1

Real Algebraic Curves

Plane curves arise naturally in numerous areas of the physical sciences (such as particle physics, engineering robotics and geometric optics) and within areas of pure mathematics itself (such as number theory, complex analysis and differential equations). In this introductory chapter, we will motivate some of the basic ideas and set up the underlying language of affine algebraic curves. That will also give us the opportunity to preview some of the material you will meet in the later chapters.

1.1 Parametrized and Implicit Curves

At root there are two ways in which a curve in the real plane $\mathbb{R}^2$ may be described. The distinction is quite fundamental.

- A curve may be defined *parametrically*, in the form $x = x(t)$, $y = y(t)$. The parametrization gives this image a dynamic structure: indeed at any parameter value t we have a *tangent vector* $(x'(t), y'(t))$ whose length is the *speed* of the curve at the parameter t. An example is the line parametrized by $x = t$, $y = t$, with constant speed $\sqrt{2}$, another parametrization such as $x = 2t$, $y = 2t$ yields the same image, but at twice the speed $2\sqrt{2}$.
- A curve may be defined *implicitly*, as the set of points (x, y) in the plane satisfying an equation $f(x, y) = 0$, where $f(x, y)$ is some reasonable function of x, y. For instance the line parametrized by $x = t$, $y = t$ arises from the function $f(x, y) = y - x$. Such a curve has no associated dynamic structure – it is simply a set of points in the plane.

Broadly speaking, the study of parametrized curves represents the beginnings of a major area of mathematics called *differential geometry*, whilst the study of curves defined implicitly represents the beginnings of another major area, *algebraic geometry*. It is the latter study which provides the material for this book, though at various junctures we will have something to say about the question of parametrization.

The common feature of many curves which appear in practice is that they are defined implicitly by equations of the form $f(x, y) = 0$ where $f(x, y)$ is a *real polynomial* in the variables x, y, i.e. given by a formula of the shape

$$f(x, y) = \sum_{i,j} a_{ij} x^i y^j$$

where the sum is finite and the coefficients a_{ij} are real numbers. There is much to gain in restricting attention to such curves, since they enjoy a number of important 'finiteness' properties. Moreover, it will be both profitable and illuminating to extend the concepts to situations where the coefficients a_{ij} lie in a more general 'ground field'. In some sense the complexity of a polynomial $f(x, y)$ is measured by its *degree*, i.e. the maximal value of $i + j$ over the indices i, j with $a_{ij} \neq 0$. Given a polynomial $f(x, y)$ we define its *zero set* to be

$$V_f = \{(x, y) \in \mathbb{R}^2 : f(x, y) = 0\}.$$

Instead of saying that a point (x, y) lies in the zero set of a curve f we may, for linguistic variety, say that (x, y) lies on the curve f, or that f passes through (x, y). Note that the zero set (and the degree) are unchanged when we multiply f by a non-zero scalar. It is for that reason that we introduce the following formal definition. A *real algebraic curve* is a non-zero real polynomial f, *up to multiplication by a non-zero scalar*. The more formally inclined reader may prefer to phrase this in terms of 'equivalence relations'. Two polynomials f, g are *equivalent*, written $f \sim g$, when there exists a non-zero scalar λ for which $g = \lambda f$. It is then trivially verified that $\sim$ has the defining properties of an equivalence relation: it is *reflexive* ($f \sim f$), it is *symmetric* (if $f \sim g$ then $g \sim f$), and it is *transitive* (if $f \sim g$ and $g \sim h$ then $f \sim h$). A real algebraic curve is then formally defined to be an equivalence class of polynomials under the relation $\sim$. So strictly speaking, a real algebraic curve is an equivalence set of all polynomials $\lambda f(x, y)$ with $\lambda \neq 0$, and any polynomial in this set is a *representative* for the curve. In this book we will usually abbreviate the term 'algebraic curve' to 'curve'. Curves of degree 1, 2, 3, 4, ... are called *lines*, *conics*, *cubics*, *quartics*, ... It is a long established convention that the curve with representative polynomial $f(x, y)$ is referred to as the 'curve' $f(x, y) = 0$. There is no harm in this provided you remember that it is a convention, and not a shorthand for the zero set. Thus $y = x$ and $2y = 2x$ represent the same curve of degree 1.

