then we get rational $x$ and $y$ which satisfy the equation $y^2 = x^3 - 432\alpha^2$. And conversely, if we have a rational solution of $y^2 = x^3 - 432\alpha^2$, then we get rational numbers satisfying $u^3 + v^3 = \alpha$. Of course, if $u = -v$, then the denominator in the expression for $x$ and $y$ is zero; but there are only a finite number of such exceptions, and they are easy to find. So the problem of finding rational points on $u^3 + v^3 = \alpha$ is the same as the problem of finding rational points on $y^2 = x^3 - 432\alpha^2$. And the general argument sketched above indicates that the same is true for any cubic. Of course, the normal form has an entirely different shape from the original equation. But there is a one-to-one correspondence between the rational points on one curve and the rational points on the other (up to a few easily catalogued exceptional points). So the problem of rational points on general cubic curves having one rational point is reduced to studying rational points on cubic curves in Weierstrass normal form.

The transformations we used to put the curve in normalized form do not map straight lines to straight lines. Since we defined the group law






A Cubic Curve with One Real Component

**Figure 1.11**

on our curve using lines connecting points, it is not at all clear that our transformation preserves the structure of the group. (That is, is our transformation a homomorphism?) It is, but that is not at all obvious. The point is that our description of addition of points on the curve is not a good one, because it seems to depend on the way the curve is embedded in the plane. But in fact the addition law is an intrinsic operation which can be described on the curve and is invariant under birational transformation. This follows from basic facts about algebraic curves, but is not so easy (virtually impossible?) to prove simply by manipulating the explicit equations.

A cubic equation in normal form looks like

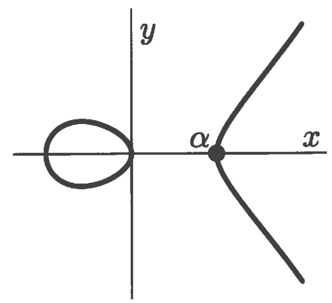
$$y^2 = f(x) = x^3 + ax^2 + bx + c.$$

Assuming that the (complex) roots of $f(x)$ are distinct, such a curve is called an *elliptic curve.* (More generally, any curve birationally equivalent to such a curve is called an elliptic curve.) Where does this name come from, because these curves are certainly not ellipses? The answer is that these curves arose in studying the problem of how to compute the arc length of an ellipse. If one writes down the integral which gives the arc length of an ellipse and makes an elementary substitution, the integrand will involve the square root of a cubic or quartic polynomial. So to compute the arc-length of an ellipse, one integrates a function involving $y = \sqrt{f(x)}$, and the answer is given in terms of certain functions on the "elliptic" curve $y^2 = f(x)$.

Now we take the coefficients $a, b, c$ of $f(x)$ to be rational, so in particular they are real; hence, the polynomial $f(x)$ of degree 3 has at least one real root. In real numbers we can factor it as

$$f(x) = (x-\alpha)(x^2 + \beta x + \gamma) \qquad \text{with } \alpha, \beta, \gamma \text{ real.}$$

Of course, it might have three real roots. If it has one real root, the curve looks something like Figure 1.11, because $y = 0$ when $x = \alpha$. If $f(x)$ has three real roots, then the curve looks like Figure 1.12. In this case the real points form two components.




A Cubic Curve with Two Real Components

**Figure 1.12**

All of this is valid, provided the roots of $f(x)$ are distinct. What is the significance of that condition? We have been assuming all along that our cubic curve is non-singular. If we write the equation as $F(x,y) = y^2 - f(x) = 0$ and take partial derivatives,

$$\frac{\partial F}{\partial x} = -f'(x), \qquad \frac{\partial F}{\partial y} = 2y,$$

then by definition the curve is non-singular, provided that there is no point on the curve at which the partial derivatives vanish simultaneously. This will mean that every point on the curve has a well-defined tangent line. Now if these partial derivatives were to vanish simultaneously at a point $(x_0, y_0)$ on the curve, then $y_0 = 0$, and hence $f(x_0) = 0$, and hence $f(x)$ and $f'(x)$ have the common root $x_0$. Thus $x_0$ is a double root of $f$. Conversely, if $f$ has a double root $x_0$, then $(x_0, 0)$ will be a singular point on the curve.

