G1CRPC: Rational Points on Curves
Course Notes 2003

Contents

1 Introduction

Much of mathematics concerns the solution of various kinds of equation: determining the existence of solutions, studying the set of solutions and its structure, and (from a practical point of view) determining ways of finding one solution or the general solution. The same is true in number theory, a substantial part of which concerns *Diophantine Equations*: polynomial equations (in one or more variables) for which solutions are sought which are integers, or rational numbers.

For example, the *Fermat equation*

$$a^n + b^n = c^n \qquad (a, b, c \in \mathbb{Z}) \tag{1}$$

is a Diophantine equation. Are there any solutions with a, b, c non-zero when $n \geq 3$? As a second example, fix a non-zero integer c and consider the *Bachet-Mordell equation*

$$\mathcal{B}: \qquad Y^2 - X^3 = c; \tag{2}$$

rational solutions (X, Y) correspond to ways of expressing c as the difference between a square and a cube.

We may view such equations geometrically: both (1) and (2) determine the equations of curves in the (X, Y) plane. In the case of (1) we write $X = a/c$, $Y = b/c$ and seek rational solutions to the equation

$$\mathcal{F}_n: \qquad X^n + Y^n = 1. \tag{3}$$

These curves (more precisely, the set of real solutions to these equations) may be plotted in the usual way in the real X, Y plane $\mathbb{R}^2$. Such pictures give us a good picture of what the *real* solutions look like, but are they much help in determining *rational* solutions? We call a point $(X, Y) \in \mathbb{R}^2$ a *rational point* if both of its coordinates are rational: $X, Y \in \mathbb{Q}$. What does a picture of just the rational points on these curves look like?

This is a much more subtle question. For example, the ordinary unit circle $\mathcal{C} = \mathcal{F}_2$ has infinitely many rational points, which are dense on the curve. Explicitly, for any rational value of the parameter t, if we set

$$X = \frac{1 - t^2}{1 + t^2}, \qquad Y = \frac{2t}{1 + t^2} \tag{4}$$

then (X, Y) is a rational point on $\mathcal{C}$, and every rational point $(X, Y) \in \mathcal{C}$ may be obtained from $t = Y/(1 + X)$, except for the point $(-1, 0)$ (which comes from letting $t \to \infty$). We will prove this later. *But* for any integer $n > 2$ the curve $\mathcal{F}_n$ has only two or four rational points, namely $\{(\pm 1, 0), (0, \pm 1)\}$ when n is even and $\{(1, 0), (0, 1)\}$ when n is odd. So the graph of $\mathcal{F}_n$ in the real plane $\mathbb{R}^2$ manages to avoid (almost) all the rational points in the plane, even though the rational points in the plane form a dense subset. Of course, proving this statement is very hard for general n: it is Fermat's Last Theorem, which was only proved in 1994, in a very indirect manner. (Many special cases were known before this; the case $n = 4$ was proved by Fermat himself 350 years ago.)

So we have seen that very different results are obtained when we seek solutions to the same equation over a different base field. There are other base fields we may wish to consider, such as the complex numbers $\mathbb{C}$ or a finite field $\mathbb{F}_q$. The set of solutions to (3) with $X, Y \in F$ for some field F will be denoted $\mathcal{F}_n(F)$. For any F the formulas (4) give all elements of $\mathcal{F}_2(F)$ apart from $(-1, 0)$, as t runs over the elements of F, provided that we avoid values of t satisfying $t^2 + 1 = 0$. (Thus we have to avoid $t = \pm i$ for $F = \mathbb{C}$, and also two values when $F = \mathbb{F}_p$ and $p \equiv 1 \pmod 4$.) Some algebraic considerations become simpler when working over the complex numbers. Over $\mathbb{C}$, our "curves" have complex dimension 1, so are in fact surfaces (with real dimension 2), though this is a little hard to visualise since the surfaces are embedded in $\mathbb{C}^2 \sim \mathbb{R}^4$. It turns out that the "circle" $\mathcal{F}_2(\mathbb{C})$ is a sphere, while in general $\mathcal{F}_n(\mathbb{C})$ is a g-holed torus where $g = (n - 1)(n - 2)/2$; the number g here is called the *genus*, and the nature of the rational points on a curve $\mathcal{C}$ depends critically on the value of its genus.

Over a finite field such as $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, any curve is just a finite set of points, since the coordinates X and Y can only take on a finite set of values: the whole plane only contains p^2 points! Geometric intuition fails us here, but the algebraic techniques which we will use work just as well in this context. For example, using the parametrisation (4) it is not hard to show that the number of points on $\mathcal{F}_2(\mathbb{F}_p)$ is

$$\#\mathcal{F}_2(\mathbb{F}_p) = \begin{cases} p+1 & \text{if } p \equiv 3 \pmod 4 \\ p-1 & \text{if } p \equiv 1 \pmod 4 \\ 2 & \text{if } p = 2. \end{cases} \tag{5}$$

Later we will study curves in the so-called *projective plane* rather than the ordinary "affine" X, Y plane; then the above formula becomes much simpler, since in the projective plane over $\mathbb{F}_p$, the curve $\mathcal{F}_2$ has exactly $p+1$ points for *every* prime p. The projective plane contains more points than the affine plane, the "points at infinity", and each curve has a finite number of extra points at infinity which exactly make up the numbers.

Studying curves over finite fields is big business these days, as they have important applications in cryptography and to error-correcting codes. So knowledge of these matters is a marketable skill, as well as being a beautiful mathematical theory in its own right. At the end of the module we will look briefly at two applications.

We have seen that studying rational solutions to polynomial equations involves an interplay between number theory and geometry; but what about algebra? At one level, algebra is certainly involved, since the equations we study are polynomial equations, and the study of polynomials is a big part of algebra. In this module we will mainly be studying special classes of curves, so we do not need more than the basic facts from the theory of polynomials; in a more advanced course on algebraic curves, surfaces or their analogues in higher dimensions (algebraic varieties), it is vital to have a substantial background in the algebra of polynomials in several variables. Several textbooks on this "commutative algebra" are written with this specifically in mind.

There is a more subtle way in which algebra will enter our study of points on curves. Consider a cubic curve (i.e., given by an equation of degree 3), such as the Bachet curve $\mathcal{B}$ defined by (2). If we have two points $P_1 = (X_1, Y_1)$, $P_2 = (Y_2, X_2)$ on $\mathcal{B}$, and join them with a straight line $\mathcal{L}$, this line will intersect the curve $\mathcal{B}$ in a third point $P_3 = (X_3, Y_3)$. (Later we will work out the explicit formulas for this.) So we have a way of combining two points on the curve together to get a third. Moreover, if both P_1 and P_2 are both rational points (and the number c in the definition of $\mathcal{B}$ is also rational), then P_3 will also be rational. Even if we only know one rational point P on $\mathcal{B}$ we still get a new point by intersecting the tangent line at P with $\mathcal{B}$, and this point will also be rational when P is. This "tangent-chord" process for obtaining new solutions from old was first carried out (for this equation) by Bachet in 1621–which is rather remarkable, since that is before Descartes introduced the idea of "Cartesian coordinates", so it is hard to see how Bachet could have had the same geometric intuition which we bring to the problem. Explicitly, Bachet found that if $P = (x, y)$ satisfies (2), and $y \neq 0$, then so does P', defined by

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

For example, when $c = -2$ the equation $Y^2 - X^3 = -2$ has the obvious rational point $P = (3, 5)$. Using Bachet's duplication formula, we can derive from this the following *infinite* sequence of solutions (or of rational points on the curve):

$$(3, 5), \left(\frac{129}{100}, \frac{-383}{1000} \right), \left(\frac{2340922881}{(7660)^2}, \frac{113259286337292}{(7660)^3} \right), \dots.$$

Assuming that the sequence never repeats (which it does not), this gives us infinitely many rational points on the curve. The number of digits needed to write down the X coordinates of these points approximately quadruples at each stage, so the solution rapidly become very large (in the sense of having many digits).

In fact, it is a theorem (due to Mordell) that this procedure will always give an infinite sequence of distinct rational solutions to Bachet's equation, for any non-zero integer value of c *except* for $c = 1$ and $c = -432$.

We will see that this operation, whereby we construct new rational points from old ones, turns the set of rational points on a cubic curve into a *group*. These cubic curves, with the group structure on their points, are called *elliptic curves*, and will form the subject of the second half of the module. What sort of abelian groups do we get this way? The main result, due to Mordell again, is that the group is always finitely-generated. In concrete terms this means that there always exists a finite set of rational points (or rational solutions), called generators, such that every other rational point may be obtained from these by a finite sequence of tangent-chord constructions. Mordell proved this in 1923; it had been conjectured by Poincaré in 1901. Even now, open questions remain: how many generators are required? It is not known whether there exist elliptic curves requiring arbitrarily many generators, though this is widely believed. The current record number is 24, for a curve found by some researchers at NSA. Another open problem is to find an algorithm which will determine the generators for any specific given curve. While this is possible for many elliptic curves, there are some for which existing techniques break down; together with related questions, this is under active study by the number theory research group in Nottingham.

To close the Introduction, we will show that the Fermat equation (3) and the Bachet equation do not have a parametric solution of the form (4), in which X and Y are given in terms of rational functions of a single parameter t. For the Bachet curve, and other elliptic curves, there do exist parametrizations using power series rather than polynomials: this is one form of the famous result of Wiles, which led (indirectly) to the proof of Fermat's Last Theorem.

Theorem 1.1. *Let $n \geq 3$ be an integer. There do not exist non-constant polynomials $F(t)$, $G(t)$, $H(t)$ satisfying*

$$F(t)^n + G(t)^n = H(t)^n.$$

Thus the Fermat equation (3) has no solutions with X, Y non-constant rational functions of t.

This contrasts with the case $n = 2$, where we have the identity

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

giving (4). No such identity holds for nth powers when $n \geq 3$. This proves Fermat's Last Theorem for polynomials. The method of proof (given in lectures) cannot be used to prove Fermat's Last Theorem for integers, since there is no operation corresponding to differentiation with respect to t.

The same technique may be used to prove the following generalisation:

Theorem 1.2. *Let n_1, n_2, n_3 be positive integers such that $1/n_1 + 1/n_2 + 1/n_3 \leq 1$, and let c_1, c_2, c_3 be non-zero constants. Then there are no non-constant coprime polynomials $F(t)$, $G(t)$, $H(t)$ satisfying*

$$c_1 F(t)^{n_1} + c_2 G(t)^{n_2} = c_3 H(t)^{n_3}.$$

From this it is not hard to show that the generalised Fermat equation

$$c_1 X^{n_1} + c_2 Y^{n_2} = c_3$$

has no parametric solution for $n_1, n_2 > 1$ except when $n_1 = n_2 = 2$. Taking $n_1 = 2$ and $n_2 = 3$ shows that the Bachet-Mordell equation (2) has no parametric solution.

2 Affine Curves

Basic Notation

$\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ will denote, as usual, the sets of integers, rational numbers, real numbers and complex numbers. The integers form a ring, the others sets are fields. The finite field with q elements is denoted $\mathbb{F}_q$. The simplest of these are the fields $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ for p a prime number.

We will use K and L to denote arbitrary fields. K^* denotes the set of non-zero elements of K (which form a group under multiplication). $K[X]$ and $K[X,Y]$ denote the rings of polynomials in one variable X and in two variables X, Y respectively (and similarly for more variables).

2.1 The affine plane

Definition 1. The *affine plane* over a field K is the set K^2 of all ordered pairs (x, y) with $x, y \in K$. Notation $\mathbb{A}^2(K)$, or just $\mathbb{A}^2$.

A *point* is an element $P = (x, y) \in \mathbb{A}^2(K)$ for some field K. Its *coordinates* are $x = X(P)$ and $y = Y(P)$. We say that P is *K-rational* if its coordinates are in K.

A *rational point* is an element of $\mathbb{A}^2(\mathbb{Q})$: a point with rational coordinates. Similarly we may talk of a *real point* in $\mathbb{A}^2(\mathbb{R})$, a *complex point* in $\mathbb{A}^2(\mathbb{C})$, etc.

Our pictures will generally be of the real affine plane $\mathbb{A}^2(\mathbb{R})$, over the real field $\mathbb{R}$. The rational points $\mathbb{A}^2(\mathbb{Q})$ form a subset of $\mathbb{A}^2(\mathbb{R})$, which in turn is a subset of $\mathbb{A}^2(\mathbb{C})$.

2.2 Affine plane curves

Roughly speaking, we wish to define a plane curve to be a subset of the affine plane consisting of the points which satisfy a single polynomial equation $f(X, Y) = 0$. However, there may be many quite different polynomials with the same set of zeros: for example $f(X, Y) = X + Y$ has the same zeros as $(X + Y)^5$ over any field, and over $\mathbb{R}$ all polynomials of the form $X^2 + Y^2 + c$ with $c > 0$ have no zeros at all. The following definition is designed to get around this: essentially, we define a plane curve to be a polynomial, up to multiplication by a non-zero constant factor.

Definition 2. An *affine plane curve defined over the field K* is an equivalence class of non-constant polynomials $f(X, Y) \in K[X, Y]$, where two polynomials f and g are said to be equivalent if $g = cf$ for some $c \in K^*$. The curve determined by f will be denoted $\mathcal{C}_f$. The *degree* of $\mathcal{C}_f$ is $\deg(\mathcal{C}_f) = \deg(f)$.

It should be obvious that being equal up to a non-zero constant factor really is an equivalence relation, and that equivalent polynomials have the same the degree.

There is a slight subtlety in this definition: a curve may be defined by a polynomial with irrational coefficients but still be rational: for example, the curve (in fact a line) defined by $\sqrt{2}X + \sqrt{2}Y = 0$ is rational since $f(X, Y) = \sqrt{2}X + \sqrt{2}Y$ is equivalent to $g = X + Y$. It is the ratios between the coefficients which are important here. Similarly, $iX^2 - iY^2$ defines a real curve (defined over $\mathbb{R}$) since it is equivalent to $X^2 - Y^2$.

Often we will use a slightly simpler language, and talk of "the curve f" or "the curve $f = 0$" for a polynomial f, or define a curve as "the curve $\mathcal{C}$ defined by (the equation) $f = 0$".

Another reason for defining a curve to be a polynomial (up to constant factor) is that it allows us to consider, for a given polynomial, its zeros in more than one field: for a curve defined over $\mathbb{Q}$ (which we will call a *rational curve* for short), we can look at its rational points, but also at its real or complex points.

Definition 3. Let $\mathcal{C} = \mathcal{C}_f$ be an affine plane curve defined by a polynomial $f(X, Y) \in K[X, Y]$, and L a field with $K \subseteq L$. The *set of L-rational points* of $\mathcal{C}$ is

$$\mathcal{C}(L) = \{(x, y) \in \mathbb{A}^2(L) \mid f(x, y) = 0\}.$$

Of course, $L = K$ is allowed, and $L \subseteq L' \implies \mathcal{C}(L) \subseteq \mathcal{C}(L')$. Evaluating $f(x, y)$ for $x, y \in L$ makes sense since $K \subseteq L$ and the coefficients of f lie in K. The condition that $f(x, y) = 0$ is unchanged if we rescale f by a constant factor.

Examples: (1) Let $f = X^2 + Y^2 - 1$. Then $\mathcal{C}_f(\mathbb{R})$ is the usual unit circle in $\mathbb{R}^2 = \mathbb{A}^2(\mathbb{R})$, while $\mathcal{C}_f(\mathbb{Q})$ consists of the infinite set of rational points on the circle (see the Introduction). For example, $(3/5, 4/5) \in \mathcal{C}_f(\mathbb{Q})$.

(2) Let $g = X^2 + Y^2 - 3$. Again, $\mathcal{C}_g(\mathbb{R})$ is a circle in $\mathbb{A}^2(\mathbb{R})$, but now $\mathcal{C}_g(\mathbb{Q}) = \emptyset$ (exercise; see G13NUM notes).

(3) Let $h = X^2 + Y^2 + 1$. Now $\mathcal{C}_h(\mathbb{Q}) = \mathcal{C}_h(\mathbb{R}) = \emptyset$; while $\mathcal{C}_h(\mathbb{C})$ is infinite (containing $(i, 0)$, for example).

Definition 4. A plane curve $\mathcal{C}_f$ defined by $f \in K[X, Y]$ is *irreducible* if the polynomial $f(X, Y)$ is irreducible, not only in $K[X, Y]$ but also in $L[X, Y]$ for every algebraic extension L of K.

For curves defined over $\mathbb{Q}$, or any subfield of $\mathbb{C}$, it may be easier to use the condition that $\mathcal{C}_f$ is irreducible if and only if $f(X, Y)$ is irreducible in $\mathbb{C}[X, Y]$. The point of making this definition more complicated than just saying that $\mathcal{C}_f$ is irreducible if $f(X, Y)$ is irreducible in $K[X, Y]$ is this: we want the condition of irreducibility to be a geometric one, which is independent of the field of definition. For example, the rational curve $\mathcal{C}$ defined by $f = X^2 + Y^2 \in \mathbb{Q}[X, Y]$ is not irreducible, since over $\mathbb{C}$ (or even over $\mathbb{Q}(i)$) we can factorize $f = (X - iY)(X + iY)$.

If a curve $\mathcal{C}_f$ defined over K is reducible, then there is a non-trivial factorization $f = gh$, where the coefficients of g, h may lie in a larger field L than K. (Non-trivial means that neither factor is constant.) In this case, for every field $L' \supseteq L$ we have

$$\mathcal{C}_f(L') = \mathcal{C}_g(L') \cup \mathcal{C}_h(L'),$$

so the curve $\mathcal{C}_f$ splits into two pieces, $\mathcal{C}_g$ and $\mathcal{C}_h$, determined by the factors of f. Since polynomials (in any number of variables over any field) form a Unique Factorization Domain (UFD), every affine curve is uniquely expressible as a union of irreducible curves, though these irreducible components may be defined over fields strictly larger than the original.

Suppose first that K is algebraically closed (e.g. $K = \mathbb{C}$). Since $K[X, Y]$ is a UFD, every non-constant $f \in K[X, Y]$ has a unique factorization

$$f = \prod_{i=1}^{n} f_i^{m_i}$$

where the factors f_i are irreducible, and each $m_i \geq 1$. Then $\mathcal{C}_f(K)$ is the union of the $\mathcal{C}_{f_i}(K)$, and each $\mathcal{C}_{f_i}$ is an irreducible curve. So $\mathcal{C}_f$ may be expressed uniquely as a union of irreducible curves, called its *irreducible components.*

If K is not algebraically closed (e.g. $K = \mathbb{Q}$ or $K = \mathbb{R}$) then the above factorization should be done over an algebraically closed field L containing K (such as $L = \mathbb{C}$ when $K = \mathbb{Q}$ or $K = \mathbb{R}$). The irreducible components of f may then not be defined over the same field K as f, but over a larger field. For example, the irreducible components of the rational curve $X^2 + Y^2$ are the two complex lines $Y = \pm iX$.

Usually we will restrict to irreducible curves.

Warning: There are two quite different meanings to the term "rational curve". We have defined a rational curve to be one which is defined over the field $\mathbb{Q}$ of *rational numbers*, i.e. a curve which is defined by a polynomial with rational coefficients (possibly up to a scaling factor). A second meaning, which we will not be using in this module, but which you may see in books, is to say that (over any field) a curve is rational if it may be parametrised by *rational functions.* In our terminology, all the Fermat curves $\mathcal{F}_n$ are rational, but with the alternative definition $\mathcal{F}_n$ is not rational for $n \geq 3$ by Theorem 1.1.

2.3 Affine lines and conics

Definition 5. An *affine line* is an affine curve of degree 1.

So a line is a curve defined by an equation of the form

$$\mathcal{L}: \quad aX + bY + c = 0, \tag{7}$$

where $(a, b) \neq (0, 0)$. If $b \neq 0$ we can write this equation in the familiar "slope-intercept" form $Y = mX + d$ where $m = -a/b$ and $d = -c/b$; the only lines not of this form are the "vertical" lines $X = e$ (where $b = 0$ and $e = -c/a$). We say that a vertical line has slope ∞.

The line $\mathcal{L}$ is rational if $a, b, c \in \mathbb{Q}$, or more generally if $ka, kb, kc \in \mathbb{Q}$ for some non-zero k. This is equivalent to the conditions that $a/b, c/b \in \mathbb{Q}$ (if $b \neq 0$) or that $b/a, c/a \in \mathbb{Q}$ (if $a \neq 0$). So a rational line has rational slope (or slope ∞, which we also regard as rational in this context).