It is an unfortunate fact of life that when dealing with the simplest possible curves of elementary geometry (such as the lines and standard conics discussed in the next section) the distinction between curves and their zero sets can be blurred without undue consequences. However, as one proceeds into algebraic geometry the relation between the two concepts becomes crucial, and leads to some of the most fundamental results in the subject. The reader is warned, even at this very early stage, to make a crystal clear mental distinction between the concept of a curve, and that of its zero set.

1.2 Introductory Examples

In this section we present a small selection of curves, illustrating some of the general concepts which will occur later. For reasons of space, it is simply not feasible to give an account of even the more significant situations (in the physical sciences, and within pure mathematics itself) where curves arise, as each such situation would demand at least some of the pertinent underlying mathematics to be developed. However, the impatient reader, wishing to see 'real' curves (in the sense of 'real' ale), might like to jump to Section 1.3 which presents some of the curves arising in planar kinematics. A good guiding philosophy is to begin at the beginning (though we will not end at the end) and work with increasing degree. Curves of degree 1 are lines, and play a fundamental role in understanding the geometry of general curves. We will recall their most important attributes via a series of examples. According to the above definition a line has the form $ax + by + c$ with at least one of a, b non-zero.

Example 1.1 Given any two distinct points $p = (p_1, p_2)$, $q = (q_1, q_2)$ in $\mathbb{R}^2$ there is a unique line $ax + by + c$ passing through p, q. We seek scalars a, b, c (not all zero) for which

$$ap_1 + bp_2 + c = 0, \qquad aq_1 + bq_2 + c = 0.$$

Since p, q are distinct, the 2×3 coefficient matrix of these two linear equations in a, b, c has rank 2. By linear algebra it has kernel rank 1, so there is a non-trivial solution (a, b, c), and any other solution is a non-zero scalar multiple of this one. Explicitly, the line joining p, q is given by

$$(p_1 - q_1)(y - p_2) = (p_2 - q_2)(x - p_1).$$

Note this point well: *the equation of a line is determined up to scalar multiples by its zero set.* For higher degree curves that can fail.

Example 1.2 Consider two lines $a_1x + b_1y + c_1$, $a_2x + b_2y + c_2$ in $\mathbb{R}^2$. The intersection points (x, y) are those points which satisfy the linear equations

$$a_1x + b_1y + c_1 = 0, \qquad a_2x + b_2y + c_2 = 0.$$

By linear algebra, when the determinant $\delta = a_1b_2 - a_2b_1 \neq 0$ these equations have a *unique* solution (x, y). Otherwise, there is no solution (*parallel lines*) or a line of solutions (*coincident lines*).

Example 1.3 Let l be a line, and let p be a point not on l. Then there is a unique line m through p parallel to l. Suppose that l has equation $ax + by + c = 0$. It follows from the previous example that the lines m parallel to l are those of the form $ax + by + d = 0$, with d arbitrary. The condition for m to pass through p then determines d uniquely.

Example 1.4 Lines can be parametrized in a natural way. Consider a line $ax + by + c$, and *distinct* points $p = (p_1, p_2)$, $q = (q_1, q_2)$ on the line. Then a brief calculation verifies that any point $p + t(q - p) = (1 - t)p + tq$ also lies on the line. Conversely, we claim that any point $r = (r_1, r_2)$ on the line has the form $r = (1 - t)p + tq$ for some scalar t. Since p, q, r all lie on the line we have

$$\begin{cases} ap_1 + bp_2 + c &= 0 \\ aq_1 + bq_2 + c &= 0 \\ ar_1 + br_2 + c &= 0. \end{cases}$$

That is a linear system of three equations in a, b, c. Since at least one of a, b is non-zero, the system has a non-trivial solution. By linear algebra, the 3×3 matrix of coefficients is singular, so the rows $(p_1, p_2, 1)$, $(q_1, q_2, 1)$, $(r_1, r_2, 1)$ are linearly dependent. However, the first two rows are linearly independent (as p, q are distinct) so the third row is a linear combination of the first two, i.e. $(r_1, r_2, 1) = s(p_1, p_2, 1) + t(q_1, q_2, 1)$ for some scalars s, t. That means $r = sp + tq$ and $1 = s + t$, so $r = (1 - t)p + tq$, as required.

