Chapter 1

DEGREE OF A CURVE

Road Map

The idea of *degree* is a fundamental concept, which will take us several chapters to explore in depth. We begin by explaining what an algebraic curve is, and offer two different definitions of the degree of an algebraic curve. Our job in the next few chapters will be to show that these two different definitions, suitably interpreted, agree.

During our journey of discovery, we will often use elliptic curves as typical examples of algebraic curves. Often, we'll use $y^2 = x^3 - x$ or $y^2 = x^3 + 3x$ as our examples.

1. Greek Mathematics

In this chapter, we will begin exploring the concept of the degree of an *algebraic curve*—that is, a curve that can be defined by polynomial equations. We will see that a circle has degree 2. The ancient Greeks also studied lines and planes, which have degree 1. Euclid limited himself to a straightedge and compass, which can create curves only of degrees 1 and 2. A "primer" of these results may be found in the *Elements* (Euclid, 1956). Because 1 and 2 are the lowest degrees, the Greeks were very successful in this part of algebraic geometry. (Of course, they thought only of geometry, not of algebra.)

Greek mathematicians also invented methods that constructed higher degree curves, and even nonalgebraic curves, such as spirals. (The latter cannot be defined using polynomial equations.) They were aware that

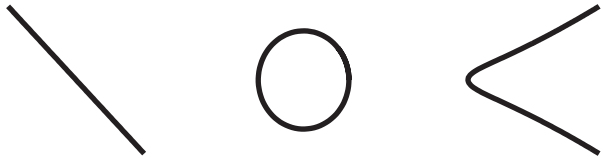

Figure 1.1. Three curves

these tools enabled them to go beyond what they could do with straightedge and compass. In particular, they solved the problems of *doubling the cube* and *trisecting angles*. Both of these are problems of degree 3, the same degree as the elliptic curves that are the main subject of this book. Doubling the cube requires solving the equation $x^3 = 2$, which is clearly degree 3. Trisecting an angle involves finding the intersection of a circle and a hyperbola, which also turns out to be equivalent to solving an equation of degree 3. See Thomas (1980, pp. 256–261, pp. 352–357, and the footnotes) and Heath (1981, pp. 220–270) for details of these constructions. Squaring the circle is beyond any tool that can construct only algebraic curves; the ultimate reason is that π is not the root of any polynomial with integer coefficients.

As in the previous two paragraphs, we will see that the degree is a useful way of arranging algebraic and geometric objects in a hierarchy. Often, the degree coincides with the level of difficulty in understanding them.

2. Degree

We have a feeling that some shapes are simpler than others. For example, a line is simpler than a circle, and a circle is simpler than a cubic curve; see figure 1.1

You might argue as to whether a cubic curve is simpler than a sine wave or not. Once algebra has been developed, we can follow the lead of French mathematician René Descartes (1596–1650), and try writing down algebraic equations whose solution sets yield the curves in which we are interested. For example, the line, circle, and cubic curve in figure 1.1 have equations $x + y = 0$, $x^2 + y^2 = 1$, and $y^2 = x^3 - x - 1$, respectively. On the other hand, as we will see, the sine curve cannot be described by an algebraic equation.

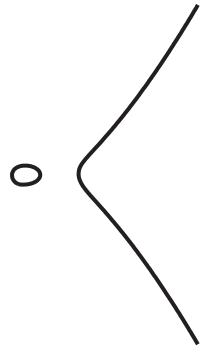

Figure 1.2. $y^2 = x^3 - x$

Our typical curve with degree 3 has the equation $y^2 = x^3 - x$. As we can see in figure 1.2, the graph of this equation has two pieces.