There are two possible pictures for the singularity. Which one occurs depends on whether $f$ has a double root or a triple root. In the case that $f$ has a double root, a typical equation is

$$y^2 = x^2(x+1),$$

and the curve has a singularity with distinct tangent directions as illustrated in Figure 1.13.

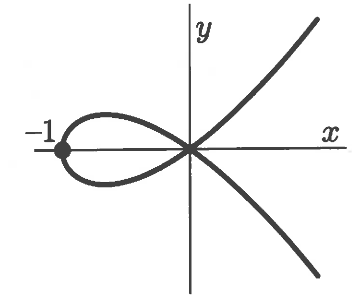
If $f(x)$ has a triple root, then after translating $x$ we obtain an equation

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

which is a semicubical parabola with a cusp at the origin. (See Figure 1.14.) These are examples of singular cubics in Weierstrass form, and the general case looks the same after a change of coordinates.

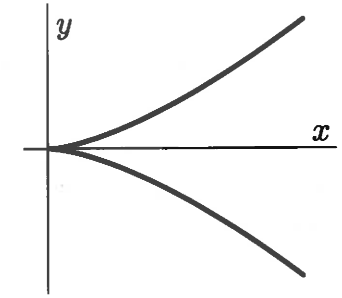
Why have we concentrated attention only on the non-singular cubics? It is not just to be fussy. The singular cubics and the non-singular cubics






A Singular Cubic with Distinct Tangent Directions

**Figure 1.13**




A Singular Cubic with A Cusp

**Figure 1.14**

have completely different types of behavior. For instance, the singular cubics are just as easy to treat as conics. If we project from the singular point onto some line, we see that the line going through that singular point meets the cubic twice at the singular point, so it meets the cubic only once more. The projection of the cubic curve onto the line is thus one-to-one. So just like a conic, the rational points on a singular cubic can be put in one-to-one correspondence with the rational points on the line. In fact, it is very easy to do that explicitly with formulas.

If we let $r = \dfrac{y}{x}$, then the equation $y^2 = x^2(x+1)$ becomes

$$r^2 = x + 1; \qquad \text{and so} \quad x = r^2 - 1 \quad \text{and} \quad y = r^3 - r.$$

If we take a rational number $r$ and define $x$ and $y$ in this way, then we obtain a rational point on the cubic; and if we start with a rational point $(x,y)$ on the cubic, then we get the rational number $r$. These operations are inverses of each other, and are defined at all rational points except for the singular point $(0,0)$ on the curve. So in this way we get all of the rational points on the curve.

The curve $y^2 = x^3$ is even simpler. We just take

$$x = t^2 \qquad \text{and} \qquad y = t^3.$$

So the singular cubics are trivial to analyze as far as rational points go, and Mordell's theorem does not hold for them. Actually we have not yet explained how to get a group law for these singular curves; but if one avoids the singularity, then one does get a group. We will see that this group is not finitely generated when we study it in more detail at the end of Chapter III.

## 4. Explicit Formulas for the Group Law

We are going to look at the group of points on a non-singular cubic a little more closely. If you are familiar with projective geometry, then you will not have any trouble; and if not, then you will have to accept a point at infinity, but only one. (If you have never studied any projective geometry, you might also want to look at the first two sections of Appendix A.)