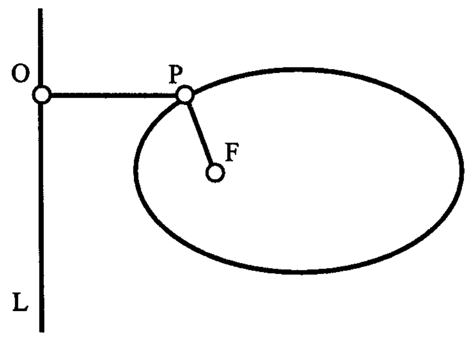

Fig. 1.1. Constructions of standard conics

Note that the parametrization of a line depends on the choice of points p, q. Given two distinct points $p = (p_1, p_2)$, $q = (q_1, q_2)$, the *parametrized* line through p, q is the specific parametric curve given by the above example, namely $x = (1-t)p_1 + tq_1$, $y = (1-t)p_2 + tq_2$.

Example 1.5 Conics will be a recurrent theme in this text. The most familiar conic by far is the circle, defined metrically as the locus of points (x, y) whose distance from a fixed point (a, b) in the plane (the *centre*) takes a constant value $r > 0$ (the *radius*). A circle is thus the zero set of a polynomial $(x-a)^2 + (y-b)^2 = r^2$. The 'standard' parametrization of the circle is $x = a + r\cos t$, $y = b + r\sin t$, but we will meet other parametrizations later.

Example 1.6 The reader has probably met the 'standard conics' of elementary geometry via a metrical construction going back to the classical Greeks. One is given a line L (the *directrix*), a point F (the *focus*) not on L, and a variable point P whose distance from F is proportional to its distance from L. O denotes the unique point on L for which L is perpendicular to the line through O, F. (See Figure 1.1.)

The locus of P is known as a 'parabola', an 'ellipse', or a 'hyperbola' according as the constant of proportionality e (the *eccentricity*) is $= 1$, < 1 or > 1. The fact that P lies on a conic is demonstrated by taking O to be the origin, L to be the y-axis, and the line OF to be the x-axis (with F on the positive axis). Then, setting $F = (2a, 0)$ with $a > 0$, the condition on $P = (x, y)$ is $(x-2a)^2 + y^2 = e^2x^2$, which is indeed a conic. For instance in the case $e = 1$ of a parabola this becomes $4a(a-x) + y^2 = 0$: the translation $x = X + a$, $y = Y$ then yields the *standard parabola* $Y^2 = 4aX$

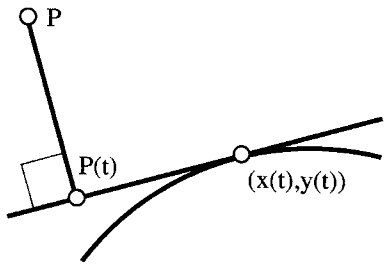

Fig. 1.2. Construction of pedal curves

with focus $F = (a, 0)$ and directrix $X = -a$. In Chapter 4 we will have more to say about the process of reducing polynomials to such 'normal' forms by applying translations, and more generally 'affine mappings'.