We can extend the concept of equations to higher dimensions also. For example a sphere of radius r can be described by the equation

$$x^2 + y^2 + z^2 = r^2. \tag{1.1}$$

A certain line in 3-dimensional space is described by the pair of simultaneous equations

$$\begin{cases} x + y + z = 5 \\ \quad x - z = 0. \end{cases} \tag{1.2}$$

The "solution set" to a system of simultaneous equations is the set of all ways that we can assign numbers to the variables and make all the equations in the system true at the same time. For example, in the equation of the sphere (which is a "system of simultaneous equations" containing only one equation), the solution set is the set of all triples of the form

$$(x, y, z) = (a, b, \pm\sqrt{r^2 - a^2 - b^2}).$$

This means: To get a single element of the solution set, you pick any two numbers a and b, and you set $x = a$, $y = b$, and $z = \sqrt{r^2 - a^2 - b^2}$ or $z = -\sqrt{r^2 - a^2 - b^2}$. (If you don't want to use complex numbers, and you

only want to look at the "real" sphere, then you should make sure that $a^2 + b^2 \leq r^2$.)

Similarly, the solution set to the pair of linear equations in (1.2) can be described as the set of all triples

$$(x, y, z) = (t, 5 - 2t, t),$$

where t can be any number.

As for our prototypical cubic curve $y^2 = x^3 - x$, we see that its solution set includes $(0, 0)$, $(1, 0)$, and $(-1, 0)$, but it is difficult to see what the entire set of solutions is.

In this book, we will consider mostly systems of *algebraic equations*. That means by definition that both sides of the equation have to be polynomial expressions in the variables. The solution sets to such systems are called "algebraic varieties." The study of algebraic geometry, which was initiated by Descartes, is the study of these solution sets. Since we can restrict our attention to solutions that are integers, or rational numbers, if we want to, a large chunk of number theory also falls under the rubric of algebraic geometry.

Some definitions:

- A *polynomial* in one or several variables is an algebraic expression that involves only addition, subtraction, and multiplication of the constants and variables. (Division by a variable is not permitted.) Therefore, each variable might be raised to a positive integral power, but not a negative or a fractional power.
- A *monomial* is a polynomial involving only multiplication, but no addition or subtraction.
- The *degree* of a monomial is the sum of the powers of the variables that occur in the monomial. The degree of the zero monomial is undefined. For example, the monomial $3xy^2z^5$ has degree 8, because $1 + 2 + 5 = 8$.
- The *degree* of a polynomial is the largest degree of any of the monomials in the polynomial. We assume that the polynomial is written without any terms that can be combined or cancelled. For example, the polynomial $3xy^2z^5 + 2x^3z^3 + xyz + 5$ has degree 8, because the other terms have degrees 6, 3, and 0, which are all

smaller than 8. The polynomial $y^5 - x^3 - y^5 + 11xy$ has degree 3, because we first must cancel the two y^5-terms before computing the degree. The polynomial $2x^7y^2 - x^7y^2 + xy - 1 - x^7y^2$ has degree 2, because it is really the polynomial $xy - 1$.

- If we have an algebraic variety defined by a system of equations of the form "some polynomial = some other polynomial," we say that the variety has degree d if the largest degree of any polynomial appearing in the system of equations is d. Again, we assume that the system of equations cannot be simplified into an equivalent system of equations with smaller degree.

EXERCISE: What is the degree of the equation for a sphere in (1.1)? What is the degree of the system of equations for the line in (1.2)?

SOLUTION: The degree of a sphere is 2. The degree of a line is 1.

Now suppose we have a geometric curve or shape. How can we tell what its degree is if we are not given its equation(s)? Or maybe it isn't given by any system of algebraic equations? We're not going to give a general answer to this question in this book, but we will explain the basic idea in the case of single equations.

Let's start by recalling another definition.

DEFINITION: Suppose that $p(x)$ is a polynomial. If b is a number so that $p(b) = 0$, then b is a *root* of the polynomial $p(x)$.

The basic idea is that we will use lines as probes to tell us the degree of a polynomial. This method is based on a very important fact about polynomials: Suppose you have a polynomial

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

where $a_n \neq 0$, so that $f(x)$ has degree n, where $n \geq 1$. Suppose that b is a root of $f(x)$. Then if you divide $x - b$ into $f(x)$, it will go in evenly, without remainder.