We start with the equation

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

and make it homogeneous by setting $x = \dfrac{X}{Z}$ and $y = \dfrac{Y}{Z}$, yielding

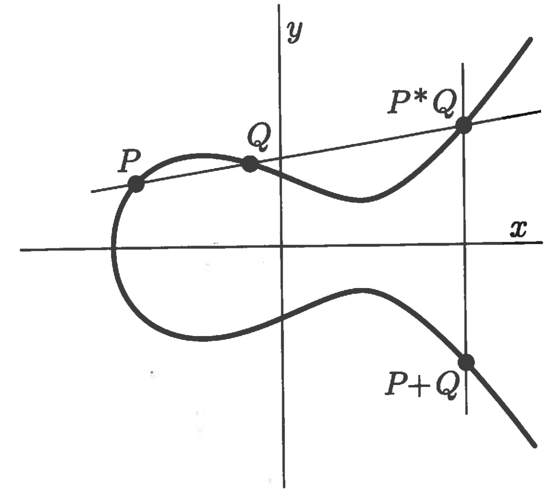
$$Y^2Z = X^3 + aX^2Z + bXZ^2 + cZ^3.$$

What is the intersection of this cubic with the line at infinity $Z = 0$? Substituting $Z = 0$ into the equation gives $X^3 = 0$, which has the triple root $X = 0$. This means that the cubic meets the line at infinity in three points, but the three points are all the same! So the cubic has exactly one point at infinity, namely, the point at infinity where vertical lines ($x =$ constant) meet. The point at infinity is an inflection point of the cubic, and the tangent line at that point is the line at infinity, which meets it there with multiplicity three. And one easily checks that the point at infinity is a non-singular point by looking at the partial derivatives there. So for a cubic in Weierstrass form there is one point at infinity; we will call that point $\mathcal{O}$.

The point $\mathcal{O}$ is counted as a rational point, and we take it as the zero element when we make the set of points into a group. So to make the game work, we have to make the convention that the points on our cubic consist of the ordinary points in the ordinary affine $xy$ plane together with one other point $\mathcal{O}$ that you cannot see. And now we find it is really true that every line meets the cubic in three points; namely, the line at infinity meets the cubic at the point $\mathcal{O}$ three times. A vertical line meets the cubic at two points in the $xy$ plane and also at the point $\mathcal{O}$. And a non-vertical line meets the cubic in three points in the $xy$ plane. Of course, we may have to allow $x$ and $y$ to be complex numbers.

Now we are going to discuss the group structure a little more closely. How do we add two points $P$ and $Q$ on a cubic equation in Weierstrass form? First we draw the line through $P$ and $Q$ and find the third intersection point $P*Q$. Then we draw the line through $P*Q$ and $\mathcal{O}$, which is just the vertical line through $P*Q$. A cubic curve in Weierstrass form is symmetric about the $x$ axis, so to find $P+Q$, we just take $P*Q$ and reflect it about the $x$ axis. This procedure is illustrated in Figure 1.15.






Adding Points on a Weierstrass Cubic

**Figure 1.15**

What is the negative of a point $Q$? The negative of $Q$ is the reflected point; if $Q = (x, y)$, then $-Q = (x, -y)$. (See Figure 1.16.) To check this, suppose that we add $Q$ to the point we claim is $-Q$. The line through $Q$ and $-Q$ is vertical, so the third point of intersection is $\mathcal{O}$. Now connect $\mathcal{O}$ to $\mathcal{O}$ and take the third intersection. Connecting $\mathcal{O}$ to $\mathcal{O}$ gives the line at infinity, and the third intersection is again $\mathcal{O}$ because the line at infinity meets the curve with multiplicity three at $\mathcal{O}$. This shows that $Q + (-Q) = \mathcal{O}$, so $-Q$ is the negative of $Q$. Of course, this formula does not apply to the case $Q = \mathcal{O}$, but obviously $-\mathcal{O} = \mathcal{O}$.