Here are some other conditions for a line to be rational:

Proposition 2.1. *The line $\mathcal{L}$ defined by* (7) *is rational if it contains at least two rational points, or if it contains at least one rational point and has rational slope.*

The essential properties of points and lines in the affine plane are as follows.

Proposition 2.2. *For every pair of distinct points $P, Q \in \mathbb{A}^2$, there is a unique line $\mathcal{L} = \mathcal{L}_{P,Q}$ passing through both P and Q; this line is rational if both points are rational.*

Proposition 2.3. *For every pair of distinct lines $\mathcal{L}_1$, $\mathcal{L}_2$ in $\mathbb{A}^2$* which do not have the same slope, *there is a unique point $P = \mathcal{L}_1 \cap \mathcal{L}_2$ lying on both $\mathcal{L}_1$ and $\mathcal{L}_2$; this point is rational if both lines are rational.*

Note the almost symmetry between the previous two results: when we consider the projective plane we will no longer need to treat parallel lines as a special case. In the affine plane, distinct lines with the same slope do *not* intersect.

The points on a line may be parametrised using linear functions of a single parameter t. For rational lines, this may be done in such a way that rational points correspond to rational values of t.

Proposition 2.4. *The line $\mathcal{L}$ defined by* (7) *may be parametrised as follows.*

$$t \mapsto P_t = (t, -(at + c)/b) \quad \textit{if } b \neq 0; \tag{8}$$

or

$$u \mapsto Q_u = (-(bu + c)/a, u) \quad \textit{if } a \neq 0. \tag{9}$$

(If $ab \neq 0$ this gives two different parametrizations.)

If $\mathcal{L}$ is rational, then these parametrizations have the property that P_t is rational if and only if $t \in \mathbb{Q}$ (and similarly for Q_u), and so give bijections between $\mathbb{Q}$ and $\mathcal{L}(\mathbb{Q})$.

So rational lines have infinitely many rational points, while a non-rational line cannot have more than one rational point by Proposition 2.1. For example, the (real) line $Y = \sqrt{2}X$ has only one rational point, namely $(0, 0)$, while the line $X + Y = \sqrt{2}$ has no rational points.

There are other ways of parametrising affine lines. If $P_0 = (x_0, y_0)$ is one point on (7) then $t \mapsto P_t = (x_0 + bt, y_0 - at)$ is a parametrisation, with the property that if both the line and P_0 are rational then P_t is rational iff $t \in \mathbb{Q}$. Alternatively, if $P_1 = (x_1, y_1)$ is a second point on the line then $t \mapsto P_t = (1 - t)P_0 + tP_1 = (x_0 + t(x_1 - x_0), y_0 + t(y_1 - y_0))$ gives a parametrisation. Note that all these parametrizations have the form

$$t \mapsto (u(t), v(t))$$

where u and v are linear polynomials in t. Conversely, given two linear polynomials $u(T) = pT + q$, $v(T) = rT + s$ which are not both constant (so $(p, r) \neq (0, 0)$), the map $t \mapsto (u(t), v(t))$ parametrises the line $aX + bY + c = 0$ where $(a, b, c) = (-r, p, qr - ps)$.

Definition 6. An *affine conic* is an affine plane curve of degree 2.

The most general equation of an affine conic is thus

$$g(X,Y) = aX^2 + bXY + cY^2 + dX + eY + f = 0,$$

with $(a,b,c) \neq (0,0,0)$. We include degenerate (reducible) cases where g is a product of two linear factors (possibly over an extension field): if $g(X,Y) = l(X,Y)m(X,Y)$ with l, m linear then $\mathcal{C}_g$ is the union of the lines $\mathcal{C}_l$ and $\mathcal{C}_m$. In the most degenerate case, $l = m$ (up to a constant factor) and then $\mathcal{C}_g$ is a "double line".

Examples: (1) $\mathcal{C}: \; X^2 - 2Y^2 = 0$ is reducible, and is the union of the lines $X = \pm\sqrt{2}Y$. Note that the lines are not rational, though $\mathcal{C}$ itself is.

(2) $\mathcal{C}: \; X^2 - aY^2 = 1$ is a conic, irreducible when $a \neq 0$. If $a \in \mathbb{R}$, this is a hyperbola when $a > 0$ and an ellipse when $a < 0$. When $a = 0$ we have the union of the lines $X = \pm 1$.

(3) $\mathcal{C}: \; X^2 + Y^2 = 3$ is a rational conic, which has no rational points. The real points form a circle (of radius $\sqrt{3}$ centred at the origin).

Later we will see how to distinguish the different kinds of irreducible real conics: ellipses, hyperbolas and parabolas. The question of determining whether an irreducible rational conic has any rational points or not is harder; but given one rational points, we will see how to parametrise all the rational points via rational functions. This will involve intersecting the conic with rational lines, so we will need to discuss this situation in some detail.

2.4 Singular points, smooth points and tangents

Intuitively, a "smooth" point on a curve is one where the curve has a (unique) tangent. Other points are "singular": in the real plane, these occur at "corners" or at places where the curve crosses itself. For example, the curve $Y^2 = X^3$ has a corner (called a "cusp") at $(0,0)$, while $Y^2 = X^2(X+1)$ crosses itself at $(0,0)$ (called a "node"). The idea of this section is to formulate an algebraic definition of singularity which fits this intuition in the case of real curves, but which applies over an arbitrary field.

Notation: for a polynomial $f(X,Y)$ we denote by f_X and f_Y its partial derivatives with respect to X and Y. (Similarly for polynomials in more variables.) The derivative is defined purely algebraically, over any field, by the usual formulae: if

$$f = \sum_{i\geq 0, j\geq 0} a_{ij} X^i Y^j$$

then

$$f_X = \sum_{i\geq 1, j\geq 0} i a_{ij} X^{i-1} Y^j \quad \text{and} \quad f_Y = \sum_{i\geq 0, j\geq 1} j a_{ij} X^i Y^{j-1}.$$

These operations satisfy the usual rules from calculus: they are both K-linear, satisfy the usual product formula $(gh)_X = g_X h + g h_X$, chain rule, and so on. In most situations, both f_X and f_Y have degree $\deg(f) - 1$. The only exceptions are in characteristic p: for example $f = X^p \in \mathbb{F}_p[X]$ has degree p, but $f_X = 0$. This pathological behaviour complicates the general theory considerably, but in this module we will usually work over subfields of $\mathbb{C}$, where such unexpected things do not occur.

Definition 7. 1. Let $\mathcal{C}$ be an affine plane curve defined over the field K by a polynomial $f(X,Y) \in K[X,Y]$, and $P = (x_0, y_0) \in \mathcal{C}(L)$ for some $L \supseteq K$, so that $f(x_0, y_0) = 0$. We say that P is a *singular* point of $\mathcal{C}$ if

$$f_X(x_0, y_0) = f_Y(x_0, y_0) = 0;$$

otherwise, P is a *non-singular* or *smooth* point.

2. An affine plane curve $\mathcal{C}$ defined over the field K is called *non-singular* or *smooth* if all points of $\mathcal{C}(L)$ are non-singular for all finite extensions L of K. Equivalently, $\mathcal{C} = \mathcal{C}_f$ is smooth if the equations

$$f(x, y) = f_X(x, y) = f_Y(x, y) = 0 \tag{10}$$

have no common solutions either in K or in any extension of K.

Note that for a curve defined over K to be smooth, it is in general not sufficient that all its K-rational points are smooth. For example, for a real curve to be smooth it is necessary that all its complex points are smooth as well as its real points. For example the real curve $f = Y^2 + (X^2 + 1)^2$ has no real points, but is not smooth since the complex points $(\pm i, 0)$ are singular. If K is an algebraically closed field (for example, $K = \mathbb{C}$) then it is sufficient to look at K-rational points, since such a field has no finite extensions.

To find singular points on a curve, one has to solve the simultaneous equations (10).

Examples: (1) The line $f = aX + bY + c = 0$ is smooth at each of its points, since $f_X = a$ and $f_Y = b$.

(2) The Fermat curves $\mathcal{F}_n$ defined by $f = X^n + Y^n - 1$ are smooth over $\mathbb{C}$. For $f_X = nX^{n-1}$ is only 0 when $X = 0$, and similarly f_Y is only 0 when $Y = 0$, but $(0, 0)$ is not on the curve. [Note that $\mathcal{F}_n$ is not smooth over a field K whose characteristic is a prime p which divides n, for then $f_X = f_Y = 0$, so every point is singular! In this case, $f = (X^{n/p} + Y^{n/p} - 1)^p$.]

(3) Let $g(X) \in K[X]$ be a non-constant polynomial and set $f(X, Y) = Y^2 - g(X)$, so $\mathcal{C} = \mathcal{C}_f$ is the curve with equation $Y^2 = g(X)$. Now $f_Y = 2Y$ and $f_X = -g'(X)$, so the singular points (if any) are the points $P = (x, 0)$ where $g(x) = g'(x) = 0$. So $\mathcal{C}$ is smooth if and only if $g(X)$ has no repeated roots (equivalently, if $\gcd(g, g') = 1$).

Warning: when $\deg(g) > 3$ these curves have singular points "at infinity", so they are never smooth when viewed as projective curves. Later we will revise our definition of smooth curve to allow for this.

(4) The Bachet-Mordell curve $Y^2 - X^3 = c$ is smooth at $P = (a, b)$ unless $3a^2 = 2b = 0$: that is, unless $a = b = 0$. But $(0, 0)$ is only on the curve when $c = 0$. So the curve is smooth if $c \neq 0$, and has the unique singular point $(0, 0)$ when $c = 0$. [This is a special case of the previous example, with $g(X) = X^3 + c$.]

Reducible curves are almost never smooth: in fact, in the projective plane they are never smooth, but in the affine plane it can happen that the singular points are all at infinity and hence invisible. For if $f = gh$, then for any point $P = (x, y) \in \mathcal{C}_g \cap \mathcal{C}_h$, it is easy to check that $f(P) = f_X(P) = f_Y(P) = 0$. So the only way to avoid singularities is for $\mathcal{C}_g \cap \mathcal{C}_h$ to be empty.

For example, take $f(X, Y) = X(X - 1)$; the affine curve $\mathcal{C}_f$ has no singular points, as the lines $X = 0$ and $X = 1$ do not intersect in the affine plane. We will see later that these lines do intersect at infinity, and the projective completion of $\mathcal{C}_f$ is singular.

At each of its smooth points, a curve has a tangent line, defined as follows.

Definition 8. Let $\mathcal{C} = \mathcal{C}_f$ be an affine curve defined over the field K, and $P = (x_0, y_0) \in \mathcal{C}(K)$ a smooth point. The *tangent line to* $\mathcal{C}$ *at* P is the line

$$f_X(x_0, y_0)(X - x_0) + f_Y(x_0, y_0)(Y - y_0) = 0. \tag{11}$$

Note that the coefficients of X and Y in (11) are not both zero since P is smooth, so (11) does define a genuine line $\mathcal{L}$. From (11) we see that $\mathcal{L}$ is also defined over K, and that $P \in \mathcal{L}(K)$. For example, the tangent to a rational curve at each of its smooth points is a rational line.

Examples: (1) The tangent to the line $\mathcal{L}: \ f = aX + bY + c = 0$ at each of its points is $\mathcal{L}$ again.

(2) The tangent to $X^2 + Y^2 = 1$ at $P = (a, b)$ is $2a(X - a) + 2b(Y - b) = 0$, which simplifies to $aX + bY = 1$.

(3) The tangent to the Bachet-Mordell curve $Y^2 = X^3 + c$ at $P = (a, b)$ is (after simplifying) $3a^2X - 2bY = a^3 - 2c$. [This is valid unless $c = 0$ and $P = (a, b) = (0, 0)$.]

To see where the equation for a tangent comes from, observe that the expansion of the polynomial $f(X, Y)$ about the point $P = (a, b)$ has the form

$$f(X, Y) = c_0 + c_{01}(X - a) + c_{10}(Y - b) + \text{terms of higher degree in } (X - a),\ (Y - b).$$

(This is just Taylor's theorem; since f is a polynomial, the number of terms is finite.) We have $f(a, b) = c_0$, $f_X(a, b) = c_{01}$ and $f_Y(a, b) = c_{10}$. Hence $P \in \mathcal{C}_f \iff c_0 = 0$, and then P is smooth iff the linear part is not identically zero. The tangent at P has equation given by the linear part of the expansion. This fits with the intuitive idea that the tangent should be a "linear approximation" to the curve, which is "close" to the curve when $X - a$ and $Y - b$ are "small".

2.5 Intersecting curves

If we intersect two curves with degrees m and n, then intuitively we expect there to be mn points of intersection, at least in general. For example, intersecting two lines ($m = n = 1$) should normally give one point of intersection; but the case of parallel lines shows that this is not necessarily the case. In fact, for curves of general degree, there are several ways in which this rule may fail to hold. First of all, we should ensure that the curves do not have a component in common; equivalently, their defining polynomials should be coprime. Then it is true that the number of intersection points is finite, and is at most mn, but the number may well be strictly less than this bound.

To prove the previous statement in general would require more algebraic tools than we have at present (though proving just the finiteness is not hard). Here we will only consider the case where one of the curves is a line. This is the only case we will need to study conics and elliptic curves.

To illustrate the ways in which the number of intersection points can be less than expected, consider the intersection of the curve

$$\mathcal{C}: \quad Y = aX^2 + bX + c$$

and the line

$$\mathcal{L}: \quad Y = 0.$$

$\mathcal{C}$ has degree 2, unless $a = 0$, and $\mathcal{L}$ has degree 1. In other words, we are counting the number of roots of the polynomial $g(X) = aX^2 + bX + c$, since the curves intersect at points $(x_0, 0)$ where $g(x_0) = 0$. The assumption that $\mathcal{C}$ and $\mathcal{L}$ have no common component is equivalent to $(a, b, c) \neq (0, 0, 0)$, so $g(X)$ is not identically zero. There are three different ways in which the number of distinct roots x_0 may fail to equal 2:

1. If $a = 0$, then $g(X)$ has degree (at most) 1; there is only one root $x_0 = -c/b$ if $b \neq 0$, and none if $b = 0$ since under our assumption we cannot also have $c = 0$.

2. If $a \neq 0$ and $\Delta = b^2 - 4ac = 0$ then there is a "double root" $x_0 = -b/(2a)$; in this case, $\mathcal{L}$ is tangent to $\mathcal{C}$ at $(x_0, 0)$.

3. If $a \neq 0$, $\Delta = b^2 - 4ac \neq 0$ and Δ is not a square in the field K, then there are no roots at all. For example, this would happen over $\mathbb{R}$ if $\Delta < 0$, so the roots of $g(X)$ are complex (non-real). Geometrically, the curve and line do not intersect at all in the real plane $\mathbb{A}^2(\mathbb{R})$, but they have two distinct intersection points in $\mathbb{A}^2(\mathbb{C})$.

More generally, suppose we intersect a curve $\mathcal{C} = \mathcal{C}_f$ of degree n with a line $\mathcal{L}$, both defined over a field K. Parametrise the line via $t \mapsto P_t = (u(t), v(t))$, where $u(T)$ and $v(T)$ are linear polynomials in $K[T]$ (or constant), as in Proposition 2.4. Then the intersection points are the points P_t where t is a root of the polynomial (in one variable)

$$h(T) = f(u(T), v(T)) \in K[T].$$

Provided that $\mathcal{L} \not\subseteq \mathcal{C}$, $h(T)$ is not identically zero, and is a polynomial of degree at most n. Hence $\mathcal{C} \cap \mathcal{L}$ consists at most n points. The exact number of points in $\mathcal{C}(L) \cap \mathcal{L}(L)$ for any field $L \supseteq K$ is the number of distinct roots of $h(T)$ in L. As in the special case $n = 2$ considered above, there are three ways this number may be less than n:

1. We may have $\deg(h) < n$.

2. There may be repeated roots.

3. The field L may not be large enough for $h(T)$ to factorize completely.

Examples of the last possibility would be if $K = \mathbb{Q}$ and some of the roots of $h(T)$ are irrational, or if $K = \mathbb{R}$ and $h(T)$ has some complex (non-real) roots.

Possibility 1 happens when $\mathcal{L}$ is parallel to an asymptote of $\mathcal{C}$. Examples when $n = 2$ will be given below. We will be able to eliminate this possibility by working in the larger projective plane: we will see that the "missing" intersection points are points "at infinity", and that asymptotes are lines which are tangent to the curve at a point at infinity.

Possibility 2 is best handled by defining the multiplicity of intersection as follows.

Definition 9. In the notation above, the *multiplicity* of a point P on $\mathcal{C} \cap \mathcal{L}$ is $m > 0$ if $P = P_t$ where t is a root of multiplicity m of $h(T)$, and 0 otherwise.

For example, when $\Delta = 0$ in the earlier example, we only have one intersection point between the line $Y = 0$ and the curve $\mathcal{C}$, but it has multiplicity 2.

Possibility 3 in unavoidable unless we wish to work exclusively in an algebraically closed field such as $\mathbb{C}$, where polynomials of degree n all have exactly n roots, when counted with the correct multiplicities. Classical algebraic geometry is usually studied in such a context; but for number-theoretical applications, where we wish to work over a field such as $\mathbb{Q}$ or a finite field, this is not the case. This is the main difference between "arithmetic algebraic geometry", to which this module is an introduction, and classical algebraic geometry.

We may sum up the above discussion as follows.

Proposition 2.5. *Let $\mathcal{C}$ be a curve of degree n and $\mathcal{L}$ a line, both defined over a field K, with $\mathcal{L} \not\subseteq \mathcal{C}$. The number of points of intersection of $\mathcal{C}$ and $\mathcal{L}$ is finite, and at most n. Over an algebraically closed field (such as $\mathbb{C}$), if we count each intersection point with its multiplicity, then the number of intersection points is exactly n* unless *$\mathcal{L}$ is parallel to an asymptote of $\mathcal{C}$.*

A more satisfactory projective version of this result will be given later, where the exceptional case is eliminated. For the moment, we merely observe that only a finite number of slopes of $\mathcal{L}$ are excluded, since $\mathcal{C}$ has only a finite number of asymptotic directions (possibly zero).

The main difficulty in generalising to an arbitrary intersection of two curves is that we can no longer reduce the definition of multiplicity to properties of a polynomial in one variable; that relied on one of the curves (the line $\mathcal{L}$) having a parametrisation, and (as we saw in the Introduction) this is not always the case. The general definition requires more advanced algebraic concepts.

In the next section we will study the possible intersections of an irreducible real conic with a general line, which will lead us to their classification into ellipses, hyperbolas and parabolas.

2.6 Classification of affine real conics

A reducible conic $\mathcal{C}$ has the form $f = gh$ where g, h are linear, so $\mathcal{C}_f$ is the union of the lines $\mathcal{C}_g$ and $\mathcal{C}_h$, which may be the same line or distinct. We call $\mathcal{C}$ a double line if $\mathcal{C}_g = \mathcal{C}_h$. Let K be a field of definition of $\mathcal{C}$, so (after scaling if necessary) $f \in K[X, Y]$. As we saw in earlier examples, g and h may not be K-rational. However, when $\mathcal{C}$ is a double line, then $f = cg^2$ with $c \in K^*$, and f is rational iff g is, since $f' = 2cg \in K[X, Y]$.

Every point on a double line is singular. Suppose that $\mathcal{C}$ is a union of distinct lines. Then $\mathcal{C}$ has at most one singular point, namely the point P (if any) where the two lines intersect. Moreover this point is K-rational, even if the lines are not: this can be seen by observing that P is also the point of intersection of the K-rational lines f_X and f_Y. We leave the details as an exercise.

From now on we will restrict our attention to irreducible conics over $\mathbb{R}$. A second exercise shows that irreducible conics are smooth.

Let the equation of the real conic $\mathcal{C}$ be

$$g(X, Y) = aX^2 + bXY + cY^2 + dX + eY + f = 0, \tag{12}$$

with $(a, b, c) \neq (0, 0, 0)$, and all coefficients real. The curve is either an ellipse, a hyperbola or a parabola, depending on the sign of $\Delta = b^2 - 4ac$. Consider the intersection of $\mathcal{C}$ with the line $\mathcal{L} :\ Y = mX + n$ of slope m. Parametrise $\mathcal{L}$ by $t \mapsto P_t = (t, mt + n)$. Substituting into g, we obtain

$$h(T) = g(T, mT + n) = (a + bm + cm^2)T^2 + (bn + 2cmn + d + em)T + (cn^2 + en + f).$$

Provided that $r(m) = a + bm + cm^2 \neq 0$, this is a genuine quadratic in T. In this case, h has at most two real roots. If $h(T)$ has distinct real roots t_1 and t_2, then $\mathcal{C}(\mathbb{R}) \cap \mathcal{L}(\mathbb{R})$ consists of the two points P_{t_1} and P_{t_2}. If $h(T)$ has one double root t_0 (necessarily real), then $\mathcal{L}$ is tangent to $\mathcal{C}$ at P_{t_0}, which is the only point of intersection. If the roots of $h(T)$ are non-real, then $\mathcal{L}$ does not intersect $\mathcal{C}$ at all in the real plane $\mathbb{A}^2(\mathbb{R})$, but there are two complex intersection points.