Example 1.7 A *regular* parameter of a parametrized curve $x = x(t)$, $y = y(t)$ is a parameter t for which the tangent vector $(x'(t), y'(t))$ is non-zero. The unique line through $P = (x(t), y(t))$ in the direction of the tangent vector is the *tangent line* to the parametrized curve at t, given parametrically as $x = x(t) + \lambda x'(t)$, $y = y(t) + \lambda y'(t)$. It is the line

$$y'(t)x - x'(t)y + \{x'(t)y(t) - x(t)y'(t)\} = 0.$$

Tangent lines play a fundamental role in studying parametrized curves. We will discuss tangent lines to algebraic curves in Chapter 7 and relate them to the concept just introduced for parametrized curves. Numerous interesting constructions are based on the tangent lines to parametrized curves, and give rise to a zoo of interesting curves. One such construction is that of the 'pedal', of considerable importance in geometric optics and kinematics. Suppose we are given a *regular* parametrized curve $x = x(t)$, $y = y(t)$, i.e. one for which every parameter t is regular, and a fixed point $P = (\alpha, \beta)$, the *pedal point*. Then the *pedal* curve of the curve with respect to P is the parametrized curve obtained by associating to the parameter t the projection $P(t)$ of P onto the tangent line at t. (Figure 1.2.)

In practice, given the tangent line, you can write down the line perpendicular to it through P and find the intersection $P(t)$ of the two lines. (Recall from elementary geometry that the lines perpendicular to a given line $ax + by + c = 0$ are the lines of the form $-bx + ay + d = 0$.) Here is a deceptively simple example giving rise to a number of interesting cubics.

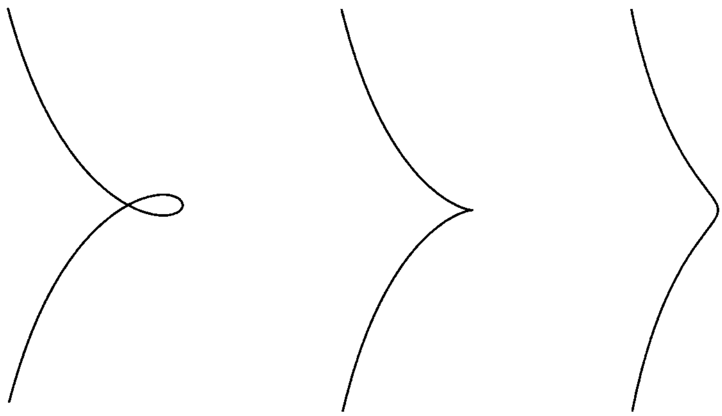

Fig. 1.3. Pedals of a parabola

Example 1.8 Consider the standard parabola $y^2 = 4ax$ with $a > 0$, parametrized as $x = at^2$, $y = 2at$. We will show that the pedal curve with respect to the point $P = (\alpha, 0)$ satisfies the equation of a cubic. The tangent line at t is $x - ty + at^2 = 0$, and the perpendicular line through P is $tx + y - \alpha t = 0$. The parametrized pedal is obtained by setting these expressions equal to zero, and then solving for x, y in terms of t, to obtain

$$x = \frac{(\alpha - a)t^2}{1 + t^2}, \qquad y = \frac{t(\alpha + at^2)}{1 + t^2}.$$

To obtain a polynomial satisfied by the points on the pedal we eliminate t instead, to obtain the cubic $x(x-\alpha)^2 + y^2(a-\alpha+x) = 0$. More precisely we have obtained a *family* of cubics, depending on α. The zero set of some of the pedal curves are illustrated in Figure 1.3.

The first thing to notice is that P always lies on the pedal, and is in some visual sense 'singular'. Thus for $\alpha < 0$ the curve has a loop, which crosses itself at P, for $\alpha = 0$ the loop contracts down to a point, giving a 'cusp' at P, and for $\alpha > 0$ the curve has an isolated point at P. Such 'singular' points play a very basic role in understanding the geometry of a curve, and will be studied in some detail in Chapter 6. The cubic $x^3 + y^2(x + a) = 0$ obtained when $\alpha = 0$ is called the *cissoid of Diocles* after the classical Greek mathematician Diocles, who derived its equation when solving the problem of 'doubling the cube'; Newton discovered a mechanical construction for the cissoid, which we will meet in Section 1.3. $\alpha = 0$ is not the only exceptional value of α. When $\alpha = a$,