Now we develop some formulas to allow us to compute $P+Q$ efficiently. Let us change notation. We set

$$P_1 = (x_1, y_1), \quad P_2 = (x_2, y_2), \quad P_1 * P_2 = (x_3, y_3), \quad P_1 + P_2 = (x_3, -y_3).$$

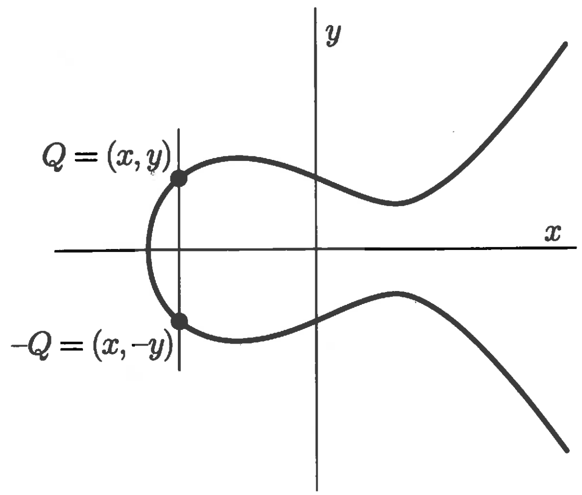
We assume that $(x_1, y_1)$ and $(x_2, y_2)$ are given, and we want to compute $(x_3, y_3)$.

First we look at the equation of the line joining $(x_1, y_1)$ and $(x_2, y_2)$. This line has the equation

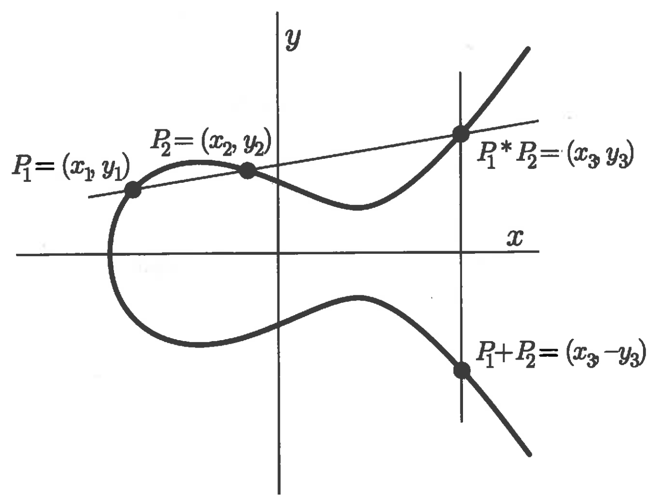
$$y = \lambda x + \nu, \qquad \text{where} \quad \lambda = \frac{y_2 - y_1}{x_2 - x_1} \quad \text{and} \quad \nu = y_1 - \lambda x_1 = y_2 - \lambda x_2.$$

By construction, the line intersects the cubic in the two points $(x_1, y_1)$ and $(x_2, y_2)$. How do we get the third point of intersection? We substitute

$$y^2 = (\lambda x + \nu)^2 = x^3 + ax^2 + bx + c.$$




The Negative of a Point on a Weierstrass Cubic
**Figure 1.16**




Deriving a Formula for the Addition Law
**Figure 1.17**

Putting everything on one side yields

$$0 = x^3 + (a - \lambda^2)x^2 + (b - 2\lambda\nu)x + (c - \nu^2).$$

This is a cubic equation in $x$, and its three roots $x_1, x_2, x_3$ give us the $x$



coordinates of the three intersections. Thus,

$$x^3 + (a - \lambda^2)x^2 + (b - 2\lambda\nu)x + (c - \nu^2) = (x - x_1)(x - x_2)(x - x_3).$$

Equating the coefficients of the $x^2$ term on either side, we find that $\lambda^2 - a = x_1 + x_2 + x_3$, and so

$$x_3 = \lambda^2 - a - x_1 - x_2, \qquad y_3 = \lambda x_3 + \nu.$$

These formulas are the most efficient way to compute the sum of two points.