There are at most two exceptional values for the slope m such that $r(m) = a+bm+cm^2 = 0$, so that $\deg(h) < 2$. (We include $m = \infty$ when $c = 0$, since then we have $h(T) = g(T, n)$ which has degree 2 unless $c = 0$.) These are called the *asymptotic* directions of $\mathcal{C}$. If $\Delta > 0$, then there are two such directions, which are the slopes of the two asymptotes of $\mathcal{C}$, which is a hyperbola. If $\Delta = 0$, then $\mathcal{C}$ is a parabola, and there is just one asymptotic direction, parallel to the axis. Finally, if $\Delta < 0$ then there are no (real) asymptotic directions, and $\mathcal{C}$ is an ellipse.

Examples: (1) $\mathcal{C} :\ X^2 + Y^2 = 1$. Here $r(m) = 1 + m^2$ has no real roots ($\Delta = -4$) so $\mathcal{C}$ is an ellipse (in fact a circle, which is a special kind of ellipse). Intersecting with $Y = 0$ gives two real intersection points $(\pm 1, 0)$. Intersecting with $Y = 1$ gives a single intersection at $(0, 1)$, with multiplicity 2: the line $Y = 1$ is tangent to $\mathcal{C}$ at $(0, 1)$. Intersection with $Y = 2$ gives no real intersections; there are two complex intersection points in this case, namely $(\pm i\sqrt{3}, 2)$.

(2) $\mathcal{C} :\ X^2 - Y^2 = 1$. Now $r(m) = 1 - m^2$ has distinct real roots $m = \pm 1$ (we have $\Delta = +4$), and $\mathcal{C}$ is a hyperbola with asymptotes having slopes ± 1. In fact the asymptotes are the lines $Y = \pm X$. If we intersect $\mathcal{C}$ with either of these lines, there are no intersection points. This is a different situation from that in the previous example, where there were complex intersection points, as here there are no complex intersection points. Instead, the intersection points are "at infinity" and we will need to enlarge the affine plane into the projective plane in order to see them. Other lines with slope ± 1 do intersect $\mathcal{C}$, but in just one point, and are *not* tangents. For example, $Y = X + 1$ intersects $\mathcal{C}$ at $(-1, 0)$ only.

(3) $\mathcal{C} :\ X - Y^2 = 0$. This is a parabola, with unique asymptotic direction $m = 0$. All horizontal lines (with slope 0) intersect the curve just once, non-tangentially.

2.7 Existence of rational points on rational conics

In the next section, we will see how to parametrise all the rational points on a rational conic, given one rational point to use as a base point. Of course, if there are no rational points at all (as with $X^2 + Y^2 = 3$) then there is no such parametrisation. In this section, we consider the related questions: given a rational conic, how do we determine whether or not it has a rational point, and how do we find a rational point in practice?

Theoretically, the most satisfactory result concerning the existence of rational points on conics is the following.

Theorem 2.6 (Hasse's Principle for conics). *Let $\mathcal{C}$ be a rational conic. Then*

$$\mathcal{C}(\mathbb{Q}) \neq \emptyset \iff \mathcal{C}(\mathbb{R}) \neq \emptyset \text{ and } \mathcal{C}(\mathbb{Q}_p) \neq \emptyset \text{ for all primes } p.$$

Here, $\mathbb{Q}_p$ denotes the field of p-adic numbers. In one direction, this result is obvious, since $\mathbb{Q} \subset \mathbb{R}$ and $\mathbb{Q} \subset \mathbb{Q}_p$ for all p. The other direction is far from obvious, and we will not prove it here, though it is not hard to show that it follows from Legendre's Theorem below. This is an example of what is called a "local-global" principle: the fields $\mathbb{R}$ and $\mathbb{Q}_p$ are examples of so-called "local fields", while $\mathbb{Q}$ is a "global field". Studying the points of a rational curve which are defined over one of these local fields is often simpler than studying the rational points directly.

What the local-global principle for conics means in more concrete terms is this: if a rational conic has no rational points, then we will be able to prove this, either by seeing that it has no real points, or by using congruences modulo some prime power. For example, the conic $\mathcal{C}: \ X^2 + Y^2 = 3$ does have points in all the local fields except for $\mathbb{Q}_2$ and $\mathbb{Q}_3$, which explains how we were able to prove that $\mathcal{C}(\mathbb{Q})$ is empty in two ways, either by using congruences modulo 4 or by using congruences modulo 9. In fact, the number of local fields in which a rational conic fails to have points is always *even* (possibly zero); the proof of this statement follows from quadratic reciprocity! We will not prove it in this module.

For curves of higher degree, the local-global principle breaks down. While in some cases one can prove that a curve has no rational points by showing that it has no local points for some prime, there do exist rational curves with real and p-adic points for all primes p, yet which have no rational points. One example of such a curve (due to Selmer) is $3X^3 + 4Y^3 = 5$. The study of these questions involves an interplay between geometry, algebra and number theory, and is called *Arithmetic Algebraic Geometry.*

Let $\mathcal{C}$ be an irreducible rational conic with equation (12). Set $\Delta = b^2 - 4ac$.

If $\Delta = 0$, so $\mathcal{C}$ is a parabola, then it is easy to write down rational points since every rational line with slope $m = -b/(2c)$ intersects $\mathcal{C}$ exactly once, at a rational point. (If $c = 0$ then use a vertical line.)

For example, consider $X^2 - 2XY + Y^2 + X - 2 = 0$. The quadratic part is $(X - Y)^2$, so the asymptotic slope is $m = 1$. The line $Y = X$ cuts the curve where $X - 2 = 0$, so we have a rational point $(2, 2)$.

If Δ is a rational square, then $\mathcal{C}$ is a hyperbola whose asymptotes have rational slopes; any rational line parallel to an asymptote will intersect $\mathcal{C}$ in one rational point.

For example, $g(X, Y) = X^2 - XY - 2Y^2 + 2X + 2 = 0$ has $\Delta = 9 = 3^2$, and the quadratic part factorizes as $(X + Y)(X - 2Y)$. Setting $Y = -X$ gives $2X + 2 = 0$, so $X = -1$ and we have a rational point $(-1, 1)$.

In other cases there are no such simple tricks to get rational points. From now on we assume that we are not in the parabolic case, so $\Delta \neq 0$.

By shifting the origin we may assume that the coefficients of X and Y are both zero; to do this, we replace X by $X + \alpha$ and Y by $Y + \beta$, where α, β satisfy the simultaneous linear equations

$$\begin{aligned} 2a\alpha + b\beta &= -d, \\ b\alpha + 2c\beta &= -e; \end{aligned}$$

a unique solution exists since the system has determinant Δ, which is non-zero. This changes the constant term f but does not affect the coefficients of the quadratic terms.

Next, we may complete the square to eliminate the XY term. Our equation now has the form

$$aX^2 + bY^2 = c.$$

The three coefficients (which are not the original a, b, c) are still rational, since the changes of variable we have carried out only involved rational numbers. We may scale the equation to make the coefficients integral. Also, we have $abc \neq 0$, since otherwise $\mathcal{C}$ is reducible. (Note that the irreducibility will not be affected by the changes of variables.) Finding a rational solution to this "diagonal" equation is equivalent to finding an integer solution other than $(0, 0, 0)$ to the homogeneous equation $aX^2 + bY^2 = cZ^2$, which is Legendre's equation.

By permuting and scaling the variables we may assume that a, b, c are all *positive square-free* integers which are *pairwise coprime*; see the G13NUM lecture notes for more details of this. Now the existence of solutions may be checked by applying Legendre's Theorem.

Theorem 2.7 (Legendre's Theorem). *Let a, b, c be positive integers which are square-free and pairwise coprime. Then the equation*

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

has a nontrivial integer solution if and only if each of these three quadratic congruences has a solution:

$$U^2 \equiv bc \pmod{a}, \quad V^2 \equiv ac \pmod{b}, \quad W^2 \equiv -ab \pmod{c}.$$

See G13NUM notes for a proof of this. One way of finding a solution follows from that proof, since when a solution exists there is always one which satisfies the inequalities $|X| \leq \sqrt{bc}$, $|Y| \leq \sqrt{ac}$, $|Z| \leq \sqrt{ab}$. (This is Holzer's Theorem.) But when the coefficients are at all large, searching all points (X, Y, Z) in this box would take too long (there are approximately $8abc$ such points to be considered). Instead, we may use a reduction method, which is in fact the method Legendre himself used to prove his theorem. (It is also the method which Maple uses to solve equations like this.)

Changing notation yet again, we may write our equation in the form

$$X^2 - aY^2 = bZ^2,$$

where a, b are square-free non-zero integers (not necessarily coprime). We seek a non-trivial integer solution, which we may assume to be primitive; that is, $\gcd(X, Y, Z) = 1$. By symmetry we may assume that $0 < |a| \leq |b|$. The idea is to replace the equation with coefficients a, b with one with smaller coefficients, repeating the process until we reach a base case which is easy to solve. At each stage we make a change of variables, which enables us to use back-substitution to recover a solution to the original equation from the solution to the base case.

Base cases: (1) If $b = 1$ then $(X, Y, Z) = (1, 0, 1)$ is a solution. (2) If $a = 1$ then $(1, 1, 0)$ is a solution. (3) If $a = -b$ then $(0, 1, 1)$ is a solution. (4) If $(a, b) = (-1, -1)$ then there are no solutions, since no real solutions exist. We may now assume $|b| \geq 2$.

Reduction step: If there is a primitive solution, then from $X^2 \equiv aY^2 \pmod{b}$ it follows that a is congruent to a square modulo b. We solve $x_0^2 \equiv a \pmod{b}$ for x_0, where we may suppose that $|x_0| \leq |b|/2$. Set $x_0^2 - a = bt$. From $|b| \geq 2$ and the bounds on x_0 we deduce that $|t| < |b|$.

Now we solve the equation $X_2^2 - aY_2^2 = tZ_2^2$, which has smaller coefficients. Let (x_2, y_2, z_2) be a solution. Set

$$x_1 = x_0x_2 - ay_2, \quad y_1 = x_2 - x_0y_2, \quad z_1 = tz_2;$$

then (x_1, y_1, z_1) satisfy $x_1^2 - ay_1^2 = bz_1^2$. The explanation for these formulas comes from the identity

$$(x_0 + \sqrt{a})(x_2 - y_2\sqrt{a}) = x_1 + y_1\sqrt{a}.$$

Moreover, $(x_1, y_1, z_1) = (0, 0, 0) \iff (x_2, y_2, z_2) = (0, 0, 0)$, provided that $x_0^2 \neq a$, which holds since a is square-free and not 1.

When we solve the smaller equation with coefficients a, t, we must first replace t by its square-free part t_0, scaling Z_2 to compensate, and also interchange t_0 with a if $|t_0| < |a|$. Unfortunately, finding t_0 involves factorizing t, which for large examples can be very time-consuming. It is possible to make this reduction method of solution much more efficient, but we will not go into further details here.

To solve a congruence such as $x_0^2 \equiv a \pmod{b}$, we first solve the congruence modulo each of the prime factors of b and then use the Chinese Remainder Theorem. Again, there are efficient algorithms for solving quadratic congruences modulo a prime. It is unavoidable to have to factorize b here. If any of these quadratic congruences fails to have a solution, then the original equation is insoluble.

See the separate handout for a complete worked example of this method.

2.8 Parametrisation of affine conics

Let $\mathcal{C}$ be an irreducible conic with equation $g(X, Y) = 0$ as in (12), and $P = (x_0, y_0)$ a point on $\mathcal{C}$. We will show how to parametrise (almost) all points on $\mathcal{C}$ by rational functions, generalising the parametrisation of the circle given in (4). The idea is to use P as a base point, consider lines through P with variable slope t, and find the second point where this line intersects $\mathcal{C}$. (See the picture in lectures.) This gives an almost bijective function between values of t and points on $\mathcal{C}$; one point, corresponding to $t = \infty$, may be excluded, and up to two values of t may not correspond to points on $\mathcal{C}$ if the line with slope t does not intersect $\mathcal{C}$ again. The latter only happens for a hyperbola, when the exceptional values of t are the asymptotic directions.

If $\mathcal{C}$ is a rational conic and P a rational point on $\mathcal{C}$, this will give a bijection between $\mathbb{Q}$ (omitting at most two values) and $\mathcal{C}(\mathbb{Q})$ (omitting at most one point).

The exceptions will go away when we work in the projective plane.

Now we may parametrise the line $\mathcal{L}_t$ through $P = (x_0, y_0)$ with slope t by $u \mapsto (x_0 + u, y_0 + tu)$. This line intersects $\mathcal{C}$ where $h(u) = g(x_0 + u, y_0 + tu) = 0$. Note that $h(0) = 0$, since $P \in \mathcal{C}$. So the polynomial $h(U)$ has the form $h(U) = AU^2 + BU = U(AU + B)$. Here, A and B depend on t; in fact, A is (at worst) quadratic in t and is zero iff t is an asymptotic direction. So provided that t is not one of these exceptional values, there is a second root of $h(U)$ at $u = -B/A$. Substituting this value into the parametrisation of $\mathcal{L}_t$ gives the point $(x_0 - B/A, y_0 - Bt/A)$ on $\mathcal{C} \cap \mathcal{L}_t$, whose coordinates are rational functions of t.

It is not very illuminating to write down a general formula for the parametrisation in terms of x_0, y_0 and the coefficients of $g(X, Y)$. Instead, we give examples.

Examples: (1) $g(X, Y) = X^2 + Y^2 - 1$, so $\mathcal{C}$ is the unit circle. Take $P = (-1, 0)$. The line with slope t through P is $Y = t(X + 1)$. Parametrise this by $u \mapsto (u - 1, tu)$ (with $u = 0$ mapping to P) and substitute $(X, Y) = (u - 1, tu)$ into the equation to get

$$0 = X^2 + Y^2 - 1 = (u - 1)^2 + t^2u^2 - 1 = (1 + t^2)u^2 - 2u = h(u).$$

One root of $h(u)$ is $u = 0$ as expected, and the second is $u = 2/(1 + t^2)$. Substituting back gives

$$(X, Y) = (u - 1, tu) = \left(\frac{1 - t^2}{1 + t^2}, \frac{2t}{1 + t^2}\right) = P_t,$$

say, as in (4). (In the notation above, we have $A = 1 + t^2$ and $B = -2$.) Exceptional values of t are those for which $t^2 + 1 = 0$; there are clearly none over $\mathbb{Q}$ or $\mathbb{R}$. The exceptional point on $\mathcal{C}$ is P itself, since the tangent at P is the vertical line $X = -1$; for any other point $Q = (x, y)$ on $\mathcal{C}$, the line from P to Q has finite slope $t = y/(x + 1)$ and is given by the above formula.

Hence every rational point $Q = (x, y) \in \mathcal{C}(\mathbb{Q})$, apart from P itself, has the form $Q = P_t$ where $t = y/(x+1) \in \mathbb{Q}$, and our parametrisation gives a bijection between $\mathbb{Q}$ and $\mathcal{C}(\mathbb{Q}) \setminus \{P\}$.

(2) Let $\mathcal{C}$ be the rational circle $X^2+Y^2=2$ with base point $P=(1,1)$. Parametrising the lines through P via $u \mapsto (1+u, 1+tu)$ gives the parametrisation

$$(X,Y)=\left(\frac{-1-2t+t^2}{1+t^2}, \frac{1-2t-t^2}{1+t^2}\right).$$

There are no exceptional rational values of t; the excluded point on $\mathcal{C}$ is not P but $Q=(1,-1)$, since the line PQ is vertical. This can also be seen by letting $t \to \infty$ in the parametrisation formula. Points $Q=(x,y)$ with $x \neq 1$ come from $t=(y-1)/(x-1)$, while $P=(1,1)$ itself comes from $t=-1$, since this is the slope of the tangent to $\mathcal{C}$ at P.

(3) Let $\mathcal{C}$ be the curve with equation $g(X,Y)=X^2-XY-2Y^2+2X+2=0$. This is a hyperbola, with rational asymptotes parallel to $X+Y$ and $X-2Y$, since the quadratic part of $g(X,Y)$ is $X^2-XY-2Y^2=(X+Y)(X-2Y)$. So the parametrisation of $\mathcal{C}$ will have two exceptional values. Taking $P=(0,1)$ and proceeding as above, we find the parametrisation

$$(X,Y)=\left(\frac{4t-1}{(1+t)(1-2t)}, \frac{1-2t+2t^2}{(1+t)(1-2t)}\right).$$

The omitted point in $(0,-1)$. The value of t mapping to $Q=(x,y)$ is $(y-1)/x$ in general; no value of t maps to $(0,-1)$, and $t=1/4$ maps to P itself.

(4) For the parabola $Y=X^2$ with $P=(0,0)$ we obtain the parametrisation $t \mapsto (t,t^2)$, so it is possible for there to be no exceptional points as well as no exceptional values of t. In this case the map $t \mapsto (t,t^2)$ is a bijection from $\mathbb{Q}$ to $\mathcal{C}(\mathbb{Q})$.

We sum up the results of this section:

Proposition 2.8. *Let $\mathcal{C}$ be a rational affine conic and P a rational point on $\mathcal{C}$. Then there are rational polynomials $q_i(T)$ for $i=1,2,3$, of degree at most 2, such that the map*

$$t \mapsto \left(\frac{q_1(t)}{q_3(t)}, \frac{q_2(t)}{q_3(t)}\right)$$

is a bijection between $\mathbb{Q} \setminus \{t \in \mathbb{Q} \mid q_3(t)=0\}$ and $\mathcal{C}(\mathbb{Q}) \setminus \{Q\}$, where the point Q is obtained by letting $t \to \infty$ in the above formula.

As with other results in this chapter, we will be able to formulate a rather simpler version of this result for projective conics, without the exceptional values and points.

3 Projective Curves

3.1 The projective line

Lemma 3.1. *Let (x_1, y_1), $(x_2, y_2) \in K^2 \setminus \{(0,0)\}$. Then*

$$x_1y_2 = x_2y_1 \iff \exists \lambda \in K^* : (x_2, y_2) = (\lambda x_1, \lambda y_1).$$

Define the relation $\sim$ on $K^2 \setminus \{(0,0)\}$ by $(x_1, y_1) \sim (x_2, y_2)$ iff these conditions hold; then $\sim$ is an equivalence relation.

Definition 10. The *projective line* $\mathbb{P}^1(K)$ over a field K is the set of equivalence classes

$$\mathbb{P}^1(K) = \{[x : y] \mid x, y \in K, (x, y) \neq (0,0)\},$$

where $(x, y) \sim (\lambda x, \lambda y)$ for all $\lambda \in K^*$ as above.

We call x, y *homogeneous coordinates* for the point $P = [x : y]$; they are only determined up to a non-zero scalar factor: $[x : y] = [\lambda x : \lambda y]$.

Lemma 3.2. *Every point in $P \in \mathbb{P}^1(K)$ may be uniquely written as either $P = [x : 1]$ with $x \in K$, or $P = [1 : 0]$.*

Identifying $[x : 1] \in \mathbb{P}^1(K)$ with $(x) \in \mathbb{A}^1(K)$, we then have $\mathbb{P}^1(K) = \mathbb{A}^1(K) \cup \{\infty\}$, where $\infty = [1 : 0]$. So the projective line is obtained by adding one new point "at infinity" to the affine line.

Lemma 3.3. *Every point $\mathbb{P}^1(\mathbb{Q})$ has homogeneous coordinates of the form $[x : y]$ where $x, y \in \mathbb{Z}$ and $\gcd(x, y) = 1$. These are unique, up to sign: $[x : y] = [-x : -y]$.*

Application: Lines through the origin in $\mathbb{A}^2$ have equations of the form $aX + bY = 0$ where $(a, b) \neq (0, 0)$, and $(\lambda a, \lambda b)$ determines the same line for all $\lambda \in K^*$. So these lines are parametrised by the point $[a : b] \in \mathbb{P}^1$, which determines the slope of the line (either $-a/b$, or ∞ if $b = 0$).