For example, if $f(x) = x^3 - x - 6$, you can check that $f(2) = 0$. Now divide $x - 2$ into $x^3 - x - 6$ using long division:

$$\begin{array}{r} x^2 + 2x + 3 \\ \hline x-2\,)\,x^3 \qquad - x - 6 \\ \underline{x^3 - 2x^2} \\ \qquad 2x^2 - x \\ \qquad \underline{2x^2 - 4x} \\ \qquad\qquad 3x - 6 \\ \qquad\qquad \underline{3x - 6} \\ \qquad\qquad\quad 0 \end{array}$$

We get a quotient of $x^2 + 2x + 3$, without remainder. Another way to say this is that $x^3 - x - 6 = (x^2 + 2x + 3)(x - 2)$.

We can prove our assertion in general: Suppose you divide $x - b$ into $f(x)$ and get the quotient $q(x)$ with remainder r. The remainder r in polynomial division is always a polynomial of degree less than the divisor. The divisor $x - b$ has degree 1, so the remainder must have degree 0. In other words, r is some *number*. We then take the true statement $f(x) = q(x)(x - b) + r$, and set $x = b$ to get $f(b) = q(b)(b - b) + r$. Since $f(b) = 0$ and $b - b = 0$, we deduce that $r = 0$. So there was no remainder, as we claimed.

Now this little fact has the momentous implication that the number of roots of a polynomial $f(x)$ cannot be greater than its degree. Why not? Suppose $f(x)$ had the roots $b_1, \ldots, b_k$, all different from each other. Then you can keep factoring out the various $x - b_i$'s and get that

$$f(x) = q(x)(x - b_1)(x - b_2) \cdots (x - b_k)$$

for some nonzero polynomial $q(x)$. Now multiply out all those factors on the right-hand side. The highest power of x you get must be at least k. Since the highest power of x on the left-hand side is the degree of $f(x)$, we know that k is no greater than the degree of $f(x)$.

Next, we interpret geometrically what it means for b to be a root of the polynomial $f(x)$ of degree n. Look at the graph G of $y = f(x)$. It is a picture in the Cartesian plane of all pairs (x, y) where $y = f(x)$. Now look at the

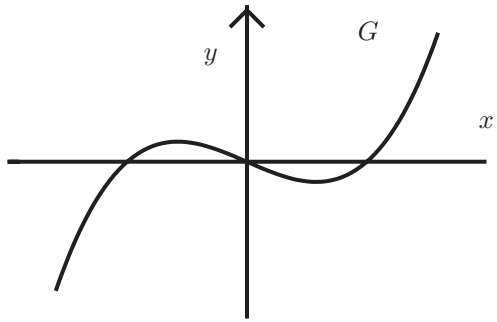

Figure 1.3. $y = x(x-1)(x+1)$

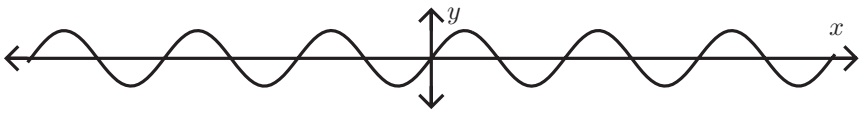

Figure 1.4. $y = \sin x$

graph of $y = 0$. It is a horizontal straight line L consisting of all pairs (x, y) where $y = 0$ and x can be anything. OK, now *look at the intersection of the graph of* $f(x)$ *and the line* L. Which points are in the intersection? They are exactly the points $(b, 0)$ where $0 = f(b)$. That means that the x-coordinates of the points of intersection, the b's, are exactly the roots of $f(x)$. There can be at most n of these roots. This means that the line L can hit the curve G in at most n points. Figure 1.3 contains an example using the function $x^3 - x = x(x-1)(x+1)$, which is the right-hand side of our continuing example cubic curve $y^2 = x^3 - x$.