Let us do an example. We look at the cubic curve

$$y^2 = x^3 + 17$$

which has the two points $P_1 = (-1, 4)$ and $P_2 = (2, 5)$. To compute $P_1 + P_2$ we first find the line through the two points, $y = \frac{1}{3}x + \frac{13}{3}$, so $\lambda = \frac{1}{3}$ and $\nu = \frac{13}{3}$. Next $x_3 = \lambda^2 - x_2 - x_1 = -\frac{8}{9}$ and $y_3 = \lambda x_3 + \nu = \frac{109}{27}$. Finally we find that $P_1 + P_2 = (x_3, -y_3) = \left(-\frac{8}{9}, -\frac{109}{27}\right)$. So computations really are not that bad.

The formulas we gave earlier involve the slope $\lambda$ of the line connecting the two points. What if the two points are the same? So suppose that we have $P_0 = (x_0, y_0)$ and we want to find $P_0 + P_0 = 2P_0$. We need to find the line joining $P_0$ to $P_0$. Because $x_1 = x_2$ and $y_1 = y_2$, we cannot use our formula for $\lambda$. But the recipe we described for adding a point to itself says that the line joining $P_0$ to $P_0$ is the tangent line to the cubic at $P_0$. From the relation $y^2 = f(x)$ we find by implicit differentiation that

$$\lambda = \frac{dy}{dx} = \frac{f'(x)}{2y},$$

so that is what we use when we want to double a point.

Continuing with our example curve $y^2 = x^3 + 17$ and point $P_1 = (-1, 4)$, we compute $2P_1$ as follows. First, $\lambda = f'(x_1)/2y_1 = f'(-1)/8 = \frac{3}{8}$. Then, once we have a value for $\lambda$, we just substitute into the formulas as above, eventually finding that $2P_1 = \left(\frac{137}{64}, -\frac{2651}{512}\right)$.

Sometimes it is convenient to have an explicit expression for $2P$ in terms of the coordinates for $P$. If we substitute $\lambda = f'(x)/2y$ into the formulas given earlier, put everything over a common denominator, and replace $y^2$ by $f(x)$, then we find that

$$x \text{ coordinate of } 2(x, y) = \frac{x^4 - 2bx^2 - 8cx + b^2 - 4ac}{4x^3 + 4ax^2 + 4bx + 4c}.$$

This formula for $x(2P)$ is often called the *duplication formula*. It will come in very handy later for both theoretical and computational purposes. We will leave it to you verify this formula, as well as to derive a companion formula for the $y$ coordinate of $2P$.

These are the basic formulas for the addition of points on a cubic when the cubic is in Weierstrass form. We will use the formulas extensively to prove many facts about rational points on cubic curves, including Mordell's theorem. Further, if you were not satisfied with our incomplete proof that the addition law is associative, you can just take three points at random and compute. Of course, there are an awful lot of special cases to consider, such as when one of the points is the negative of the other or two of the points coincide. But in a few days you will be able to check associativity using these formulas. So we need say nothing more about the proof of the associative law!

## EXERCISES

1.1. (a) If $P$ and $Q$ are distinct rational points in the $(x, y)$ plane, prove that the line connecting them is a rational line.
(b) If $L_1$ and $L_2$ are distinct rational lines in the $(x, y)$ plane, prove that their intersection is a rational point (or empty).

1.2. Let $C$ be the conic given by the equation

$$F(x,y) = ax^2 + bxy + cy^2 + dx + ey + f = 0,$$

and let $\delta$ be the determinant

$$\det\begin{pmatrix} 2a & b & d \\ b & 2c & e \\ d & e & 2f \end{pmatrix}.$$

(a) Show that if $\delta \neq 0$, then $C$ has no singular points. That is, show that there are no points $(x, y)$ where

$$F(x,y) = \frac{\partial F}{\partial x}(x,y) = \frac{\partial F}{\partial y}(x,y) = 0.$$

(b) Conversely, show that if $\delta = 0$ and $b^2 - 4ac \neq 0$, then there is a unique singular point on $C$.
(c) Let $L$ be the line $y = \alpha x + \beta$ with $\alpha \neq 0$. Show that the intersection of $L$ and $C$ consists of either zero, one, or two points.
(d) Determine the conditions on the coefficients which ensure that the intersection $L \cap C$ consists of exactly one point. What is the geometric significance of these conditions? (Note. There will be more than one case.)