3.2 Homogeneous polynomials and their roots

Definition 11. 1. A polynomial $F(X, Y) \in K[X, Y]$ is *homogeneous of degree n* iff every term of F has degree n, so

$$F(X, Y) = \sum_{i=0}^{n} a_i X^i Y^{n-i}$$

with not all the a_i zero. A homogeneous polynomial (in two variables) is also called a (binary) *form.*

2. The point $P = [x : y] \in \mathbb{P}^1(K)$ is a *root* of the homogeneous polynomial $F(X, Y)$ iff $F(x, y) = 0$; this is well-defined since $F(\lambda x, \lambda y) = \lambda^n F(x, y)$.

A point of the form $[x : 1]$ is a root of $F(X, Y)$ iff x is a root of the dehomogenised one-variable polynomial $f(x) = F(x, 1)$. The point $\infty = [1 : 0]$ is a root iff $F(1, 0) = a_n = 0$, which is iff $\deg(f) < \deg(F)$. Conversely, given $f(X) \in K[X]$ of degree n, we can form the associated homogeneous polynomial $F(X, Y) = Y^n f(X/Y)$, which has no root at ∞, and a root at $[x : 1] \in \mathbb{P}^1$ iff x is a root of $f(X)$.

Over an algebraically closed field such as $\mathbb{C}$, every homogeneous $F(X, Y)$ of degree n factorizes as a product of n linear factors, uniquely up to the order and scaling of the factors. So F has exactly n roots in $\mathbb{P}^1(\mathbb{C})$, if counted with multiplicity. We have

$$F(X, Y) = \prod_{j=1}^{n} (u_j X - t_j Y)$$

where the (not necessarily distinct) roots are $[t_j : u_j]$ for $1 \leq j \leq n$. For a general field K, such a factorization will exist after extending to a larger field $L \supset K$ if necessary.

Examples 1. $F(X, Y) = X^2Y - 2XY^2 = X \cdot Y \cdot (X - 2Y)$ is homogeneous of degree 3. It has roots $[0 : 1]$ (from the factor X), $[1 : 0]$ (from the factor Y), and $[2 : 1]$ (from the factor $X - 2Y$). The dehomogenised polynomial $f(X) = F(X, 1) = X^2 - 2X$ has only two roots, at 0 and 2. The root $[1 : 0]$ of F is “at infinity”.

2. $F(X, Y) = X^2 + Y^2$ has no roots in $\mathbb{P}^1(\mathbb{R})$, but two roots $[\pm i : 1]$ in $\mathbb{P}^1(\mathbb{C})$.

3. $F(X, Y) = X^2Y^2 - 2Y^4 = Y^2(X - \sqrt{2}Y)(X + \sqrt{2}Y)$ has simple roots at $[\pm\sqrt{2} : 1]$ and a double root at $[1 : 0]$. The dehomogenised $f(X) = X^2 - 2$ has only two simple roots $\pm\sqrt{2}$; the double root has gone to infinity. The only rational root is the double root at $[1 : 0]$.

3.3 Projective space and the projective plane

Lemma 3.4. *Let* $(x_0, x_1, \ldots, x_n)$, $(y_0, y_1, \ldots, y_n) \in K^{n+1} \setminus \{(0, 0, \ldots, 0)\}$. *Then*

$$\big(\forall i, j : x_iy_j = x_jy_i\big) \iff \big(\exists \lambda \in K^* : \forall j : y_j = \lambda x_j\big).$$

Define the relation $\sim$ *on* $K^{n+1} \setminus \{0\}$ *by* $(x_0, x_1, \ldots, x_n) \sim (y_0, y_1, \ldots, y_n)$ *iff these conditions hold; then* $\sim$ *is an equivalence relation.*

Definition 12. *Projective n-dimensional space* $\mathbb{P}^n(K)$ over a field K is the set of equivalence classes

$$\mathbb{P}^n(K) = \{[x_0 : x_1 : \cdots : x_n] \mid x_i \in K, \text{not all } 0\},$$

where $(x_0, x_1, \ldots, x_n) \sim (\lambda x_0, \lambda x_1, \ldots, \lambda x_n)$ for all $\lambda \in K^*$ as above.

We call $x_0, x_1, \ldots, x_n$ *homogeneous coordinates* for the point $P = [x_0 : x_1 : \cdots : x_n]$; they are only determined up to a non-zero scalar factor: $[x_0 : x_1 : \cdots : x_n] = [\lambda x_0 : \lambda x_1 : \cdots : \lambda x_n]$.

The case $n = 1$ gives the projective line $\mathbb{P}^1$ as before. The case $n = 2$ defines the projective plane $\mathbb{P}^2$, where we will use x, y, z as names for the coordinates instead of x_0, x_1, x_2:

$$\mathbb{P}^2(K) = \{[x : y : z] \mid x, y, z \in K, \text{not all } 0\},$$

with $[x : y : z] = [\lambda x : \lambda y : \lambda z]$ for all $\lambda \in K^*$.

The following results hold for general n, but to ease notation we state them for $n = 2$. The proofs are almost the same as for $n = 1$.

Lemma 3.5. *Every point in* $P \in \mathbb{P}^2(K)$ *may be uniquely written as either* $P = [x : y : 1]$ *with* $x, y \in K$, *or* $P = [x : 1 : 0]$ *with* $x \in K$, *or* $P = [1 : 0 : 0]$.

Identifying $[x : y : 1] \in \mathbb{P}^2(K)$ with $(x, y) \in \mathbb{A}^2(K)$, and $[x : y : 0]$ with $[x : y] \in \mathbb{P}^1(K)$, we then have $\mathbb{P}^2(K) = \mathbb{A}^2(K) \cup \mathbb{P}^1(K) = \mathbb{P}^2(K) = \mathbb{A}^2(K) \cup \mathbb{A}^1(K) \cup \{[1 : 0 : 0]\}$. So the projective plane is obtained by adding a projective line of points $[x : y : 0]$ “at infinity” to the affine plane.

In the general case, every point in $\mathbb{P}^n(K)$ has a unique set of homogeneous coordinates in which the last non-zero coordinate is 1; the points for which $x_n \neq 0$ may be identified with $\mathbb{A}^n(K)$, and the points with $x_n = 0$ form a “hyperplane at infinity” which may be identified with $\mathbb{P}^{n-1}$.

Lemma 3.6. *Every point* $\mathbb{P}^2(\mathbb{Q})$ *has homogeneous coordinates of the form* $[x : y : z]$ *where* $x, y, z \in \mathbb{Z}$ *and* $\gcd(x, y, z) = 1$. *These are unique, up to sign:* $[x : y : z] = [-x : -y : -z]$.

3.4 Projective plane curves

Projective plane curves are defined via homogeneous polynomials in three variables X, Y, Z.

Definition 13. 1. A polynomial $F(X, Y, Z) \in K[X, Y, Z]$ is *homogeneous of degree* n iff every term of F has degree n, so

$$F(X,Y,Z) = \sum_{i=0}^{n}\sum_{j=0}^{i} a_{ij} X^j Y^{i-j} Z^{n-i}$$

with not all the a_{ij} zero. A homogeneous polynomial (in three variables) is also called a (ternary) *form*.

2. A point $P = [x : y : z] \in \mathbb{P}^2(K)$ is a *zero* of the homogeneous polynomial $F(X, Y, Z)$ if $F(x, y, z) = 0$.

Note that we cannot define the value of $F(X, Y, Z)$ at $P = [x : y : z]$ to be $F(x, y, z)$, since the homogeneous coordinates x, y, z are only defined up to a non-zero constant factor. But the homogeneity of F implies that

$$F(\lambda x, \lambda y, \lambda z) = \lambda^n F(x, y, z),$$

so the condition $F(x, y, z) = 0$ is independent of the homogeneous coordinates chosen.

Definition 14. A *projective plane curve defined over the field* K is an equivalence class of non-constant forms (homogeneous polynomials) $F(X, Y, Z) \in K[X, Y, Z]$, where two forms F and G are said to be equivalent if $G = cF$ for some $c \in K^*$. The curve determined by F will be denoted $\mathcal{C}_F$. The *degree* of $\mathcal{C}_F$ is $\deg(\mathcal{C}_F) = \deg(F)$.

Definition 15. Let $\mathcal{C} = \mathcal{C}_F$ be a projective plane curve defined by a form $F(X, Y, Z) \in K[X, Y, Z]$, and L a field with $K \subseteq L$. The *set of L-rational points* of $\mathcal{C}$ is

$$\mathcal{C}(L) = \{[x : y : z] \in \mathbb{P}^2(L) \mid F(x, y, z) = 0\}.$$

Since we have the decomposition $\mathbb{P}^2(L) = \mathbb{A}^2(L) \cup \mathbb{P}^1(L)$, we may decompose the set of points on a projective curve $\mathcal{C}_F$ into the "affine points", of the form $[x : y : 1]$, and the "points at infinity", of the form $[x : y : 0]$. To each form $F(X, Y, Z)$ we may associate its *dehomogenisation*

$$f(X, Y) = F(X, Y, 1) \in K[X, Y]$$

with associated affine curve $\mathcal{C}_0 = \mathcal{C}_f$; then under the identification $(x, y) \leftrightarrow [x : y : 1]$ we have

$$(x, y) \in \mathcal{C}_0(L) \iff [x : y : 1] \in \mathcal{C}(L).$$

We think of the affine curve $\mathcal{C}_0$ as the intersection of the projective curve $\mathcal{C}$ with the affine part $\mathbb{A}^2 \subset \mathbb{P}^2$. Note that we may have $\deg(f) < \deg(F)$; this will happen iff Z is a factor of F. More precisely, write

$$F(X,Y,Z) = \sum_{i=0}^{n} F_i(X,Y) Z^{n-i}, \tag{13}$$

where $F_i(X, Y) = \sum_{j=0}^{i} a_{ij} X^j Y^{i-j}$ is a binary form of degree i, or is identically zero. Then

$$f(X,Y) = F(X,Y,1) = \sum_{i=0}^{n} F_i(X,Y),$$

which has degree m, where $m \leq n$ is the largest index for which F_m is not identically 0. Alternatively, $n - m \geq 0$ is the power of Z dividing $F(X, Y, Z)$.

The points "at infinity" on $\mathcal{C}$ are those of the form $[x : y : 0]$, lying on the line $Z = 0$. We have $F(X, Y, 0) = F_n(X, Y)$, which is either a homogeneous binary form of degree n (if $m = n$), or identically 0 if ($m < n$). In the latter case, Z is a factor of F and the whole line at infinity is a component of $\mathcal{C}$. Otherwise, $\mathcal{C}$ has at most n points $[x : y : 0]$ at infinity, where $[x : y]$ is a root of the binary form $F_n(X, Y)$ in $\mathbb{P}^1$. We sometimes refer to these extra points in $\mathcal{C} \setminus \mathcal{C}_0$ as the points at infinity on the affine curve $\mathcal{C}_0$, though of course they are not literally points on $\mathcal{C}_0$.

Examples: 1. $\mathcal{C} : F(X, Y, Z) = X^2 - Y^2 - Z^2$, with dehomogenised polynomial $f(X, Y) = X^2 - Y^2 - 1$. The affine part $\mathcal{C}_0$ of $\mathcal{C}$ consists of the points $[x : y : 1]$ satisfying $x^2 - y^2 = 1$, which is an affine hyperbola. The points at infinity are the points $[x : y : 0]$ satisfying $x^2 - y^2 = 0$; there are two of these, namely $[\pm 1 : 1 : 0]$.

2. $\mathcal{C} : F(X, Y, Z) = XYZ$. This is the union of the three lines $X = 0$, $Y = 0$ and $Z = 0$. The affine part $\mathcal{C}_0$ is defined by $f(X, Y) = XY$, which is the union of the two affine lines $X = 0$ and $Y = 0$; the whole line at infinity ($Z = 0$) is contained in $\mathcal{C}$, and $\deg(f) < \deg(F)$.

Now we go in the other direction, starting with an affine curve $\mathcal{C}_0 = \mathcal{C}_f$ defined by a polynomial $f(X, Y) \in K[X, Y]$ of degree n. Write $f(X, Y) = \sum_{i=0}^{n} \sum_{j=0}^{i} a_{ij} X^j Y^{i-j}$, where not all the a_{nj} are zero. The *homogenisation* of $f(X, Y)$ is

$$F(X, Y, Z) = Z^n f(X/Z, Y/Z) = \sum_{i=0}^{n} \sum_{j=0}^{i} a_{ij} X^j Y^{i-j} Z^{n-i}.$$

Note that $F(X, Y, Z)$ is homogeneous of degree n, and does not have Z as a factor. Also, $F(X, Y, 1) = f(X, Y)$, so if we first homogenise and then dehomogenise, then we recover the original polynomial f. In the other direction, first dehomogenising a form $F(X, Y, Z)$ and then homogenising the result, we only recover the original form $F(X, Y, Z)$ when it does not have Z as a factor.

Definition 16. Let $\mathcal{C}_0$ be an affine plane curve defined by a polynomial $f(X, Y) \in K[X, Y]$. The *projective completion* or *projective closure* of $\mathcal{C}_0$ is the projective plane curve $\mathcal{C} = \overline{\mathcal{C}_0}$, defined by the homogenisation $F(X, Y, Z)$ of $f(X, Y)$.

The points on $\mathcal{C}(L)$ consist of the affine points on $\mathcal{C}_0(L)$ together with a finite set of points at infinity. Firstly (as above), $f(x, y) = F(x, y, 1)$, so $(x, y) \in \mathcal{C}_0(L) \iff [x : y : 1] \in \mathcal{C}(L)$. Secondly, $F(X, Y, 0) = F_n(X, Y) = \sum_{j=0}^{n} a_{nj} X^j Y^{n-j}$, which is a (nonzero) form of degree n. It has exactly n roots, if we count multiplicities and enlarge the field if necessary. These are the "points at infinity" of the original curve $\mathcal{C}_0$.

To summarise:

Proposition 3.7. *1. To every affine plane curve $\mathcal{C}_0$ of degree n, defined by an equation $f(X, Y) = 0$, there is an associated projective plane curve $\mathcal{C} = \overline{\mathcal{C}_0}$, defined by the homogenised equation $Z^n f(X/Z, Y/Z) = 0$. Over any extension field L, $\mathcal{C}(L)$ consists of the union of $\mathcal{C}_0(L)$ together with a finite set of at most n points "at infinity".*

2. To every plane projective curve $\mathcal{C}$ of degree n, defined by a homogeneous equation $F(X, Y, Z) = 0$, the affine points (with $Z \neq 0$) of $\mathcal{C}$ form an affine curve $\mathcal{C}_0$, with equation $f(X, Y) = F(X, Y, 1) = 0$, of degree $m \leq n$. If the line $Z = 0$ is not a component of $\mathcal{C}$, we have $m = n$, and then $\mathcal{C}(L) \setminus \mathcal{C}_0(L)$ consists of a finite set of at most n points.

3. We have $(\overline{\mathcal{C}_0})_0 = \mathcal{C}_0$ for every affine curve $\mathcal{C}_0$, and $\overline{(\mathcal{C}_0)} = \mathcal{C}$ for every projective curve $\mathcal{C}$ not containing $Z = 0$ as a component.

4. The field of definition of $\overline{\mathcal{C}_0}$ is the same as that of $\mathcal{C}_0$; in particular, if one is rational then so is the other.

Examples: 1. The affine line $X-Y+3=0$ has projective completion $X-Y+3Z=0$. There is one point at infinity, $[1:1:0]$. Note that all affine lines of slope 1, which have affine equations of the form $X-Y+c=0$, all have projective completions with this same point at infinity.

2. The affine circle $X^2+Y^2-1=0$ has projective completion $X^2+Y^2-Z^2=0$, and points at infinity $[\pm i:1:0]$. Similarly, every real affine ellipse has two distinct non-real points at infinity.

3. The affine hyperbola $X^2-Y^2-1=0$ has projective completion $X^2-Y^2-Z^2=0$, and points at infinity $[\pm 1:1:0]$. Similarly, every real affine hyperbola has two distinct real points at infinity.

4. The affine parabola $X^2-Y=0$ has projective completion $\mathcal{C}:\ X^2-YZ=0$, and one point $[0:1:0]$ at infinity. The intersection of $\mathcal{C}$ and the line $Z=0$ has multiplicity two at this point (since substituting $Z=0$ gives $X^2=0$, with a double root). Similarly, every real affine parabola has one real "double" point at infinity.

5. The affine Bachet-Mordell curve $Y^2=X^3+c$ has projective completion $\mathcal{C}:\ Y^2Z=X^3+cZ^3$, and one point $[0:1:0]$ at infinity. The intersection of $\mathcal{C}$ and the line $Z=0$ has multiplicity three at this point.

3.5 Lines in the projective plane

Lines in $\mathbb{P}^2$ have equations of the form $aX+bY+cZ=0$, where $(a,b,c)\neq(0,0,0)$. If $(a,b)=(0,0)$ this is the line $Z=0$; otherwise it is the projective completion of the affine line $aX+bY+c=0$ with slope $-a/b$, and unique point $[-b:a:0]$ at infinity.

The three coefficients a,b,c of a plane projective line determine a unique point $[a:b:c]\in\mathbb{P}^2$, since they are not all zero and scaling by a nonzero constant factor gives the same line. Conversely, each point $[a:b:c]\in\mathbb{P}^2$ determines uniquely the equation of a line $aX+bY+cZ=0$. So the space of all lines in $\mathbb{P}^2$ is parametrised by another "dual" copy of $\mathbb{P}^2$. This duality between points and lines in $\mathbb{P}^2$ gives a remarkable symmetry between results about points and results about lines in the projective plane.

Proposition 3.8. *1. Let P_1, P_2 be distinct points in $\mathbb{P}^2(K)$. Then there is a unique line $\mathcal{L}$ in $\mathbb{P}^2(K)$ passing through both P_1 and P_2.*

2. Let $\mathcal{L}_1$, $\mathcal{L}_2$ be distinct lines in $\mathbb{P}^2(K)$. Then there is a unique point P in $\mathbb{P}^2(K)$ lying on both $\mathcal{L}_1$ and $\mathcal{L}_2$.

Definition 17. 1. Three points P_i in $\mathbb{P}^2(K)$ $(i=1,2,3)$ are *collinear* if there is a line $\mathcal{L}$ which passes through all three of them.

2. Three lines $\mathcal{L}_i$ in $\mathbb{P}^2(K)$ $(i=1,2,3)$ are *concurrent* if there is a point P which lies on all three of them.

Proposition 3.9. *For $i=1,2,3$, let $\mathcal{L}_i$ be the line $a_iX+b_iY+c_iZ=0$ in $\mathbb{P}^2(K)$, and let P_i be the point $[a_i:b_i:c_i]$ in $\mathbb{P}^2(K)$. The following are equivalent:*

1. $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ are concurrent;

2. P_1, P_2, P_3 are collinear;

3. $\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = 0.$

Lines in $\mathbb{P}^2$ may be parametrised as follows: the parameter is a point in $\mathbb{P}^1$.

Proposition 3.10. *Let $\mathcal{L}$ be a line in $\mathbb{P}^2(K)$, and $P_1,P_2\in\mathcal{L}(K)$ distinct points on $\mathcal{L}$. Fix homogeneous coordinates $P_i=[x_i:y_i:z_i]$. The map given by*

$$[t:u]\mapsto tP_1+uP_2=[tx_1+ux_2:ty_1+uy_2:tz_1+uz_2]$$

is a bijection from $\mathbb{P}^1(K)$ to $\mathcal{L}(K)$.

In practice, note that at least two of $[-b : a : 0]$, $[-c : 0 : a]$, $[0 : -c : b]$ give valid points on the line $aX + bY + cZ = 0$.

Example: The line $\mathcal{L} :\ 3X - 2Y + 5Z = 0$ contains the points $[2 : 3 : 0]$ and $[0 : 5 : 2]$. Hence it may be parametrised by

$$[t : u] \mapsto [2t : 3t + 5u : 2u].$$

Note that the affine part $\mathcal{L}_0$ has equation $3X - 2Y + 5 = 0$, and is parametrised by $t \mapsto (t, (3t + 5)/2)$. Homogenising this by replacing t by t/u gives $(t/u, (3(t/u) + 5)/2) = [t/u : (3(t/u) + 5)/2 : 1] = [2t : 3t + 5u : 2u]$, where we identified $\mathbb{A}^2(K)$ with a subset of $\mathbb{P}^2(K)$ in the usual way. So we obtain a projective parametrisation by homogenising an affine parametrisation, homogenising the parameter from $K = \mathbb{A}^1(K)$ to $\mathbb{P}^1(K)$ as well as the image point.