On the other hand, let G be the graph of $y = \sin(x)$. As we can see in figure 1.4, the line L hits G in infinitely many points (namely $(b, 0)$, where b is any integral multiple of π). Therefore, G cannot be the graph of any polynomial function $f(x)$. So the sine wave is not an algebraic curve.

For another example, let H be the graph of $x^2 + y^2 = 1$ (a circle). As we can see in figure 1.5, the line L_1 hits the graph H in 2 points, the tangent line L_2 hits H in one point, and the line L_3 hits H in zero points.

In a case like this, we take the *maximum* number of points of intersection, and call it the *geometric degree* of the curve. So the geometric degree of the circle H is 2. In subsequent chapters, we will discuss how

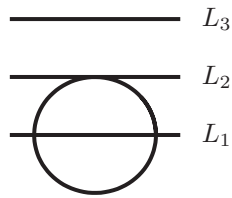

Figure 1.5. Three lines intersecting a circle

mathematicians dealt with the initially unpleasant but ultimately very productive fact that the number of intersection points is not always the same, but depends on the line we choose as probe. The desire to *force* this number to be constant, independent of the probing line, turned out to be a fruitful source of new mathematics, as we will see.

Similarly to the preceding example, we can look at a sphere, for instance the graph of $x^2 + y^2 + z^2 = 1$. Again, a line may intersect the sphere in 2, 1, or 0 points. We say the geometric degree of the sphere is 2.

To repeat, we provisionally *define* the *geometric degree* of a geometric object to be the maximum number of points of intersection of any line with the object. Suppose we have any polynomial $f(x, y, \dots)$ in any number of variables. It looks very likely from our examples so far that the geometric degree of the graph of $f(x, y, \dots) = 0$ should equal the degree of f.

3. Parametric Equations

Let's investigate the possibility that the geometric degree of a curve according to the definition of the probing line will equal the degree of the equation of the curve. To do so, we will have to write our probing lines in parametric form.

A *parameter* is an extra variable. A system of equations in parametric form is one where all of the other variables are set equal to functions of the parameter(s). For example, when we describe curves in the xy-plane, we will often use t as a parameter, and write $x = f(t)$ and $y = g(t)$.

We will only use parametric form with a single parameter. A single parameter will suffice to describe curves, and particularly lines. We use parametric form because that method of describing curves makes it easy

to find intersection points of two curves by substituting one equation into another, as you will see. Other reasons to use parametric form will also soon become apparent.

Here's an important assumption: *We will always parametrize lines linearly*. What does this assumption mean? Whenever we parametrize a line, we will always set $x = at + b$ and $y = ct + e$, where a and c cannot both be 0. It is possible to pick other more complicated ways to parametrize a line, and we want to rule them out.

We can describe a line or a curve in the plane in two different ways. We can give an equation for it, such as $y = mx + b$. Here, m and b are fixed numbers. You probably remember that m is the slope of the line and b is its y-intercept. A problem with this form is that a vertical line cannot be described this way, because it has "infinite" slope. The equation of a vertical line is $x = c$, for some constant c. We can include all lines in a single formula by using the equation $ax + by = c$ for constants a, b, and c. Depending on what values we assign to a, b, and c, we get the various possible lines.

The other way to describe a line or a curve in the plane is to use a parameter. A good way to think about this is to pretend that the line or curve is being described by a moving point. At each moment of time, say at time t, the moving point is at a particular position in the plane, say at the point $P(t)$. We can write down the coordinates of $P(t) = (x(t), y(t))$. In this way we get two functions of t, namely $x(t)$ and $y(t)$. These two functions describe the line or curve "parametrically," where t is the parameter. The line or curve is the set of all the points $(x(t), y(t))$, as t ranges over a specified set of values.

Parametric descriptions are very natural to physicists. They think of the point moving in time, like a planet moving around the sun. The set of all points successively occupied by the moving point is its "orbit."