1.3. Let $C$ be the conic given by the equation

$$x^2 - 3xy + 2y^2 - x + 1 = 0.$$

(a) Check that $C$ is non-singular (cf. the previous exercise).
(b) Let $L$ be the line $y = \alpha x + \beta$. Suppose that the intersection $L \cap C$ contains the point $P_0 = (x_0, y_0)$. Assuming that $L \cap C$ consists of two distinct points, find the second point of $L \cap C$ in terms of $\alpha, \beta, x_0, y_0$.
(c) If $L$ is a rational line and $P_0$ is a rational point (i.e., $\alpha, \beta, x_0, y_0 \in \mathbb{Q}$), prove that the second point of $L \cap C$ is also a rational point.



1.4. Find all primitive integral right triangles whose hypotenuse has length less than 30.

1.5. Find all of the rational points on the circle

$$x^2 + y^2 = 2$$

by projecting from the point $(1, 1)$ onto an appropriate rational line. (Your formulas will be simpler if you are clever in your choice of the line.)

1.6. (a) Let $a, b, c, d, e, f$ be non-zero real numbers. Use the substitution $t = \tan(\theta/2)$ to transform the integral

$$\int \frac{a + b\cos\theta + c\sin\theta}{d + e\cos\theta + f\sin\theta}\,d\theta$$

into the integral of a rational function of $t$.
(b) Evaluate the integral

$$\int \frac{a + b\cos\theta + c\sin\theta}{1 + \cos\theta + \sin\theta}\,d\theta.$$

1.7. For each of the following conics, either find a rational point or prove that there are no rational points.
(a) $x^2 + y^2 = 6$
(b) $3x^2 + 5y^2 = 4$
(c) $3x^2 + 6y^2 = 4$

1.8. Prove that for every exponent $e \geq 1$, the congruence

$$x^2 + 1 \equiv 0 \pmod{5^e}$$

has a solution $x_e \in \mathbb{Z}/5^e\mathbb{Z}$. Prove further that these solutions can be chosen to satisfy

$$x_1 \equiv 2 \pmod 5, \qquad \text{and} \quad x_{e+1} \equiv x_e \pmod{5^e} \quad \text{for all } e \geq 1.$$

(This is equivalent to showing that the equation $x^2 + 1 = 0$ has a solution in the 5-adic numbers. It is a special case of Hensel's lemma. *Hint.* Use induction on $e$.)

1.9. Let $C_1$ and $C_2$ be the cubics given by the following equations:

$$C_1 : x^3 + 2y^3 - x - 2y = 0, \qquad C_2 : 2x^3 - y^3 - 2x + y = 0.$$

(a) Find the nine points of intersection of $C_1$ and $C_2$.
(b) Let $\{(0,0), P_1, \ldots, P_8\}$ be the nine points from part (a). Prove that if a cubic curve goes through $P_1, \ldots, P_8$, then it must also go through the ninth point $(0, 0)$.

1.10. Define a composition law on the points of a cubic $C$ by the following rule as described in the text: Given $P, Q \in C$, then $P * Q$ is the point on $C$ so that $P$, $Q$, and $P * Q$ are colinear.
(a) Explain why this law is commutative, $P * Q = Q * P$.
(b) Prove that there is no identity element for this composition law; that is, there is no element $P_0 \in C$ such that $P_0 * P = P$ for all $P \in C$.
(c) Prove that this composition law is not associative; that is, in general, $P * (Q * R) \neq (P * Q) * R$.
(d) Explain why $P * (P * Q) = Q$.
(e) Suppose that the line through $\mathcal{O}$ and $S$ is tangent to $C$ at $\mathcal{O}$. Explain why
$$\mathcal{O} * (Q * (Q * S)) = \mathcal{O}.$$
(This is an algebraic verification that the point we called $-Q$ is the additive inverse of $Q$.)