There is a dual result to the previous one, parametrising all lines through a given point. This is sometimes called a *pencil* of lines.

Proposition 3.11. *Let P be a point in $\mathbb{P}^2(K)$, and $\mathcal{L}_1, \mathcal{L}_2 \in \mathcal{L}(K)$ distinct lines through P. Fix equations for $\mathcal{L}_i :\ = a_iX + b_iY + c_iZ = 0$. The map given by*

$$[t : u] \mapsto t\mathcal{L}_1 + u\mathcal{L}_2 :\ (ta_1 + ua_2)X + (tb_1 + ub_2)Y + (tc_1 + uc_2)Z = 0$$

is a bijection from $\mathbb{P}^1(K)$ to the set of all lines in $\mathbb{P}^2(K)$ through P.

3.6 Projective transformations

A non-singular linear map $T : K^{n+1} \to K^{n+1}$ induces (or determines) a well-defined map $\mathbb{P}^n(K) \to \mathbb{P}^n(K)$, since

1. $T(v) = 0 \iff v = 0$, for $v \in K^{n+1}$;
2. $T(\lambda v) = \lambda T(v)$, for $v \in K^{n+1}$ and $\lambda \in K^*$, so $v \sim v' \iff T(v) \sim T(v')$.

We use the same letter T to denote the induced map on $\mathbb{P}^n(K)$.

Definition 18. A *projective transformation* or *projective change of coordinates* of $\mathbb{P}^n(K)$ is a map $T : \mathbb{P}^n(K) \to \mathbb{P}^n(K)$ induced by a nonsingular linear transformation T on K^{n+1} as above.

The set of all projective transformations of $P^n(K)$ forms a group, denoted $\mathrm{PGL}(n+1, K)$.

Projective transformations are invertible, since they are induced by invertible linear maps. In terms of projective coordinates on $P^n(K)$ they are given by $(n + 1) \times (n + 1)$ nonsingular matrices, and two matrices determine the same projective transformation if (and only if) one is a nonzero scalar multiple of the other. After fixing a basis for K^{n+1}, we may identify $\mathrm{PGL}(n + 1, K)$ with the quotient group $\mathrm{GL}(n + 1, K)/H$, where $\mathrm{GL}(n + 1, K)$ is the group of invertible $(n + 1) \times (n + 1)$ matrices with entries in K, and H is the subgroup of scalar matrices.

Just as with matrices representing linear transformations of a vector space, we may view an element $T \in \mathrm{PGL}(n+1, K)$ as either transforming points in $P^n(K)$ into new points, or as giving new coordinates for the same point. Often, choosing coordinates suitable can simplify the algebra in a problem.

When using matrix notation for projective transformations, we will write the homogeneous coordinates for points as column vectors with the matrix multiplying on the left. In the case $n = 2$, each projective transformation is represented by a 3×3 matrix, determined up to non-zero scalar multiple. The three columns of the matrix $T \in \mathrm{GL}(3, K)$ give the images under T of the “basic triangle” of points $[1 : 0 : 0]$, $[0 : 1 : 0]$, $[0 : 0 : 1]$.

Proposition 3.12. *Let P_i for $i = 1, 2, 3$ be non-collinear points in $\mathbb{P}^2(K)$. Then there exists $T \in \mathrm{PGL}(3, K)$ such that $T(P_1) = [1 : 0 : 0]$, $T(P_2) = [0 : 1 : 0]$, $T(P_3) = [0 : 0 : 1]$.*

More generally, if Q_i (for $i = 1, 2, 3$) is another set of three non-collinear points, then there exists $T \in \mathrm{PGL}(3, K)$ such that $T(P_i) = Q_i$ for $i = 1, 2, 3$.

Dually, we can always find a projective transformation taking any three lines which are not concurrent to the lines $X = 0$, $Y = 0$, $Z = 0$ (exercise).

3.7 Smooth points and tangents

Lemma 3.13 (Euler's identity). *Let $F(X, Y, Z) \in K[X, Y, Z]$ be homogeneous of degree n. Then*

$$XF_X + YF_Y + ZF_Z = nF.$$

Definition 19. Let $\mathcal{C}_F$ be a plane projective curve defined by the form $F \in K[X, Y, Z]$. Let $P \in \mathcal{C}_F(L)$, where L is a field containing K. We say that P is a *singular point* of $\mathcal{C}_F$ if

$$F_X(P) = F_Y(P) = F_Z(P) = 0.$$

Otherwise, P is a *non-singular* or *smooth* point of $\mathcal{C}_F$.

The curve $\mathcal{C}_F$ itself is called *non-singular*, or *smooth*, if all its points are smooth (over all extensions L of K).

Euler's identity shows that the four conditions $F(P) = 0$, $F_X(P) = 0$, $F_Y(P) = 0$, $F_Z(P) = 0$ are not independent. The first always follows from the last three; unless $Z(P) = 0$, the condition $F_Z(P) = 0$ follows from the other three; and so on.

Definition 20. Let P be a smooth point on the projective curve $\mathcal{C}_F$. The *tangent to $\mathcal{C}_F$ at P* is the line

$$F_X(P)X + F_Y(P)Y + F_Z(P)Z = 0.$$

It is an exercise to show that a smooth point P on an affine curve $\mathcal{C}_0$ is still smooth when considered on the projective completion $\mathcal{C}$ of $\mathcal{C}_0$, and the tangent line to $\mathcal{C}$ at P is the projective completion of the tangent line to $\mathcal{C}_0$ at P.

It will be useful later to have a finer classification of singular points.

Definition 21. Let $\mathcal{C}_F$ be a plane projective curve defined by the form $F \in K[X, Y, Z]$. Let $P \in \mathbb{P}^2(L)$, where L is a field containing K. We say that P has *multiplicity* m on $\mathcal{C}_F$ if m is the least integer such that not all the mth order partial derivatives of F vanish at P. Notation: $m(P, F)$, or $m(P, \mathcal{C})$ if $\mathcal{C} = \mathcal{C}_F$.

For example, $m(P, \mathcal{C}) = 0$ iff $P \notin \mathcal{C}$; $m(P, \mathcal{C}) = 1$ iff P is a smooth point of $\mathcal{C}$; $m(P, \mathcal{C}) > 1$ iff P is a singular point on $\mathcal{C}$. We call P a *double point* of $\mathcal{C}$ if $m(P, \mathcal{C}) = 2$, a *triple point* if $m(P, \mathcal{C}) = 3$, and so on.

Consider the point $P = [0 : 0 : 1]$. Recall the decomposition (13) of a form $F \in K[X, Y, Z]$, where $F_i(X, Y)$ is either identically zero or is a binary form of degree i. Then the constant F_0 is zero iff $P \in \mathcal{C}_F$, and P is a smooth point of $\mathcal{C}_F$ iff $F_0 = 0$ and $F_1 \neq 0$. In the latter case, F_1 is the equation of the tangent to $\mathcal{C}_F$ at P. More generally, $m(P, F)$ is the least integer m for which F_m is not identically zero. The binary form F_m has m linear factors, counting multiplicities, after extending the field if necessary. The lines determined by these linear factors are called the *tangent lines* to $\mathcal{C}$ at P. (When $m = 1$ this agrees with the earlier definition of tangent line.)

A double point is called a *node* of $\mathcal{C}$ if $\mathcal{C}$ has two distinct tangents at P, and a *cusp* if it has a double tangent. For example the real cubic curves $Y^2 = X^2(X + 1)$ and $Y^2 = X^3$ have a node and a cusp at $(0, 0)$, respectively. The real cubic curve $Y^2 = X^2(X - 1)$ has a double point at $(0, 0)$ but no real tangents there.

A tedious exercise in polynomial algebra shows that these notions of smoothness, multiplicity and tangents are all invariant under projective changes of coordinates.

3.8 Intersecting lines and curves: Bezout's Theorem

Let $\mathcal{L}$ be a line in $\mathbb{P}^2(K)$ parametrised by the function

$$\varphi : [t : u] \mapsto [L_1(t,u) : L_2(t,u) : L_3(t,u)]$$

from $\mathbb{P}^1(K)$ onto $\mathcal{L}(K)$. Here the $L_i(T,U) \in K[T,U]$ are either linear forms or identically zero, and at most one is zero.

Let $\mathcal{C} = \mathcal{C}_F$ be a plane projective curve of degree n, defined by the form $F(X,Y,Z)$ of degree n. Assume that $\mathcal{L} \not\subseteq \mathcal{C}$. The points of intersection of $\mathcal{L}$ and $\mathcal{C}$ are determined by the roots of

$$G(T,U) = F(L_1(T,U), L_2(T,U), L_3(T,U)),$$

which is a binary form (in T, U) of degree n (not identically zero since $\mathcal{L} \not\subseteq \mathcal{C}$).

Let the roots of G be $[t_j : u_j] \in \mathbb{P}^2(L)$ for $1 \leq j \leq n$, counted with multiplicity, where L may be an extension of K. Then the intersection points of $\mathcal{L}$ and $\mathcal{C}$ are the points $P_j = \varphi([t_j : u_j]) \in \mathbb{P}^2(L)$. There are exactly n of them, counted with multiplicity.

Definition 22. In the above notation, the *multiplicity of $P = P_j$ on $\mathcal{L} \cap \mathcal{C}$* is the multiplicity of $[t_j : u_j]$ as a root of the binary form $G(T,U)$. Notation: $I(P, \mathcal{L} \cap \mathcal{C})$. We set $I(P, \mathcal{L} \cap \mathcal{C}) = 0$ if P is not one of the P_j.

Proposition 3.14 (Bezout's Theorem (special case)). *Let $\mathcal{C}$ be a plane projective curve of degree n and $\mathcal{L}$ a line in $\mathbb{P}^2$. Suppose that $\mathcal{L} \not\subseteq \mathcal{C}$. Then*

$$\sum_{P \in \mathbb{P}^2(L)} I(P, \mathcal{L} \cap \mathcal{C}) = n.$$

Here the sum is over all points in $\mathbb{P}^2(L)$ for all extensions L of K, but only a finite number of terms are non-zero.

The general version of Bezout's Theorem (over $\mathbb{C}$) states that if $\mathcal{C}_1$ and $\mathcal{C}_2$ are curves in $\mathbb{P}^2(\mathbb{C})$ of degrees m and n, with no common component, then $\mathcal{C}_1 \cap \mathcal{C}_2$ consists of exactly mn points, counted with multiplicity; but defining the general intersection multiplicity $I(P, \mathcal{C}_1 \cap \mathcal{C}_2)$ is harder.

Examples: Distinct lines intersect in one point with intersection multiplicity 1.

If $\mathcal{C}$ is an irreducible conic then every line intersects $\mathcal{C}$ either in two distinct points (possibly over an extension field) or in one point with intersection multiplicity 2.

If $\mathcal{C}$ is an irreducible cubic and $\mathcal{L}$ a line, then $\mathcal{C} \cap \mathcal{L}$ consists of either three distinct points with multiplicity 1 each, or two distinct points, one with multiplicity 2, or in a single point with multiplicity 3.

Proposition 3.15. *Let $\mathcal{C}$ be an irreducible projective plane curve, P a smooth point on $\mathcal{C}$, and $\mathcal{L}$ a line through P. Then $I(P, \mathcal{L} \cap \mathcal{C}) \geq 1$, with equality unless $\mathcal{L}$ is the tangent to $\mathcal{C}$ at P.*

More generally, if $m = m(P, \mathcal{C}) \geq 1$, then $I(P, \mathcal{L} \cap \mathcal{C}) \geq m$, with equality unless $\mathcal{L}$ is one of the tangents to $\mathcal{C}$ at P.

Corollary 3.16. *An irreducible conic is smooth. An irreducible cubic is either smooth, or has a unique singular point which is a double point.*

3.9 Conics in the projective plane

A conic $\mathcal{C}$ in $\mathbb{P}^2$ is a curve of degree 2, so has an equation of the form

$$F(X,Y,Z) = aX^2 + bXY + cY^2 + dXZ + eYZ + fZ^2 = 0,$$

where the coefficients a, b, c, d, e, f are not all zero and are determined up to a nonzero scalar multiple. Thus, $\mathcal{C}$ is determined by the point $[a : b : c : d : e : f] \in \mathbb{P}^5$.

We do not require $(a, b, c) \neq (0, 0, 0)$; but if $a = b = c = 0$ then $F = Z(dX + eY + fZ)$, so $\mathcal{C}$ is reducible, consisting of the union of the line at infinity $Z = 0$ and a second line.

We have

$$\begin{aligned} F_X &= 2aX + bY + dZ, \\ F_Y &= bX + 2cY + eZ, \\ F_Z &= dX + eY + 2fZ; \end{aligned}$$

recall from Euler's identity that $F(P) = 0$ follows from $F_X(P) = F_Y(P) = F_Z(P) = 0$. Hence the singular points on $\mathcal{C}$ are those $[x : y; z]$ which satisfy the equation

$$\begin{pmatrix} 2a & b & d \\ b & 2c & e \\ d & e & 2f \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

So a necessary and sufficient condition for nonsingularity is that $\det(M) \neq 0$, where M is the above matrix. (Note that the solution $x = y = z = 0$ is irrelevant as it does not determine a point in $\mathbb{P}^2$.)

Proposition 3.17. *Let $\mathcal{C}$ be a conic with coefficient matrix $M = \begin{pmatrix} 2a & b & d \\ b & 2c & e \\ d & e & 2f \end{pmatrix}$. Then*

1. $\mathcal{C}$ is smooth iff $\det(M) \neq 0$ iff $\operatorname{rank}(M) = 3$;

2. $\mathcal{C}$ has a unique singular point P, and is the union of distinct lines through P, iff $\operatorname{rank}(M) = 2$;

3. $\mathcal{C}$ is a double line, with every point singular, iff $\operatorname{rank}(M) = 1$.

Note that the equation for $\mathcal{C}$ may be written in the form

$$\begin{pmatrix} X & Y & Z \end{pmatrix} \begin{pmatrix} 2a & b & d \\ b & 2c & e \\ d & e & 2f \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

If we make a projective change of coordinates, this amounts to substituting $(X\,Y\,Z)^t = T(X'\,Y'\,Z')^t$, where T is a nonsingular 3×3 matrix. In the new variables, the equation of $\mathcal{C}$ has equation with coefficient matrix $M' = T^t M T$. For suitable choice of T, this will be diagonal, by the following fact from algebra:

Lemma 3.18. *Let M be an $n \times n$ symmetric matrix over a field K (with $\operatorname{char}(K) \neq 2$). Then there exists $T \in \operatorname{GL}(n, K)$ such that $T^t M T$ is diagonal.*

If $\mathcal{C}$ is nonsingular then $\det(M) \neq 0$, so the diagonal entries of M' are nonzero. By further scaling and permuting the variables and scaling the matrix, we can simplify the equation still further, depending on the field. Note that scaling the variables multiplies the diagonal entries of M' by squares.

Proposition 3.19. *Let $\mathcal{C}$ be a nonsingular plane conic over the field K, with $\operatorname{char}(K) \neq 2$. After a suitable projective transformation, $\mathcal{C}$ has one of the following standard forms.*

$K = \mathbb{C}$: $X^2 + Y^2 = Z^2$.

$K = \mathbb{R}$: either $X^2 + Y^2 = Z^2$ or $X^2 + Y^2 + Z^2 = 0$; the latter has no real points.

$K = \mathbb{Q}$: $aX^2 + bY^2 + cZ^2 = 0$, where a, b, c are nonzero square-free pairwise coprime integers.

So in $\mathbb{P}^2(\mathbb{C})$ all nonsingular conics are essentially the same; over $\mathbb{R}$ there are just two types, one of which has no real points. In $\mathbb{P}^2(\mathbb{R})$ we eliminate the affine distinction between real ellipses, hyperbolas and parabolas, which comes from the three different ways in which the line at infinity may intersect a conic in $\mathbb{P}^2(\mathbb{R})$.

Finally we show how to parametrise all points on a nonsingular conic, given one point. The idea is the same as in the affine case, but works out more simply since there are no exceptional points or parameter values.

Proposition 3.20. *Let $\mathcal{C}$ be a nonsingular conic defined over a field K, and $P \in \mathcal{C}(K)$. Then there are binary quadratic forms $Q_i(T,U) \in K[T,U]$ for $i = 1,2,3$, without common factors, such that the map*

$$\theta : [t : u] \mapsto [Q_1(t,u) : Q_2(t,u) : Q_3(t,u)]$$

is a bijection from $\mathbb{P}^1(K)$ to $\mathcal{C}(K)$.

Sketch proof: details in lectures. Shift coordinates so that $P = [0 : 0 : 1]$. The equation of $\mathcal{C}$ then has no Z^2 term, so has the form

$$F(X,Y,Z) = Q(X,Y) + L(X,Y)Z = 0$$

where Q is a nonzero quadratic form and L is a nonzero linear form, and L is not a factor of Q. The map θ is defined with

$$Q_1(T,U) = -TL(T,U), \qquad Q_2(T,U) = -UL(T,U), \qquad Q_3(T,U) = Q(T,U);$$

the inverse map takes $[x : y : z] \in \mathcal{C}(K) \mapsto [x : y]$, except that $\theta^{-1}(P) = [-b : a]$ where $L(X,Y) = aX + bY$ (which is the tangent to $\mathcal{C}$ at P), so $L(-b,a) = 0$. □

Example: Let $\mathcal{C}$ be the rational conic $3X^2 + 6Y^2 = 4Z^2$, with point $P = [2 : 2 : 3] \in \mathcal{C}(\mathbb{Q})$. To shift P to $[0 : 0 : 1]$ we use the transformation

$$\begin{aligned} X &= 3X_1 + 2Z_1 \\ Y &= 3Y_1 + 2Z_1 \\ Z &= 3Z_1. \end{aligned}$$

Substituting in $F(X,Y,Z) = 3X^2 + 6Y^2 - 4Z^2$ gives the equation (after simplifying)

$$3(X_1^2 + 2Y_1^2) + 4(X_1 + 2Y_1)Z_1 = 0,$$

so $Q(X_1,Y_1) = 3(X_1^2 + 2Y_1^2)$ and $L(X_1,Y_1) = 4(X_1 + 2Y_1)$. Hence the parametrisation is given by

$$[X_1 : Y_1 : Z_1] = [-4T(T + 2U) : -4U(T + 2U) : 3(T^2 + 2U^2)].$$

Substituting back to the original coordinates, we have the parametrisation

$$\begin{aligned}[X : Y : Z] &= [3X_1 + 2Z_1 : 3Y_1 + 2Z_1 : 3Z_1] \\ &= [2(-T^2 - 4TU + 2U^2) : 2(T^2 - 2TU - 2U^2) : 3(T^2 + 2U^2)].\end{aligned}$$

In other words, the map

$$[t : u] \mapsto [2(-t^2 - 4tu + 2u^2) : 2(t^2 - 2tu - 2u^2) : 3(t^2 + 2u^2)]$$

is a bijection $\mathbb{P}^1(\mathbb{Q}) \to \mathcal{C}(\mathbb{Q})$.

The inverse map is given by setting $[t : u] = [x_1 : y_1]$ away from P, and $[t : u] = [-2 : 1]$ at P (to make $L(t,u) = 0$). To express this in terms of the original coordinates, we have to invert the change of coordinates. This gives (after scaling the homogeneous coordinates)

$$[X_1 : Y_1 : Z_1] = [3X - 2Z : 3Y - 2Z : 3Z],$$

so $[t : u] = [x_1 : y_1] = [3x - 2z : 3y - 2z]$ maps to $Q = [x : y : z] \in \mathcal{C}(\mathbb{Q})$ when $Q \neq P$, while (from above) $[t : u] = [-2 : 1]$ maps to $P = [2 : 2 : 3]$.

The corresponding affine parametrisation is

$$t \mapsto \left(\frac{2(-t^2 - 4t + 2)}{3(t^2 + 2)}, \frac{2(t^2 - 2t - 2)}{3(t^2 + 2)} \right),$$

which omits the point $\varphi([1 : 0]) = [-2 : 2 : 3] = (-2/3, 2/3)$ as well as the images of $t = \pm\sqrt{-2}$, which are the points $[\pm\sqrt{-2} : 1 : 0]$ at infinity.

4 Elliptic Curves

4.1 Definition and examples

Definition 23. An *elliptic curve* defined over the field K is a nonsingular plane cubic curve $\mathcal{E}$ defined over K which has at least one K-rational point.