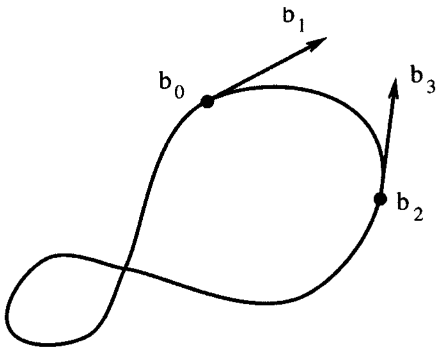

Fig. 1.4. Bézier curve with control points b_0, b_1, b_2, b_3

i.e. the pedal point is the focus F of the parabola, the equation factorizes as $x\{(x-a)^2+y^2\}=0$, whose zero set is simply the y-axis together with the point F. Another exceptional value is $\alpha=-a$, i.e. the pedal point is the point of intersection of the axis and directrix of the parabola: the pedal curve is then known as the *right strophoid*, and is characterized geometrically within the family by the fact that the tangents to the pedal curve at P are perpendicular.

Cubic curves play important roles in numerous areas of mathematics and the physical sciences. An interesting class of naturally parametrized cubics arises in Computer Aided Design (CAD). The idea is as follows. One is given a plane 'curve', for instance part of an artist's visualization of an industrial product, and one seeks a useful mathematical model for this curve which can be handled on a computer. The underlying idea was developed in the late 1950s by two design engineers working for rival French car companies, namely Bézier (working for Renault) and de Casteljau (working for Citröen). A first crude step is to take a sequence of points b_0, b_2, …, b_{2n} on the curve and interpolate a polynomially parametrized curve. However, this process is intrinsically unsatisfactory: as n increases, so the degrees of the polynomials increase, and the interpolating curve may oscillate wildly. The idea is to control this oscillation by specifying the tangent direction at each point. One way of doing this is to associate to each point b_{2k} another point b_{2k+1} and stipulate that the tangent direction of the interpolating curve at b_{2k} should be the direction of the line segment joining b_{2k}, b_{2k+1}. Let us illustrate this for the case of two points b_0, b_2. In that case there are four *control points* b_0, b_1, b_2, b_3 to which is associated the *Bézier curve* defined by

$$B(t) = (1-t)^3 b_0 + 3t(1-t)^2 b_1 + 3t^2(1-t)b_3 + t^3 b_2.$$

Note that $B(0) = b_0$, $B(1) = b_2$, $B'(0) = 3(b_1 - b_0)$, $B'(1) = 3(b_2 - b_3)$ so the curve passes through the points b_0, b_2 and has tangent directions $b_1 - b_0$, $b_2 - b_3$ at those points. (Figure 1.4.) What is not obvious is that *Bézier curves are algebraic*. For the moment we will content ourselves with a numerical example. Later (Example 14.4) when we have a little more algebra available, we will be able to prove this in full generality.

Example 1.9 In the above discussion take $b_0 = (0,0)$, $b_1 = (1/3, 0)$, $b_2 = (2,2)$, $b_3 = (1, 1/3)$ so $B(t) = (t + t^2, t^2 + t^3)$. Write $x = t + t^2$, $y = t^2 + t^3$. Note that $y = tx$. Eliminating t we see that each point (x, y) on the Bézier cubic lies on the cubic curve $x^3 = y(x+y)$. Conversely, we will show that any point (x, y) satisfying the equation $x^3 = y(x+y)$ necessarily has the form $x = t + t^2$, $y = t^2 + t^3 = tx$ for some real number t. Indeed if $x \neq 0$ *define* t by the relation $y = tx$; then, substituting for y we obtain $x = t + t^2$, and hence $y = tx = t^2 + t^3$. Finally, if $x = 0$ then $y = 0$ and we can choose either $t = 0$ or $t = -1$.

Numerous examples of quartic curves arise in the physical sciences. For the moment we will content ourselves with a particularly interesting family of quartics.