1.11. Let $S$ be a set with a composition law $*$ satisfying the following two properties:

(i) $P * Q = Q * P$ for all $P, Q \in S$.

(ii) $P * (P * Q) = Q$ for all $P, Q \in S$.

Fix an element $\mathcal{O} \in S$, and define a new composition law $+$ by the rule
$$P + Q = \mathcal{O} * (P * Q).$$
(a) Prove that $+$ is commutative and has $\mathcal{O}$ as identity element (i.e., prove that $P + Q = Q + P$ and $P + \mathcal{O} = P$).
(b) Prove that for any given $P, Q \in S$, the equation $X + P = Q$ has the unique solution $X = P * (Q * \mathcal{O})$ in $S$. In particular, if we define $-P$ to be $P * (\mathcal{O} * \mathcal{O})$, then $-P$ is the unique solution in $S$ of the equation $X + P = \mathcal{O}$.
(c) Prove that $+$ is associative (and so $(S, +)$ is a group) if and only if

(iii) $R * (\mathcal{O} * (P * Q)) = P * (\mathcal{O} * (Q * R))$ for all $P, Q, R \in S$.

(d) Let $\mathcal{O}' \in S$ be another point, and define a composition law $+'$ by $P +' Q = \mathcal{O}' * (P * Q)$. Suppose that both $+$ and $+'$ are associative, so we obtain two group structures $(S, +)$ and $(S, +')$ on $S$. Prove that the map
$$P \longmapsto \mathcal{O} * (\mathcal{O}' * P)$$
is a (group) isomorphism from $(S, +)$ to $(S, +')$.
(e) * Find a set $S$ with a composition law $*$ satisfying (i) and (ii) such that $(S, +)$ is not a group.



1.12. The cubic curve $u^3 + v^3 = \alpha$ (with $\alpha \neq 0$) has a rational point $[1, -1, 0]$ at infinity. (That is, this is a point on the homogenized equation $U^3 + V^3 = \alpha W^3$.) Taking this rational point to be $\mathcal{O}$, we can make the points on the curve into a group.
(a) Derive a formula for the sum $P_1 + P_2$ of two points $P_1 = (u_1, v_1)$ and $P_2 = (u_2, v_2)$.
(b) Derive a duplication formula for $2P$ in terms of $P = (u, v)$.

1.13. (a) Verify that if $u$ and $v$ satisfy the relation $u^3 + v^3 = \alpha$, then the quantities
$$x = \frac{12\alpha}{u + v} \quad \text{and} \quad y = 36\alpha \frac{u - v}{u + v}$$
satisfy the relation $y^2 = x^3 - 432\alpha^2$.
(b) Part (a) gives a birational transformation from the curve $u^3 + v^3 = \alpha$ to the curve $y^2 = x^3 - 432\alpha^2$. Each of these cubic curves has a group law defined on it. Prove that the birational transformation described in (a) is an isomorphism of groups.

1.14. Let $C$ be the cubic curve $u^3 + v^3 = u + v + 1$. In the projective plane, this curve has a point $[1, -1, 0]$ at infinity. Find rational functions $x = x(u, v)$ and $y = y(u, v)$ so that $x$ and $y$ satisfy a cubic equation in Weierstrass normal form with the given point still at infinity.

1.15. Let $g(t)$ be a quartic polynomial with distinct (complex) roots, and let $\alpha$ be a root of $g(t)$. Let $\beta \neq 0$ be any number.
(a) Prove that the equations
$$x = \frac{\beta}{t - \alpha}, \qquad y = x^2 u = \frac{\beta^2 u}{(t - \alpha)^2}$$
give a birational transformation between the curve $u^2 = g(t)$ and the curve $y^2 = f(x)$, where $f(x)$ is the cubic polynomial
$$f(x) = g'(\alpha)\beta x^3 + \tfrac{1}{2} g''(\alpha)\beta^2 x^2 + \tfrac{1}{6} g'''(\alpha)\beta^3 x + \tfrac{1}{24} g''''(\alpha)\beta^4.$$
(b) Prove that if $g$ has distinct (complex) roots, then $f$ also has distinct roots, and so $u^2 = g(t)$ is an elliptic curve.