The last condition, that the set $\mathcal{E}(K)$ be not empty, is an important part of the definition when K is not algebraically closed. For $K = \mathbb{Q}$, it can be very hard to determine whether a nonsingular cubic has any rational points: there is no known method which is guaranteed to answer this question in all cases! For example, the Selmer curve $3X^3 + 4Y^3 = 5Z^3$ is nonsingular but has no rational points, so is *not* an elliptic curve. Over a finite field, it is true that every nonsingular cubic has points and so is an elliptic curve, but this is a hard theorem. Over $\mathbb{C}$, obviously every curve has infinitely many points.

Singular cubics: If $F(X, Y, Z)$ is reducible (possibly over an extension field) then $\mathcal{C}_F$ is singular; for if $F = LG$ with $\deg(L) = 1$ and $\deg(G) = 2$, then any point on the intersection of L and G is singular. There are irreducible singular cubics: such a curve can only have one singular point, which must be a double point. To see this (see Corollary 3.16): if P and Q are both singular on a cubic $\mathcal{C}$, then the line $\mathcal{L}_{PQ}$ intersects $\mathcal{C}$ with total multiplicity at least 4, hence by Bezout's Theorem (Proposition 3.14) must be a component of $\mathcal{C}$, so that $\mathcal{C}$ is reducible. Similarly, if P has multiplicity 3 on $\mathcal{C}$ then for any other point Q on $\mathcal{C}$ the line $\mathcal{L}_{PQ}$ must be a component of $\mathcal{C}$ (so $\mathcal{C}$ is a union of three lines through P in this case).

Examples of irreducible singular cubics are the curves $Y^2 = X^3 + X^2$ and $Y^2 = X^3$, which have double points at $(0, 0)$ (a node and a cusp respectively).

Examples of elliptic curves: 1. The curve $aX^3+bY^3+cZ^3 = 0$ is nonsingular provided $abc \neq 0$, so is an elliptic curve provided that it does have at least one point. If $a = b = 1$ then $[1 : -1 : 0]$ is such a point, for example.

2. Let $f(X) \in K[X]$ be a polynomial of degree 3 with no repeated roots. Then the affine curve $Y^2 = f(X)$ is smooth (see Example (3) on page 9), and its projective completion has a single point at infinity $[0 : 1 : 0]$ which is also nonsingular, hence is an elliptic curve. Most of the elliptic curves we will study will have equations of this form, and we will see later how to transform a given cubic curve (with a point) into this form.

4.2 The group law

Let $\mathcal{E}$ be an elliptic curve defined over the field K, with point $\mathcal{O} \in \mathcal{E}(K)$. The set $\mathcal{E}(K)$ of K-rational points on $\mathcal{E}$ is therefore nonempty, as it contains $\mathcal{O}$. We will see that it is possible to define a group structure on the set $\mathcal{E}(K)$, in which the given point $\mathcal{O}$ is the identity. The construction is originally geometric, but can also be described purely algebraically, and works over arbitrary fields (including finite fields).

Let $P, Q \in \mathcal{E}(K)$. They determine a unique line $\mathcal{L}_{PQ}$ as follows: if $P \neq Q$ then $\mathcal{L}_{PQ}$ is the unique line through P and Q, while if $P = Q$ then $\mathcal{L}_{PP}$ is the tangent to $\mathcal{E}$ at P (which exists since $\mathcal{E}$ is smooth). In either case, the line $\mathcal{L}_{PQ}$ is defined over K, and intersects $\mathcal{E}$ at both P and Q. By Bezout's Theorem, since $\deg(\mathcal{E}) = 3$, $\mathcal{L}_{PQ}$ intersects $\mathcal{E}$ in one more point, say R. Here we count multiplicities, so R may equal either P or Q (or both). The point is that we have a line, so by Bezout's Theorem it intersects $\mathcal{E}$ in exactly three points counting multiplicities, and these points are P, Q and R.

Definition 24. Let P, Q, R be points on the elliptic curve $\mathcal{E}$. We write $P * Q = R$ if there is a line $\mathcal{L}$ which intersects $\mathcal{E}$ in the three points P, Q, R counting multiplicity. We say that P, Q, R are *collinear on* $\mathcal{E}$ if this is the case.

Note that when P, Q, R are distinct then there is no distinction between being collinear on $\mathcal{E}$ and being collinear in the usual sense, but this is not true when two or more of the points are equal.

Although we have defined the operation $*$ as a binary operation, in fact the relation $P * Q = R$ is entirely symmetric.

Proposition 4.1. *1. $P * Q = Q * P$;*

2. $P * Q = R \implies P * R = Q$ and $Q * R = P$;

3. If $P, Q \in \mathcal{E}(K)$ then $P * Q \in \mathcal{E}(K)$ also.

So we have a commutative binary operation defined on the set $\mathcal{E}(K)$. This is *not* a group operation, as there is no identity. To get a group operation (or group law) we proceed as follows.

Definition 25. Let $\mathcal{E}$ be an elliptic curve defined over the field K, with base point $\mathcal{O} \in \mathcal{E}(K)$. We define the binary operation $\oplus$ on $E(K)$ by

$$P \oplus Q = (P * Q) * \mathcal{O}.$$

Note that this definition *does* depend on the particular base point $\mathcal{O}$, and that changing the base point will give a *different* operation.

It is immediate from the preceding Proposition that $P \oplus Q \in \mathcal{E}(K)$ for all $P, Q \in \mathcal{E}(K)$. With the operation $\oplus$, $\mathcal{E}(K)$ now does form a group:

Theorem 4.2. *Let $\mathcal{E}$ be an elliptic curve defined over the field K, with base point $\mathcal{O} \in \mathcal{E}(K)$. The set $\mathcal{E}(K)$ forms a group under the operation $\oplus$, with identity $\mathcal{O}$, where the inverse of $P \in \mathcal{E}(K)$ is*

$$\ominus P = P * (\mathcal{O} * \mathcal{O}).$$

The hard part is to prove that $\oplus$ is associative. This relies on the following fact:

Lemma 4.3. *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be cubics which intersect in the nine points P_j, ($1 \leq j \leq 9$). Then every cubic which passes through P_j for $1 \leq j \leq 8$ also passes through P_9.*

Definition 26. Let P be a smooth point on a plane projective curve $\mathcal{C}$, with tangent line $\mathcal{L}$. We say that P is a *flex* on $\mathcal{C}$ if $I(P, \mathcal{L} \cap \mathcal{C}) \geq 3$.

We always have $I(P, \mathcal{L} \cap \mathcal{C}) \geq 2$ since $\mathcal{L}$ is the tangent at P (see Proposition 3.15). For a flex on a cubic curve we have $I(P, \mathcal{L} \cap \mathcal{C}) = 3$. On an elliptic curve $\mathcal{E}$, the point P is a flex if the tangent to $\mathcal{E}$ intersects $\mathcal{E}$ at P with multiplicity 3 and nowhere else; equivalently, if $P * P = P$.

The group law becomes simpler if we use a flex as base-point $\mathcal{O}$. Now in general, $\mathcal{E}$ has exactly nine flexes (see Exercise 3-7 for a special case of this), but they need not be rational. We will see later how to handle the case where $\mathcal{E}(K)$ has no flexes. When $\mathcal{O}$ is a flex, the negation formula becomes

$$\ominus P = P * \mathcal{O}$$

and the addition law becomes

$$P \oplus Q = (P * Q) * \mathcal{O} = \ominus(P * Q).$$

Proposition 4.4. *Let $\mathcal{E}$ be an elliptic curve defined over the field K with base-point $\mathcal{O} \in \mathcal{E}(K)$ which is a flex. Then*

$$P \oplus Q \oplus R = \mathcal{O} \iff P, Q, R \text{ are collinear on } \mathcal{E}.$$

4.3 The structure of $\mathcal{E}(K)$

When K is a finite field, then certainly $\mathcal{E}(K)$ is a finite group. We will discuss this case, which is the one used for cryptographic applications, in more detail below.

When $K = \mathbb{C}$ or $K = \mathbb{R}$ then $\mathcal{E}(K)$ is an infinite "continuous" group. Over $\mathbb{R}$ elliptic curves are examples of so-called Lie Groups. The group $\mathcal{E}(\mathbb{R})$ is isomorphic either to the circle group $\mathbb{R}/\mathbb{Z}$, or to the product of this with a cyclic group of order two. Over $\mathbb{C}$, the group $\mathcal{E}(\mathbb{C})$ is a torus: a surface shaped like a doughnut. Algebraically, $\mathcal{E}(\mathbb{C}) \cong \mathbb{C}/L$ where L is a "lattice" in $\mathbb{C}$. The complex theory of elliptic curves involves the study of doubly-periodic functions called *elliptic functions*, using complex analysis. In fact, historically, elliptic functions came first, being invented to solve certain questions about ellipses, and elliptic curves are named after them. (Elliptic curves are **not** ellipses!)

When $K = \mathbb{Q}$ the group $\mathcal{E}(\mathbb{Q})$ may be trivial, finite or infinite. It is hard to tell whether $\mathcal{E}(\mathbb{Q})$ is finite or infinite in general, though we will see special cases of this. However, even though $\mathcal{E}(\mathbb{Q})$ may be infinite, it is always *finitely-generated*:

Theorem 4.5 (Mordell's Theorem). *Let $\mathcal{E}$ be an elliptic curve defined over the field $\mathbb{Q}$ of rational numbers. Then the group $\mathcal{E}(\mathbb{Q})$ is finitely-generated. In other words, there exists a finite set of rational points $P_1, P_2, \ldots, P_r \in \mathcal{E}(\mathbb{Q})$ (where $r \geq 0$), such that every rational point P may be expressed as*

$$P = n_1P_1 \oplus n_2P_2 \oplus \cdots \oplus n_rP_r$$

where the $n_j \in \mathbb{Z}$.

It is an unsolved problem to determine whether the number of generators required can be arbitrarily large, though this is suspected to be the case. There exist curves which require at least 24 generators.

When $\mathcal{E}$ is defined over $\mathbb{Q}$, the group $\mathcal{E}(\mathbb{Q})$ is called the *Mordell group* of the curve in Mordell's honour. From the structure theory for finitely-generated abelian groups, we deduce that

$$\mathcal{E}(\mathbb{Q}) \cong T \times \mathbb{Z}^r$$

where $r \geq 0$ and T is a finite group. The integer r appearing here is called the *rank* of $\mathcal{E}$. Determining the rank for a specific elliptic curve can be very difficult, thought there do exist computer programs which can compute r (from an equation for the curve) for many curves. The finite group T is (isomorphic to) the *torsion subgroup* of $\mathcal{E}(\mathbb{Q})$, which is the subgroup of all points of finite order. We will study this in more detail later on.

4.4 Weierstrass equations

The simplest type of equation for an elliptic curve is (in affine form) $Y^2 = g(X)$, where $g(X) \in K[X]$ has degree 3 and no repeated roots. There is a unique point at infinity, namely $\mathcal{O} = [0 : 1 : 0]$, and the line $Z = 0$ is the tangent at $\mathcal{O}$, which is a flex. We will do most of our work with elliptic curves using an equation of this form. This form of equation is called a *Weierstrass equation*.

Given an arbitrary elliptic curve $\mathcal{E}$, we now see how to transform its equation into Weierstrass form. This is easiest when $\mathcal{E}$ has a rational flex; in the general case, we will require a coordinate transformation more general than the projective changes of coordinates we have met so far. This more general transformation is called a *birational transformation*: it is not linear, but defined by polynomials of degree greater than 1.

Case 1: rational flex

Suppose that $\mathcal{E}$ is an elliptic curve defined over a field K, and that $\mathcal{O}$ is a flex in $\mathcal{E}(K)$ with tangent $\mathcal{L}$. After a projective change of coordinates we may assume that $\mathcal{O} = [0 : 1 : 0]$ and that $\mathcal{L}$ is the line $Z = 0$. The change of coordinates will have coefficients in K, so the new equation for $\mathcal{E}$ still has coefficients in K. The general cubic equation has the form

$$F(X,Y,Z) = aX^3 + bX^2Y + cXY^2 + dY^3 + (eX^2 + fXY + gY^2)Z + (hX + iY)Z^2 + jZ^3.$$

Set $\mathcal{E} = \mathcal{C}_F$. The condition that $\mathcal{O} = [0:1:0] \in \mathcal{E}$ implies $d = 0$; that the tangent at $\mathcal{O}$ is Z implies $c = 0$, $g \neq 0$; that $\mathcal{O}$ is a flex implies $b = 0$, $a \neq 0$. We can scale the variables so that $g = 1$ and $a = -1$. Dehomogenizing with respect to Z gives an equation of the form

$$Y^2 + a_1XY + a_3Y = X^3 + a_2X^2 + a_4X + a_6$$

with the $a_i \in K$. (There is no a_5.) This is a *long Weierstrass equation.*

If the characteristic of K is not 2, so that division by 2 is possible, we may complete the square: replace Y by $Y - (a_1X + a_3)/2$ to get an equation of the form

$$Y^2 = X^3 + aX^2 + bX + c$$

with a, b, $c \in K$. This is a *Weierstrass equation.* The condition that this equation defines a smooth curve is that $\Delta \neq 0$, where

$$\Delta = \Delta_{a,b,c} = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2.$$

If the characteristic of K is also not 3, so that division by 3 is possible, we may complete the cube: replace X by $X - a/3$ to get an equation of the form

$$Y^2 = X^3 + bX + c$$

with b, $c \in K$. This is a *short Weierstrass equation.* The condition that this equation defines a smooth curve is that $\Delta \neq 0$, where

$$\Delta = \Delta_{b,c} = -4b^3 - 27c^2.$$

Case 2: no rational flex

Suppose that $\mathcal{E}$ is an elliptic curve defined over a field K, that P_1 is a point in $\mathcal{E}(K)$ with tangent $\mathcal{L}$, and that P_1 is *not* a flex. Then the line $\mathcal{L}$ intersects $\mathcal{E}$ twice at P_1 and once at a distinct point P_2. Let P_3 be a point on the tangent at P_2. Then P_1, P_2 and P_3 are not collinear, so after a projective change of coordinates (with coefficients in K) we may assume that $P_1 = [1:0:0]$, $P_2 = [0:1:0]$ and $P_3 = [0:0:1]$. Now the tangent to $\mathcal{E}$ at P_1 is $Z = 0$ and the tangent at P_2 is $X = 0$. The equation for $\mathcal{E}$ thus has the form

$$cXY^2 + eX^2Z + fXYZ + hXZ^2 + iYZ^2 + jZ^3 = 0 \tag{14}$$

where $c \neq 0$ and $e \neq 0$. Now comes the nonlinear change of variables. Multiply by XZ^2, and introduce new variables $U = XZ$, $V = XY$, $W = Z^2$, to give

$$cV^2W + eU^3 + fUVW + hU^2W + iVW^2 + jUW^2 = 0. \tag{15}$$

Scaling as before, and dehomogenising with respect to W, gives a long Weierstrass equation in the (scaled) variables U, V.

The transformation we used:

$$\varphi : [X:Y:Z] \mapsto [U:V:W] = [XZ:XY:Z^2]$$

is quadratic. It does map rational points to rational points, but it is not defined everywhere on $\mathbb{P}^2$; in fact it is not defined precisely at $[0:1:0]$ and $[1:0:0]$. Also, the whole line $Z = 0$ maps to the single point $[0:1:0]$. There is an inverse map, also not defined everywhere, given by

$$\psi : [U:V:W] \mapsto [X:Y:Z] = [U^2:VW:UW]$$

which is not defined at $[0 : 0 : 1]$ and $[0 : 1 : 0]$. Wherever the composites $\varphi\psi$ and $\psi\varphi$ are defined they act as the identity. Thus φ and ψ are mutually inverse bijections between points on (14) apart from $[1 : 0 : 0]$, $[0 : 1 : 0]$, and $[0 : -j : i]$ and points on (15) apart from $[0 : 1 : 0]$, $[0 : 0 : 1]$ and $[-i : 0 : 1]$. Rational points map to rational points.

φ is an example of a birational isomorphism between curves. Quadratic maps like this are in fact called *Cremona*[1] *transformations.* Hence we have proved the following result:

Theorem 4.6. *Every elliptic curve is birationally isomorphic to a curve in long Weierstrass form.*

In the first case, where all we needed was a projective change of coordinates, it is clear that the group law is unaffected by our change of variables, since it is defined in terms of straight lines. In the second case however, the birational transformation φ certainly does not take straight lines to straight lines. Miraculously, the group law is also preserved, since the condition that three points on the curve are "collinear on the curve" is preserved. We will not prove that here, and from now on we will only use Weierstrass equations.

4.5 The group law for Weierstrass equations

Let $\mathcal{E}$ be an elliptic curve defined over a field K by an affine Weierstrass equation of the form

$$Y^2 = X^3 + aX^2 + bX + c. \tag{16}$$

Write $g(X) = X^3 + aX^2 + bX + c$. The condition that $g(X)$ should not have repeated roots, so that $\mathcal{E}$ is smooth, is that the discriminant $\Delta = \Delta_{a,b,c} = \operatorname{disc}(g) \neq 0$.

The unique point at infinity on $\mathcal{E}$ is $\mathcal{O} = [0 : 1 : 0]$. It is a flex, since it is the only intersection point of $\mathcal{E}$ with the line $Z = 0$ at infinity. It is simpler to use affine coordinates, and remember that there is this one extra point on the curve. We always take this point $\mathcal{O}$ to be the identity for the group law, so that we can use the simpler forms developed above, where three points add to $\mathcal{O}$ if and only if they are collinear on the curve.

Negating points

Let $P = (x, y)$ be an affine point on $\mathcal{E}$. In homogeneous coordinates $P = [x : y : 1]$, so the line through P and $\mathcal{O}$ is the vertical line with homogeneous equation $X - xZ = 0$ or affine equation $X = x$. This line intersects $\mathcal{E}$ at P, $\mathcal{O}$, and $P' = (x, -y)$. It follows that, for $P = (x, y) \in \mathcal{E}(K)$ with $P \neq \mathcal{O}$,

$$\ominus P = P * \mathcal{O} = (x, -y).$$

So negating points on a curve in Weierstrass form is very easy: one just changes the sign of the Y-coordinate. Geometrically, this is just reflecting on the X-axis, which is a line of symmetry of the curve.

Adding points

Let $P_i = (x_i, y_i)$ be two points in $\mathcal{E}(K)$. We will see explicitly how to find their sum $P_1 \oplus P_2$. It is recommended that you do not memorise the resulting formulas, but rather remember the method so as to be able to apply it to numerical examples.

If $x_1 = x_2$ then $y_2 = \pm y_1$, so either $P_2 = P_1$ or $P_2 = \ominus P_1$. In the latter case, $P_1 \oplus P_2 = \mathcal{O}$ and there is nothing more to do. We leave the case $P_1 = P_2$ for later, so suppose now that $x_1 \neq x_2$.

We start by finding $P_1 * P_2$. The line $\mathcal{L} = \mathcal{L}_{P_1P_2}$ has slope $\lambda = \frac{y_2 - y_1}{x_2 - x_1}$, so has equation $Y = \lambda X + \mu$ where $\mu = y_2 - \lambda x_2 = y_1 - \lambda x_1$. Note that $\mathcal{L}$ has coefficients in K. To find the third point of intersection of $\mathcal{L}$ with $\mathcal{E}$, substitute $Y = \lambda X + \mu$ into the equation for $\mathcal{E}$ to get

$$(\lambda X + \mu)^2 = X^3 + aX^2 + bX + c,$$

[1] no relation

or

$$X^3 + (a - \lambda^2)X^2 + (b - 2\lambda\mu)X + (c - \mu^2).$$

The roots of this cubic add up to $\lambda^2 - a$, and two of the roots are x_1 and x_2, hence the third is

$$x_3 = \lambda^2 - a - x_1 - x_2.$$

Setting $y_3 = \lambda x_3 + \mu$, we have $P_3 = P_1 * P_2 = (x_3, y_3)$.

Finally,

$$P_1 \oplus P_2 = \ominus(P_1 * P_2) = \ominus P_3 = (x_3, -y_3).$$

Next we return to the case $P_1 = P_2$, and compute $2P_1$. If $y_1 = 0$, then $P_2 = P_1 = \ominus P_1$ and then $2P_1 = \mathcal{O}$. Otherwise, we take for $\mathcal{L}$ the tangent at P_1, which has slope

$$\lambda = \frac{3x_1^2 + 2ax_1 + b}{2y_1}.$$

The rest is now as before.