Example 1.10 The unit circle $x^2 + y^2 = 1$ is parametrized as $x = \cos t$, $y = \sin t$. We will find the pedal curve with respect to a point $p = (\alpha, 0)$ on the x-axis with $\alpha > 0$. The tangent line at t is $(\cos t)x + (\sin t)y - 1 = 0$, and the perpendicular line through p is $\sin t(x - \alpha) - (\cos t)y = 0$. An equation for the pedal can be found by solving these relations for $\sin t$, $\cos t$, and then substituting in the identity $\cos^2 t + \sin^2 t = 0$. The result is the quartic curve

$$\{x(x-\alpha) + y^2\}^2 = (x-\alpha)^2 + y^2$$

known as a *limacon*. The zero set of the limacon (Figure 1.5) depends on the value of α. The point p always lies on the pedal, and is in some sense 'singular'. (Compare with the pedals of the parabola.) For $\alpha > 1$ the curve has two loops, whilst for $\alpha < 1$ it has just one. The intermediate case $\alpha = 1$ gives rise to a curve with a 'cusp', known as a *cardioid*.

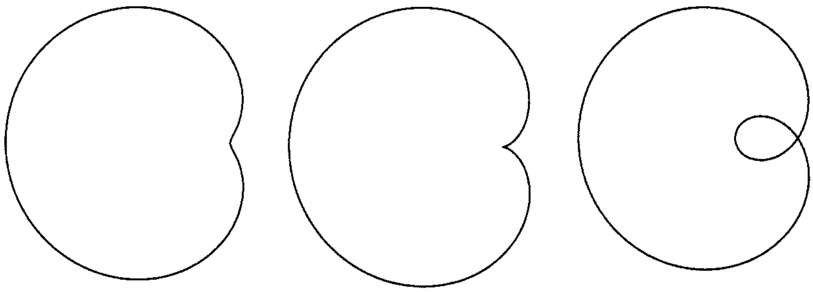

Fig. 1.5. The three forms of a limacon

Exercises

1.2.1 Show that the Bézier curve defined by the points $b_0 = (-9, 0)$, $b_1 = (-9, -1)$, $b_2 = (-6, -2)$, $b_3 = (-8, -2)$ is given parametrically by $x = 3(t^2 - 3)$, $y = t(t^2 - 3)$. By eliminating t from these relations, show that every point lies on the zero set of *Tschirnhausen's cubic* $27y^2 = x^2(x + 9)$.

1.2.2 A parametrized curve is defined by $x(t) = t^2 + t^3$, $y(t) = t^3 + t^4$. Find a polynomial $f(x, y)$ of degree 4 such that $f(x(t), y(t)) = 0$ for all t. (It helps to observe that $y = tx$.) Conversely, show that for any point (x, y) with $f(x, y) = 0$ there exists a real number t with $x = x(t)$, $y = y(t)$. (Again, it helps to observe that you seek a t for which $y = tx$.)

1.2.3 Show that there exists a cubic curve $f(x, y)$ such that every point on the parametrized curve $x(t) = 1 + t^2$, $y(t) = t + t^3$ satisfies the equation $f(x(t), y(t)) = 0$. Conversely, show that for any point (x, y) with $f(x, y) = 0$, *with one exception*, there exists a real number t with $x = x(t)$, $y = y(t)$.

1.2.4 Let $a > 0$, let C be the circle of radius a with centre $(a, 0)$ and let D be the line $x = 2a$. For each line L through the origin O (except the y-axis) let C_L, D_L denote the points where L meets C, D respectively, and let B_L denote the point on the line segment joining O, D_L for which $OB_L = C_L D_L$. Taking t to be the angle between L and the x-axis, find the coordinates of C_L, D_L in terms of t, and hence show that the locus of B_L has the parametric form $x(t) = 2a\sin^2 t$, $y(t) = 2a\sin^2 t \tan t$ with $-\pi/2 < t < \pi/2$. Verify that every point on this parametrized curve lies on the zero set of the cubic $x^3 = (2a - x)y^2$. Conversely, show that any