1.16. Let $0 < \beta \leq \alpha$, and let $E$ be the ellipse
$$\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} = 1.$$
(a) Prove that the arc length of $E$ is given by the integral
$$4\alpha \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \theta}\, d\theta.$$

for an appropriate choice of the constant $k$ depending on $\alpha$ and $\beta$.
(b) Check your value for $k$ in (a) by verifying that when $\alpha = \beta$, the integral yields the correct value for the arc length of a circle.
(c) Prove that the integral in (a) is also equal to

$$4\alpha \int_0^1 \sqrt{\frac{1-k^2t^2}{1-t^2}}\,dt = 4\alpha \int_0^1 \frac{1-k^2t^2}{\sqrt{(1-t^2)(1-k^2t^2)}}\,dt.$$

(d) Prove that if the ellipse $E$ is not a circle, then the equation

$$u^2 = (1-t^2)(1-k^2t^2)$$

defines an elliptic curve (cf. the previous exercise). Hence the problem of determining the arc length of an ellipse comes down to evaluating the integral

$$\int_0^1 \frac{1-k^2t^2}{u}\,dt \qquad \text{on the "elliptic" curve } u^2 = (1-t^2)(1-k^2t^2).$$

1.17. Let $C$ be a cubic curve in the projective plane given by the homogeneous equation

$$Y^2Z = X^3 + aX^2Z + bXZ^2 + cZ^3.$$

Verify that the point $[0,1,0]$ at infinity is a non-singular point of $C$.

1.18. The cubic curve

$$y^2 = x^3 + 17$$

has the following five rational points:

$$P_1 = (-2,3),\ P_2 = (-1,4),\ P_3 = (2,5),\ P_4 = (4,9),\ P_5 = (8,23).$$

(a) Show that $P_2$, $P_4$, and $P_5$ can each be expressed as $mP_1 + nP_3$ for an appropriate choice of integers $m$ and $n$.
(b) Compute the points

$$P_6 = -P_1 + 2P_3 \qquad \text{and} \qquad P_7 = 3P_1 - P_3.$$

(c) Notice that the points $P_1, P_2, P_3, P_4, P_5, P_6, P_7$ all have integer coordinates. There is exactly one more rational point on this curve which has integer coordinates and $y > 0$. Find that point. (You will probably need at least a programmable calculator or else a lot of patience.)
(d) ** Prove the assertion in (c) that there are exactly eight rational points $(x,y)$ on this curve with $y > 0$ and $x, y$ both integers. (This is an extremely difficult problem, and you will almost certainly not be able to do it with the tools we have developed. But it is also an extremely interesting problem which is well worth thinking about.)



1.19. Suppose that $P = (x,y)$ is a point on the cubic curve

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

(a) Verify that the $x$ coordinate of the point $2P$ is given by the *duplication formula*

$$x(2P) = \frac{x^4 - 2bx^2 - 8cx - 4ac + b^2}{4y^2}.$$

(b) Derive a similar formula for the $y$ coordinate of $2P$ in terms of $x$ and $y$.
(c) Find a polynomial in $x$ whose roots are the $x$ coordinates of the points $P = (x,y)$ satisfying $3P = \mathcal{O}$. (*Hint.* The relation $3P = \mathcal{O}$ can also be written $2P = -P$.)
(d) For the particular curve $y^2 = x^3 + 1$, solve the equation in (c) to find all of the points satisfying $3P = \mathcal{O}$. Note you will have to use complex numbers.

1.20. Consider the point $P = (3,8)$ on the cubic curve $y^2 = x^3 - 43x + 166$. Compute $P$, $2P$, $3P$, $4P$, and $8P$. Comparing $8P$ with $P$, what can you conclude?