Note that in all cases the sum $P_1 \oplus P_2$ has coordinates in K, as we expected, since the only operations we have carried out are the four arithmetic ones (addition, subtraction, multiplication and division). To summarise, we have:

Proposition 4.7. *Let $\mathcal{E}$ be an elliptic curve in Weierstrass form (16) and let $P_i = (x_i, y_i)$ be two points on $\mathcal{E}$.*

If $x_1 = x_2$ and $y_1 = -y_2$ then $P_1 \oplus P_2 = \mathcal{O}$.

If $x_1 = x_2$ and $y_1 = y_2 \neq 0$, set $\lambda = \frac{3x_1^2+2ax_1+b}{2y_1}$, otherwise set $\lambda = \frac{y_2-y_1}{x_2-x_1}$; set $\mu = y_1 - \lambda x_1$. Then

$$P_1 \oplus P_2 = (x_3, -y_3)$$

where $x_3 = \lambda^2 - a - x_1 - x_2$ and $y_3 = \lambda x_3 + \mu$.

4.6 Points of finite order I

Definition 27. A point on the elliptic curve $\mathcal{E}$ is said to have *finite order* if it has finite order in the group $\mathcal{E}(K)$, so that $nP = \mathcal{O}$ for some $n > 0$. Points of finite order are also called *torsion points.*

The set of all points of finite order is denoted $\mathcal{E}(K)_{\text{tors}}$; it is a subgroup of $\mathcal{E}(K)$.

For fixed $n > 0$, the set

$$\mathcal{E}(K)[n] = \{P \in \mathcal{E}(K) : nP = \mathcal{O}\}$$

of points of order dividing n is also a subgroup, called the n-torsion subgroup of $\mathcal{E}(K)$.

Here, nP means $P \oplus P \oplus \cdots \oplus P$ with n summands. The fact that $\mathcal{E}(K)_{\text{tors}}$ and $\mathcal{E}(K)[n]$ are subgroups is just elementary group theory.

Points of order two

Proposition 4.8. *Let $\mathcal{E}$ be the elliptic curve defined over the field K by the affine equation $Y^2 = g(X)$, where $g(X) \in K[X]$ has degree 3 and no repeated roots. A point $P = (x, y) \in \mathcal{E}(K)$ has order 2 iff $y = 0$ and $g(x) = 0$.*

$\mathcal{E}(K)[2]$ has order 1, 2 or 4, according as the number of roots of $g(X)$ in K is 0, 1 or 3 respectively.

Hence, if we extend the field K (if necessary) to a field L containing the roots of the cubic $g(X)$, in $\mathcal{E}(L)$ we will have $\mathcal{E}(L)[2]$ of order 4; and since all nontrivial elements of $\mathcal{E}(L)[2]$ have order 2, the structure is that of a Klein 4-group (the direct product of two cyclic groups of order 2).

The three points of order 2 are collinear on $\mathcal{E}$, being the intersection of $\mathcal{E}$ with the line $Y = 0$. Group-theoretically, it is clear that the sum of two distinct elements of order 2 again has order 2.

Examples: 1. $Y^2 = X^3 - X$ has three points of order 2, namely $(0, 0)$, $(\pm 1, 0)$, so $\#\mathcal{E}(K)[2] = 4$.

2. $Y^2 = X^3 - 2X$ has only one rational point of order 2, namely $(0, 0)$, but two additional real points of order 2: $(\pm\sqrt{2}, 0)$.

3. $Y^2 = X^3 - 2$ has no rational points of order 2, one real point $(\sqrt[3]{2}, 0)$, and two conjugate complex points of order 2.

Since a real cubic always has at least one real root, $\mathcal{E}(\mathbb{R})[2]$ will never be the trivial group. It has order 2 or 4 according to the sign of the discriminant $\Delta = \mathrm{disc}(g)$.

Points of order three

A point P on an elliptic curve satisfies $3P = \mathcal{O}$ iff $P \oplus P = \ominus P$ which is iff $P * P = P$, or equivalently P is a flex (point of inflection) on $\mathcal{E}$. There are at most 9 such points (including $\mathcal{O}$) forming a subgroup $\mathcal{E}(K)[3]$ which is either trivial, or cyclic of order 3, or the direct product of two cyclic groups of order 3.

Over $\mathbb{C}$ we always have $\#\mathcal{E}(\mathbb{C})[3] = 9$; over $\mathbb{R}$ we always have $\#\mathcal{E}(\mathbb{C})[3] = 3$ (see Exercise 2.2 on page 58 of Silverman and Tate). Over $\mathbb{Q}$ either $\#\mathcal{E}(\mathbb{C})[3] = 3$ or $\#\mathcal{E}(\mathbb{C})[3] = 1$. For example, $P = (0, 1)$ has order 3 on $Y^2 = X^3 + 1$ (the tangent at P is $Y = 1$). The curve $Y^2 = X^3 + 2$ has no rational points of order 3; its real points of order 3 are $(0, \pm\sqrt{2})$.

In general, the situation over $\mathbb{C}$ is very simple. For each $n \geq 1$, $\mathcal{E}(\mathbb{C})[n]$ is isomorphic to $C_n \times C_n$. To prove this one can use the parametrization of $\mathcal{E}(\mathbb{C})$ by elliptic functions. Over $\mathbb{R}$ one finds that $\mathcal{E}(\mathbb{R})[n] \cong C_n$ for odd n; the situation for even n depends on the sign of the discriminant.

Over any field K, the structure of the finite group $\mathcal{E}(K)$ is quite restricted. It is always either cyclic or a product of two cyclic factors. This can be seen (briefly) as follows. First, the equation $nP = \mathcal{O}$ cannot have more than n^2 solutions P. This is because (when n is odd, say) the X-coordinate of P must satisfy an equation of degree $(n^2 - 1)/2$, using explicit techniques as above. Now the structure theorem for finite abelian groups implies that an abelian group with this property cannot have more than 2 cyclic factors.

So for any field K, $\mathcal{E}(K)[n] \cong C_{n_1} \times C_{n_2}$, where $n_1 | n_2 | n$.

Over $\mathbb{Q}$ the only relatively easy result is that since $\mathbb{Q} \subset \mathbb{R}$ we have that $\mathcal{E}(\mathbb{Q})[n]$ is a subgroup of $\mathcal{E}(\mathbb{R})[n]$; in particular, this is cyclic if n is odd or $\Delta < 0$. Beyond that things become more complicated. There are examples of rational points on elliptic curves over $\mathbb{Q}$ with order n for any $n \leq 10$ and also $n = 12$; and it turns out that these are the only possibilities. For example, $P = (-9, 49)$ has order 12 on the curve $Y^2 + XY + Y = X^3 - X^2 - 122X + 1721$. Proving that there are no points of other orders is difficult, beyond the scope of this module.

4.7 Points of finite order II

Let $\mathcal{E}$ be an elliptic curve defined over $\mathbb{Q}$ by an affine equation of the form $Y^2 = g(X) = X^3 + aX^2 + bX + c$. By scaling the equation suitably, we may assume that all the coefficients are integers (if not, multiply through by d^6 for a suitable positive integer d, and use new variables d^2X and d^3Y). Recall that the condition for $\mathcal{E}$ to be smooth is that the discriminant $\Delta = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2$, which is now an integer, is not 0.

Proposition 4.9. *Let $P = (x, y) \in \mathcal{E}(\mathbb{Q})$ with $\mathcal{E}$ as above. Then $x = \frac{u}{w^2}$ and $y = \frac{v}{w^3}$ where $u, v, w \in \mathbb{Z}$, $w \geq 1$, and $\gcd(u, w) = \gcd(v, w) = 1$.*

In other words, if x is integral then so is y, while if not then the denominator of x is a perfect square, say w^2, and the denominator of y is w^3. See page 3 for examples.

Definition 28. An *integral point* on the curve $Y^2 = X^3 + aX^2 + bX + c$ (where $a, b, c \in \mathbb{Z}$) is a point $P = (x, y)$ where $x, y \in \mathbb{Z}$.

There is a connection between whether a rational point is integral and whether it has finite order in the group $\mathcal{E}(\mathbb{Q})$. The Lutz-Nagell Theorem says that points of finite order are integral; we can also bound their Y-coordinate, so they are finite in number and we can find them easily for a given curve. The rough idea is that if a point P is not integral, then successive multiples of P have larger and larger denominators so can never reach $\mathcal{O}$.

Theorem 4.10 (Nagell-Lutz). *Let $\mathcal{E}:\ Y^2 = X^3 + aX^2 + bX + c$ be an elliptic curve with integer coefficients a, b, c. Set $\Delta = -4a^3c + a^2b^2 + 18abc - 4b^3 - 27c^2$. Let $P = (x, y) \in \mathcal{E}(\mathbb{Q})$ be a rational point of finite order. Then P is integral, and either $y = 0$ or $y^2 | \Delta$.*

The first part is quite hard to prove: see the handout for details. The second part follows quite easily from it.

Proposition 4.11. *Let $\mathcal{E}$ be the elliptic curve with equation as above. Let $P = (x, y) \in \mathcal{E}(\mathbb{Q})$ be a point such that both P and $2P$ are integral. Then $y^2 | \Delta$.*

To deduce the second part of Nagell-Lutz, observe that if $P = (x, y)$ has finite order then either $2P = \mathcal{O}$, in which case $y = 0$, or $2P$ is another point of finite order, in which case $2P$ is also integral.

Warning: not all integral points have finite order! It is even possible for both P and $2P$ to be integral while P has infinite order. Also, the Theorem only applies to curves in Weierstrass form with integral coefficients. A curve in long Weierstrass form with integer coefficients can have non-integral points of finite order, but only of order 2, and they are "almost" integral, of the form $(u/4, v/8)$ with $u, v \in \mathbb{Z}$.

Corollary 4.12. *Let $\mathcal{E}$ be an elliptic curve defined over $\mathbb{Q}$. Then the torsion subgroup $\mathcal{E}(\mathbb{Q})_{tors}$ is finite.*

We may use the Lutz-Nagell Theorem to find all torsion points on a given curve. Compute the integer Δ, list the integers y such that $y = 0$ or $y^2 | \Delta$, and for each solve the cubic equation $g(X) - y^2 = 0$ for X. This gives a finite list of points. For each point P on the list, to see if it really does have finite order, compute nP for $n = 2, 3, 4 \ldots$ until one of the following happens: if nP is not integral then nP and hence P has infinite order; if $nP = (x, 0)$ then nP has order 2 so P has order $2n$; if $nP = -mP$ (where $n > m > 0$) then P has order (at most) $m + n$. (For example, if P has order 5 we will notice that $3P = -2P$ since $3P$ and $2P$ have the same X-coordinate.)

Finally, the complete result for torsion on elliptic curves over $\mathbb{Q}$ is given by the following Theorem, proved in 1977 by Mazur. Its proof requires techniques far more advanced than the ones we have been using.

Theorem 4.13. *Let $\mathcal{E}$ be an elliptic curve defined over $\mathbb{Q}$. Then $\mathcal{E}(\mathbb{Q})_{tors}$ is isomorphic to one of the following finite abelian groups: C_n for $1 \le n \le 10$, C_{12}, $C_2 \times C_{2n}$ for $1 \le n \le 4$.*

Examples of all these 15 possible torsion subgroups may be found in Exercise 2.2 on page 62 of Silverman and Tate.

4.8 Elliptic Curves over Finite Fields

In this section we take the field of definition of our curves to be the finite field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. An elliptic curve defined over $\mathbb{F}_p$ will have a long Weierstrass equation; if $p > 2$ it will even have a standard Weierstrass equation of the form $Y^2 = g(X)$, where now $g(X) \in \mathbb{F}_p[X]$. Since $\mathbb{F}_p$ is finite, clearly the group of points $\mathcal{E}(\mathbb{F}_p)$ is a finite group.

In this section, we will consider the basic properties of elliptic curves over $\mathbb{F}_p$. (These also apply to other finite fields.) The finite groups $\mathcal{E}(\mathbb{F}_p)$ are of use in cryptography, and we will return to this application in the last chapter. In the next section, we will see how by using "reduction modulo p" we can obtain information about elliptic curves over $\mathbb{Q}$.

Since $\mathbb{P}^2(\mathbb{F}_p)$ is finite (with p^2+p+1 points) is is clear that an elliptic curve over $\mathbb{F}_p$ cannot have more than about p^2 points. In fact, this rough estimate can easily be improved. Given the equation

$$\mathcal{E}: \; Y^2 = g(X) = X^3 + aX^2 + bX + c,$$

to each value of x there are at most 2 corresponding values of y for which $(x, y) \in \mathcal{E}(\mathbb{F}_p)$, depending on whether or not $g(x)$ is a square in $\mathbb{F}_p$. This implies that $\#\mathcal{E}(\mathbb{F}_p) \leq 2p+1$ (including the point at infinity). In fact the number of x for which $g(x)$ is a square in $\mathbb{F}_p$ turns out to be roughly equal to the number for which $g(x)$ is not a square, so that on average there is one affine point per value of x, and the number of points on the curve is approximately $p+1$. This is made precise in the following Theorem, whose proof is omitted as it is far from elementary.

Theorem 4.14 (Hasse's estimate). *Let $\mathcal{E}$ be an elliptic curve defined over the finite field $\mathbb{F}_q$. Then*

$$|\#\mathcal{E}(\mathbb{F}_q) - (q+1)| \leq 2\sqrt{q}.$$

So the number of points lies between $q+1-2\sqrt{q} = (1-\sqrt{q})^2$ and $q+1+2\sqrt{q} = (1+\sqrt{q})^2$. For example, over $\mathbb{F}_5$ the number of points is between 2 and 10.

To count the number of points on an elliptic curve defined over $\mathbb{F}_p$, the simplest way is to make a table of $g(x)$ for $0 \leq x \leq p-1$, and see which values are squares. Count 2 for a non-zero square and 1 when $g(x) = 0$. Then add 1 for the point at infinity.

Example: Let $\mathcal{E}$ be the curve $Y^2 = g(X) = X^3 - X + 1$ over $\mathbb{F}_3$. (This is smooth since $\Delta = -23 \neq 0$ in $\mathbb{F}_3$). Hasse's estimate tells us that $\mathcal{E}(\mathbb{F}_3)$ has between 1 and 7 points. Now $g(0) = g(1) = g(-1) = 1$ (in $\mathbb{F}_3$) so we have 6 affine points and 7 points altogether. The group $\mathcal{E}(\mathbb{F}_3)$ must therefore be cyclic, and one can check that $P = (0, 1)$ is indeed a generator.

4.9 Reduction modulo a prime

Recall from Lemma 3.6 that every point in $\mathbb{P}^2(\mathbb{Q})$ can be written, uniquely up to sign, in the form $[x : y : z]$ with $x, y, z \in \mathbb{Z}$ and $\gcd(x, y, z) = 1$. Using a bar to denote the map $\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ this gives us a well-defined function

$$\mathbb{P}^2(\mathbb{Q}) \to \mathbb{P}^2(\mathbb{F}_p) \qquad [x : y : z] \mapsto [\overline{x} : \overline{y} : \overline{z}].$$

Similarly for $\mathbb{P}^n(\mathbb{Q})$. So we can reduce rational points in projective space modulo a prime p.

We can also reduce rational curves. Given a homogeneous polynomial $F \in \mathbb{Q}[X, Y, Z]$ defining a curve $\mathcal{C} = \mathcal{C}_F$, we can scale F so that its coefficients are coprime integers. Again, this is unique up to sign. Then by reducing all the coefficients modulo p, we obtain a *nonzero* homogeneous polynomial in $\overline{F} \in \mathbb{F}_p[X, Y, Z]$ with $\deg(\overline{F}) = \deg(F)$. The projective curve defined by $\overline{F}$ over $\mathbb{F}_p$ is called the *reduction of $\mathcal{C}$ modulo p*, and will be denoted by $\overline{\mathcal{C}}$, or by $\overline{\mathcal{C}}_p$ if we need to be explicit about the prime.

Both these operations are much easier in the projective context. Reducing a rational point in the affine plane $\mathbb{A}^2(\mathbb{Q})$ modulo p is not possible when p divides the denominator of one of the coordinates. Reducing a non-homogeneous polynomial $f(X, Y) \in \mathbb{Q}[X, Y]$ modulo p is better, since we are still free to scale f first, but the degree can go down. For example, $5X^2 + Y - 1$ reduces to the linear polynomial $Y - 1$ modulo 5, but the homogenised form $5X^2 + YZ - Z^2$ reduces to $YZ - Z^2$ which is still quadratic.

Lemma 4.15. *Let $F(X, Y, Z) \in \mathbb{Q}[X, Y, Z]$ be a form of degree n defining a rational curve $\mathcal{C}$ of degree n. Let $P \in \mathbb{P}^2(\mathbb{Q})$. Fix a prime p. Define the reductions $\overline{F}$ and $\overline{P}$ of F and P modulo p as above. Then the curve $\overline{\mathcal{C}}_p = \mathcal{C}_{\overline{F}}$ is a curve of degree n in $\mathbb{P}^2(\mathbb{F}_p)$, and*

$$P \in \mathcal{C}(\mathbb{Q}) \implies \overline{P} \in \overline{\mathcal{C}}(\mathbb{F}_p).$$

Note that the implication here is one-way only. For example, let $\mathcal{C}$ be the line defined by $F = X - Y$, and consider the rational points $P = [1 : 1 : 4]$ and $Q = [7 : 2 : 3]$. Then $P \in \mathcal{C}(\mathbb{Q})$ while $Q \notin \mathcal{C}(\mathbb{Q})$; however, if we reduce modulo 5 we find that $\overline{P} = \overline{Q} \in \overline{\mathcal{C}}_5$.

Our aim now is to show that if $\mathcal{E}$ is an elliptic curve defined over $\mathbb{Q}$ then for all but a finite number of primes p the reduction map $\mathcal{E} \to \overline{\mathcal{E}}_p$ is a *group homomorphism.* We will also be able to say exactly what happens to torsion points under the reduction, and hence use reduction modulo carefully chosen primes p to get information about the torsion in $\mathcal{E}(\mathbb{Q})$.

The main reason that some primes must be excluded is that the reduced curve $\overline{\mathcal{E}}_p$ is not necessarily an elliptic curve, since it may be singular. The concept of smoothness does not necessarily survive reduction modulo p. For example, the rational conic $X^2 - Y^2 = 5Z^2$ is nonsingular, but modulo 5 becomes the union of the two lines $Y \pm X$.

Theorem 4.16. *Let $\mathcal{E}$ be an elliptic curve defined over $\mathbb{Q}$ by the affine equation $Y^2 = g(X) = X^3 + aX^2 + bX + c$, where $a, b, c \in \mathbb{Z}$ and $\Delta \neq 0$. For all primes $p \not| 2\Delta$, the reduced curve $\overline{\mathcal{E}}_p$ is an elliptic curve over the field $\mathbb{F}_p$, and the reduction map $\mathcal{E}(\mathbb{Q}) \mapsto \overline{\mathcal{E}}_p(\mathbb{F}_p)$ is a group homomorphism.*

We call primes p satisfying the condition of the Theorem *good primes*, or *primes of good reduction* for $\mathcal{E}$; the finite set of primes dividing 2Δ are called *bad primes*, or *primes of bad reduction.* (We are simplifying things slightly here: with our definition the prime 2 is always bad. This is because we are only using the standard Weierstrass equations. For a proper treatment of the prime 2, one needs to use long Weierstrass equations.)

The key lemma which is needed to give the homomorphism property of reduction is the following, which says that the property of three points being "collinear on $\mathcal{C}$" is preserved under reduction.

Lemma 4.17. *Let $\mathcal{C}$ be rational curve of degree n, and $\mathcal{L}$ a rational line such that $\mathcal{C} \cap \mathcal{L}$ consists of the n rational points P_j $(1 \leq j \leq n)$ counted with multiplicity. Let p be a prime. Assume that $\overline{\mathcal{L}} \not\subseteq \overline{\mathcal{C}}$. Then $\overline{\mathcal{C}} \cap \overline{\mathcal{L}}$ consists of the n points $\overline{P_j}$, also with the correct multiplicities.*

The next result is the promised application of reduction modulo p to the determination of the torsion points in $\mathcal{E}(\mathbb{Q})$.

Theorem 4.18. *Let $\mathcal{E}$ be an elliptic curve defined over $\mathbb{Q}$ by the affine equation $Y^2 = g(X) = X^3 + aX^2 + bX + c$, where $a, b, c \in \mathbb{Z}$ and $\Delta \neq 0$. Let p be a prime such that $p \not| 2\Delta$. Then the reduction homomorphism $\varphi : \mathcal{E}(\mathbb{Q}) \to \overline{\mathcal{E}}(\mathbb{F}_p)$ is injective on torsion. In other words, $\mathcal{E}(\mathbb{Q})_{tors} \cap \ker(\varphi) = \{\mathcal{O}\}$.*

Corollary 4.19. *With the notation of the theorem, for all $p \not| 2\Delta$ we have*

$$\#\mathcal{E}(\mathbb{Q})_{tors} \mid \#\overline{\mathcal{E}}(\mathbb{F}_p).$$

To use this in practice, suppose that we have found some torsion points in $\mathcal{E}(\mathbb{Q})$ and suspect that they are all of them. Let m be the number of torsion points we know. If we can find a good prime p such that $\#\overline{\mathcal{E}}(\mathbb{F}_p) = m$ then we know that there are no more torsion points. Alternatively, without looking for any rational points on $\mathcal{E}$ at all, we may count $\#\overline{\mathcal{E}}(\mathbb{F}_p)$ for several good primes p, and take the gcd of these numbers. This gives us an upper bound for $\#\mathcal{E}(\mathbb{Q})_{\text{tors}}$.

Examples 1. The curve $\mathcal{E}: \; Y^2 = X^3 + 3$ has $\Delta = -243 = -3^5$ so the only bad primes are 2 and 3. The number of points modulo 5 and 7 is 6 and 13 respectively. Since $\gcd(6, 13) = 1$ it follows that $\mathcal{E}(\mathbb{Q})_{\text{tors}}$ is trivial. Now the point $P = (1, 2) \in \mathcal{E}(\mathbb{Q})$ cannot have finite order, so $\mathcal{E}(\mathbb{Q})$ is infinite. In effect, we have proved that the equation $Y^2 = X^3 + 3$ has infinitely many rational solutions by producing just one solution, and by solving a couple of congruences.

2. $\mathcal{E}: \; Y^2 = X^3 + X$ has $\Delta = -4$ so only $p = 2$ is bad. Reduction modulo 5 shows (with some care) that $\mathcal{E}(\mathbb{Q})_{\text{tors}} = \{\mathcal{O}, (0, 0)\}$, of order 2.

3. $\mathcal{E}: \; Y^2 = X^3 - 43X + 166$ has $\Delta = -2^{15}13$. It has 7 points modulo 3. The point $P = (3, 8) \in \mathcal{E}(\mathbb{Q})$ has order 7, so the torsion subgroup is cyclic of order 7, generated by P.

5 Applications

In this final chapter we give a brief introduction to two of the recent practical applications for elliptic curves, in cryptography and in the factorization of large integers.

5.1 Application 1: Elliptic Curves in Cryptography

Elliptic curves are increasingly being used in cryptography. This is called *Elliptic Curve Cryptography* or *ECC*. For further reading on this subject, see the book "Elliptic Curve Cryptography" by Blake, Seroussi and Smart (CUP).

The cryptographic situation

Suppose that two people, Alice and Bill, wish to send secret messages to each other. Their communications may be intercepted by Eve (an eavesdropper) so must be encrypted in such a way that Alice and Bill know how to decrypt each other's messages, but Eve cannot do this – at least, not quickly or easily. They have a way of encrypting their messages which involves both of them knowing a certain secret *key*, which is a large integer. We will not discuss any specific cryptographic systems which use such a private key here, but concentrate on the following problem: how can Alice and Bill obtain shared knowledge of a private key in the first place? We assume that it is impractical for them to communicate directly (e.g. by whispering in each other's ear) but that they must communicate at a distance, by letter or electronically, throughout. Once they both know their shared private key they can start to encrypt messages knowing that only the other can decrypt them.

One method for doing this uses elliptic curves. It is a special case of a more general scheme, based on the *Discrete Logarithm problem* or *DLP*.

The mathematical setup

Let G be a very large finite abelian group, and g an element of G of large order n. Given an integer m with $1 < m < n$, set $h = g^m$ (using the group operation in G). The *Discrete Logarithm Problem* (*DLP*) for G is:

Given g and h in G such that $h = g^m$, find m.

The reason for the name "discrete logarithm" is that m is the "logarithm to base g" of h. In fact, if we were to take $G = \mathbb{R}_+^*$, the multiplicative group of positive real numbers (which admittedly is not finite), we could just compute $m = \log_g(h) = \log(h)/\log(g)$ as usual.

There is no essential difference if G is an additive group; then we are given g and $h = m.g$ and must find m. In ECC we are in this situation, taking $G = \mathcal{E}(\mathbb{F}_q)$ where $\mathcal{E}$ is an elliptic curve defined over the finite field $\mathbb{F}_q$.

In practice, the group G and the base element g are all publicly available, and the element $h = g^m$ is also known, but the exponent m is secret. To be useful, the group G must satisfy certain conditions:

1. it must be possible to compute $h = g^m$ quickly, given g and m, when m is large;
2. it must be very hard to recover m from knowledge of g and g^m: in other words, the DLP for G must be hard.

The first condition is no problem, since g^m can always be computed in about $\log(m)$ steps. See below for some details and examples. The second condition is the crucial one. Much research effort is currently being spent on finding new ways to "break" the DLP for various classes of groups, making those groups unsuitable for cryptographic purposes. So far, general elliptic curve groups have resisted these attacks, though there are special families of curves where the DLP has a much easier solution, and these must clearly be avoided.

What do Alice and Bill do?

Alice and Bill agree on which group G to use and which element g. These need not be kept secret. They both know the order n of g, which is large.

Alice chooses a *secret* integer a in the range $1 < a < n$ with $\gcd(a, n) = 1$ and computes $h_1 = g^a$ using the group operation in G. She sends h_1 to Bill, over an open line. At the same time, Bill also chooses a secret integer b in the range $1 < b < n$ with $\gcd(b, n) = 1$, computes $h_2 = g^b$, and sends h_2 to Alice. Eve can intercept h_1 and h_2, but even though she also knows g she cannot recover a and b from them, provided that the DLP for G is hard.

Now Bill takes the element h_1 received from Alice and computes $h = h_1^b$, while Alice takes h_2 and computes $h = h_2^a$. Note that these elements h really are the same, since

$$h = h_1^b = (g^a)^b = g^{ab} = (g^b)^a = h_2^a.$$

Now both Alice and Bill know h, but no-one else does and no-one else can find it without expending a very large amount of effort to break the DLP, so Alice and Bill can use h as a secret key for their secret purposes.

Possible Groups to use

Example 1: additive group modulo n. As a first example, which will turn out to be useless for cryptographic purposes, take the additive group $G = \mathbb{Z}/n\mathbb{Z}$, and g any integer coprime to n. Now for $a \in \mathbb{Z}$ set $h = ag \pmod{n}$. Given g and h it is easy to find a, since this is the same as solving the congruence $ag \equiv h \pmod{n}$ for a. Since g is coprime to n we may use the Extended Euclidean Algorithm to solve $1 = ug + vn$ and then set $a = uh$. So in this group the DLP is trivial to solve (thanks to the EEA), making it useless for cryptography.

Example 2: multiplicative group modulo n. Choose a large prime p and let $G = \mathbb{F}_p^* = (\mathbb{Z}/p\mathbb{Z})^*$, the multiplicative group of the finite field $\mathbb{F}_p$. This is a cyclic group of order $n = p - 1$. We take g to be a primitive root modulo p, which is (by definition) a generator of G. The DLP of G can be now expressed as follows: given $h \in G = \langle g \rangle$, find the exponent a such that $h = g^a$. This problem has been studied intensively, and there are quite efficient methods for solving it. So for cryptographic use one needs to use a *very* large prime p. One drawback of this choice is that the prime p, and hence the order of G, are public knowledge, and this helps the attacks, particularly if $p - 1$ can be factorized easily.

According to a recent announcement (17 April 2001), a DLP was solved modulo a prime with 120 digits using the latest techniques; the computation took 10 weeks of computing time on a 4-processor Digital Alpha Server 8400 computer. It involved algebraic number theory, and the solution of a system of 2685597 equations in 1242551 unknowns. This work was done in Paris by Joux and Lercier.

Example 3: ECDLP. Again, choose a large prime p, and let $G = \mathcal{E}(\mathbb{F}_p)$ for a suitably chosen elliptic curve $\mathcal{E}$ defined over $\mathbb{F}_p$. The group G is "almost" cyclic (it is either cyclic or the product of two cyclic factors), and we assume that we have a "base" point $P \in \mathcal{E}(\mathbb{F}_p)$ with large order n, around the same size as p.

Given $a \in \mathbb{Z}$ coprime to n, we compute $Q = aP$ using the group law on $\mathcal{E}$, where all computations are done modulo p. Even if a is large, this can be done efficiently as we explain below. Recovering a from P and Q is hard in general. Note that while p, $\mathcal{E}$ and P are public, the orders of $\mathcal{E}(\mathbb{F}_p)$ and of P are not, and they can be very time-consuming to compute.

An alternative to using a large prime p is to use a large prime power q, usually a large power of 2, and work over the field $\mathbb{F}_q$. Computer chips have been designed for use on smart cards on which the arithmetic of points on an elliptic curve over such a field is wired into the electronic circuit!

5.1.1 Computing large powers and multiples quickly

To compute a large multiple $Q = aP$ of an element P in a group is much faster than it seems at first sight. The naive method would involve starting with $Q = P$ and adding P to

Q repeatedly: this takes $a - 1$ steps, which is certainly impractical for large a, even if the individual steps can be done quickly.

Instead one uses a method which involves approximately $\log(a)$ steps. By repeated doubling one computes $2P$, $4P$, $8P$, ..., $2^k P$ where $2^k \leq a < 2^{k+1}$. Then by adding a suitable combination of these (based on the binary representation of a) one obtains aP.

For example,

$$59P = (1 + 2 + 8 + 16 + 32)P = P + 2P + 8P + 16P + 32P,$$

which can be computed using 5 doublings and a further 4 additions, making only 9 operations in all compared with 58 for the naive method. For larger a, the saving is even more significant.

The same method works in a multiplicative group, with repeated squaring instead of doubling.

5.2 Application 2: Elliptic Curves in Factorization

Elliptic curves may be used to factorize large integers. This factorization method is called the *Elliptic Curve Method* or *ECM*. For more details of these factorization methods, refer to several of the books on the G13NUM reading list, or "A course in Computational Number Theory" by H. Cohen (Springer).

Statement of the problem

> Given an odd, positive integer N which is known to be a composite, find a non-trivial factorization $N = ab$ (with $a, b > 1$).

In practice one can easily prove that a composite number N is indeed composite, without factorizing N, using what are usually called primality tests. The simplest of these are the pseudo-primality or Fermat tests based on Fermat's Little Theorem. If for some integer $a > 1$ with $\gcd(a, N) = 1$ we have $a^{N-1} \not\equiv 1 \pmod{N}$, then N is certainly not prime. Unfortunately there are numbers which pass all these tests but are not prime; however there are stronger tests known, so that in practice we never try to factorize a number unless we are certain that it is composite.

Trial Division

The most obvious way to find a factorization of N is to test each possible factor $a = 2, 3, \ldots$ in turn. One can restrict to prime values of a (since the smallest divisor of N greater than 1 is certainly prime), and can stop when all divisors $\leq \sqrt{N}$ have been tested. So this method takes about $\sqrt{N}$ steps. It also requires a large list of known primes, unless we are willing to waste time by testing non-prime divisors as well.

This method is fine for small numbers, up to a million or so, and the more sophisticated methods usually assume that the number N being factorized has no small prime factors.

Pollard's $p - 1$ method

Almost all factorization methods currently in use, other than some classical ones (due to Fermat and others) and the Elliptic Curve Method, were first thought up by the English mathematician John Pollard. Pollard's $p-1$ method is simple to describe, and uses a simpler version of the same ideas as we will see again in the ECM.

A number N will be successfully factorized by the $p - 1$ method if some prime divisor p of N has the property that $p-1$ is *smooth*, meaning that $p-1$ has no large prime factors. In contrast with Trial Division, the size of N itself is almost irrelevant; instead, it is this special property of certain primes which makes it easy for them to be detected as factors of larger numbers.

Choose a base a which is coprime to N. If $g = \gcd(a, N) > 1$ then we have a factor g of N and can stop anyway. Compute successively the sequence (a_k) for $k = 1, 2, 3, \ldots$, where

$$a_k \equiv a^{k!} \pmod N$$

by setting $a_1 = a$ and successively $a_k = a_{k-1}^k \pmod N$. Raising to the power k mod N is done using the repeated squaring method of the previous section, with reduction modulo N at each intermediate step. At each stage also compute

$$g_k = \gcd(a_k - 1, N).$$

If ever $1 < g_k < N$, then g_k is a factor of N and we have succeeded. If ever $g_k = N$, then $a_k \equiv 1 \pmod N$ and we have failed; either start again with a new base a, or give up and use another method.

Why does this work? Suppose that p is a prime divisor of N, and that $p-1$ is smooth. Then for a fairly small value of k we will have $(p-1)|k!$. Now $\gcd(a, p) = 1$, since $\gcd(a, N) = 1$ and $p|N$, so by Fermat's Little Theorem we have $a^{p-1} \equiv 1 \pmod p$. Since $(p-1)|k!$, we also have $a_k \equiv a^{k!} \equiv 1 \pmod p$, and so p divides $g_k = \gcd(a_k - 1, N)$.

In practice we will decide in advance a maximum value K of k and give up if $g_k = 1$ for all $k \leq K$. Trying a different base a is rather unlikely to succeed, since failure usually means that none of the prime factors of N have the crucial property that $p-1$ is smooth.

Examples. 1. $N = 403$, $a = 2$. This leads to $a_k = 2, 4, 64, 326, \ldots$ and $g_k = 1, 1, 1, 13$ so $13|N$: in fact, $N = 13 \cdot 31$. This works because $13 - 1 = 12|4!$ while $31 - 1 = 30 \not| 4!$ so when $k = 4$, g_k is divisible by 13 but not by 31.

2. $N = 1891$. With $a = 2$ we get $a_k = 2, 4, 64, 264, 1$ so $g_5 = N$ and we fail. But with $a = 11$ we get $a_k = 11, 121, 1585, 1465, \ldots$ and $g_4 = 61$ giving $N = 61 \cdot 31$. What happened here was that $30 = 31 - 1$ and $60 = 61 - 1$ are equally smooth, but with $a = 11$ we were lucky since $11^{4!} - 1$ is divisible by 61 but not by 31.

3. $N = 5157437$. Now $a = 2$ gives $g_k = 1$ for $k \leq 8$ but $a_9 = 4381440$, $g_9 = \gcd(4381439, N) = 2269$, and $N = 2269 \cdot 2273$. Here $p = 2269$ has $p - 1 = 2^2 \cdot 3^4 \cdot 7$ which is very smooth, while $2273 - 1 = 2^5 \cdot 71$ is much less smooth. What makes the method work here is the smoothness of 2269, so this method will always be good at finding 2269 as a prime factor, whatever the size of N. For example, in only 9 steps again we can factor the number pq where $p = 2269$ and $q = 10^{99} + 289$, the smallest prime with 100 digits.

ECM

The drawback of the Pollard $p-1$ method is that each possible prime factor p has only "one chance" of being discovered, namely if $p-1$ is smooth. The quantity $p-1$ is significant here as it is the order of the multiplicative group $\mathbb{F}_p^*$. Using the group of points on an elliptic curves modulo p instead of the multiplicative group gives us many chances, since by changing the curve $\mathcal{E}$ we can use many groups whose orders range between $p+1-2\sqrt{p}$ and $p+1+2\sqrt{p}$, and if one of these orders is smooth then we will succeed.

More precisely, in Pollard's method we work in the multiplicative group $U_N = (\mathbb{Z}/N\mathbb{Z})^*$, whose order we do not know in practice, since there is no way to find its order $\varphi(N)$ which is faster than to factorize N and use the standard formula. Now if $p|N$ then there is a group homomorphism $U_N \to U_p = \mathbb{F}_p^*$. We take a random $a \in U_N$, compute high powers of it, and hope that we get non-trivial elements of the kernel of the reduction map. This will happen if the order of a in U_p is much less than the order of a in U_N, which is quite likely since the order of U_p itself is much less than the order of U_N.

Given our odd composite integer N, pick "at random" an elliptic curve $\mathcal{E}$ of the form $Y^2 = X^3 + bX + c$ with a point $P = (x, y)$ on $\mathcal{E}$ modulo N, so $y^2 \equiv x^3 + bx + c \pmod N$. To do this, fix b, x, y arbitrarily and set $c = y^2 - x^3 - bx$. Now compute $P_k = k!P$ for $k = 2, 3, 4, \ldots$, using the repeated doubling trick to compute $P_k = kP_{k-1}$ in about $\log(k)$ steps. All the arithmetic is done modulo N. Since N is not prime, so that $\mathbb{Z}/N\mathbb{Z}$ is not a field, we may find along the way that we need to divide by a number d which is non-zero but not invertible modulo N; but if this happens, then $g = \gcd(d, N) > 1$, and so g will be a factor of N and we can successfully stop.

Otherwise, we continue to compute the points P_k successively. Eventually, k will be large enough so that for some prime divisor p of N, $k!$ will be a multiple of the order of the group $\mathcal{E}_p(\mathbb{F}_p)$; this will happen sooner rather than later if the order $\#\mathcal{E}_p(\mathbb{F}_p)$ is smooth. But then the point P_k reduces to the identity $\mathcal{O}$ modulo p, and we will detect this since in this case the computation of P_k will involve an attempt to divide by an integer which is divisible by p and hence not coprime to N.

This is in fact much easier in practice than it sounds in theory. One simply writes down a curve $\mathcal{E}$ and a point P on $\mathcal{E}$ modulo N, computes successively higher multiples $k!P$ of P using arithmetic modulo N, and **either** we will reach P_K (for our predetermined maximum value K of k) without mishap, in which case we restart with another curve, **or** we find a divisor of N along the way, during an attempt to divide by an integer not coprime to N.

In the examples below, we use a variation where instead of $P_k = k!P$ we use $P_k = l_kP$, where $l_k = \operatorname{lcm}(1, 2, 3, \dots, k)$. We do not compute the P_k successively, but go straight for P_K using the repeated doubling method. A factor of N will be revealed if any of the necessary steps of doubling or adding two points involves a division by an integer not coprime to N, and this is certain to occur if l_K is a multiple of the order of $\mathcal{E}(\mathbb{F}_p)$ for some prime $p|N$.

Examples. 1. $N = 1715761513$. This is composite since $2^{N-1} \not\equiv 1 \pmod{N}$. We take $P = (2, 1)$ and vary b, setting $c = -7 - 2b$. So the elliptic curve is $Y^2 = X^3 + bX - (7 + 2b)$ on which P lies for all b.

First take $b = 1$. We compute lP where $l = l_{100} = \operatorname{lcm}(1, 2, 3, \dots, 100)$. Nothing interesting happens, so we try the next value of b. With this value of l it takes until $b = 3$ before we find that we have to try to invert 26927 modulo N, which is impossible, and in fact $N = 26927 \cdot 63719$. To see what is happening here, we count the number of points on the curves modulo 26927, for each b. We find that the order for $b = 1$ is $2 \cdot 3^2 \cdot 1493$, for $b = 2$ it is $2 \cdot 5 \cdot 2699$, while for $b = 3$ it is $2^3 \cdot 3 \cdot 19 \cdot 59$, which is much smoother than the preceding values (and in fact $l = l_{59}$ would have sufficed).

The $b = 6$ curve has an even smoother order modulo the other prime factor 63719, namely $2^4 \cdot 3^4 \cdot 7^2$. However, on account of the higher powers appearing here, the smallest l_k which is divisible by this order is l_{81}.

2. $N = 7560636089$. This is composite since $2^{N-1} \not\equiv 1 \pmod{N}$. We take $P = (1, 3)$ and vary b, setting $c = 8 - b$. Using l_{17} as the multiplier we have to wait until $b = 120$ before finding the factor $p = 15121$ of N; the order of the associated elliptic curve modulo p is the smooth number $2^4 \cdot 5 \cdot 11 \cdot 17$. Increasing the multiplier to l_{25} brings earlier success, with $b = 1$, where the order modulo p is the ultra-smooth $5 \cdot 7 \cdot 19 \cdot 23$.

3. $N = pq$ where $p = 10^6 + 3$ and $q = 10^7 + 19$. We take $P = (1, 3)$ as before. Using l_{25} as the multiplier we find the factor p with $b = 1278$; using l_{50} finds p at $b = 80$; and using l_{100} finds p at $b = 58$.

Clearly, an efficient implementation of this method needs careful tuning of the parameters.