For example, the line with equation $y = mx + b$ can be expressed parametrically by the pair of equations $x = t$ and $y = mt + b$. The parametric description may seem redundant: We used two equations where formerly we needed only one. Each kind of representation has its advantages and disadvantages.

One advantage of parametric representation is when we go to higher dimensions. Suppose we have a curve in 6-dimensional space. Then we would need five (or perhaps more) equations in six variables to describe

it. But since a line or curve is intrinsically only 1-dimensional, we really should be able to describe it with one independent variable. That's what the parametric representation does for us: We have one variable t and then 6 equations of the form $x_i = f_i(t)$ for each of the coordinates $x_1, \ldots, x_6$ in the 6-dimensional space in which the curve lies.

As we already mentioned, a second advantage of the parametric form is that it gives us a very clear way to investigate the intersection of the line or curve with a geometric object given in terms of equations.

Some curves are easy to express in either form. Consider, for example, the curve with equation $y^2 = x^3$, which can be seen in figure 3.7. This curve can be expressed parametrically with the pair of equations $x = t^2$, $y = t^3$.

On the other hand, the curve defined by the equation $x^4 + 3x^3y + 17y^2 - 5xy - y^7 = 0$ is pretty hard to express parametrically in any explicit way. Conversely, it's hard to find an equation that defines the parametric curve $x = t^5 - t + 1$, $y = e^t + \cos(t)$ as t runs over all real numbers.

For future use, we pause and describe how to parametrize a line in the xy-plane. If a, b, c, and e are any numbers, then the pair of equations

$$x = at + b$$

$$y = ct + e$$

parametrizes a line as long as a or c (or both) are nonzero.

Conversely, any line in the xy-plane can be parametrized in this way. In particular: If the line is not vertical, it can be described with an equation of the form $y = mx + b$, and then the pair of equations

$$x = t$$

$$y = mt + b$$

parametrizes the same line. If the line is vertical, of the form $x = e$, then the pair of equations

$$x = e$$

$$y = t$$

gives a parametrization.

4. Our Two Definitions of Degree Clash

First, let's look at a simple example to show that the degree of an equation doesn't always equal the geometric degree of the curve it defines. In this section, we only consider real, and not complex, numbers. Let K be the graph of $x^2 + y^2 + 1 = 0$. Because squares of real numbers cannot be negative, K is the empty set. Our definition of the probing line would tell us that K would have geometric degree zero. But the polynomial defining K has degree 2.

If you object that we needn't consider empty curves consisting of no points, we can alter this example as follows: Let M be the graph of $x^2 + y^2 = 0$. Now M consists of a single point, the origin: $x = 0$, $y = 0$. By our definition of the probing line, M would have geometric degree 1. But the polynomial defining M has degree 2.

You may still object: M isn't a "curve"—it's got only 1 point. But we can beef up this example: Let N be the graph of $(x - y)(x^2 + y^2) = 0$. Now N consists of the 45°-line, given parametrically by $x = t$, $y = t$. By our definition of the probing line, N would have geometric degree 1. But the polynomial defining N has degree 3.

Another type of example would be the curve defined by $(x - y)^2 = 0$. This is again the 45°-line, but the degree of the equation is 2, not 1.

There is one important observation we can make at this point: If a curve is given by an equation of degree d, *any* probing line will intersect it in *at most* d points. Let's see this by looking at our continuing example. Suppose the curve E is given by the equation $y^2 = x^3 - x$. Take the probing line L given by $x = at + b$, $y = ct + e$ for some real constants a, b, c, and e. As we already mentioned, any line in the plane can be parametrized this way for some choice of a, b, c, and e.

When you substitute the values for x and y given by the parametrization into the equation for E, you will get an equation that has to be satisfied by any parameter value for t corresponding to a point of intersection. If we do the substitution, we get

$$(ct + e)^2 = (at + b)^3 - (at + b).$$

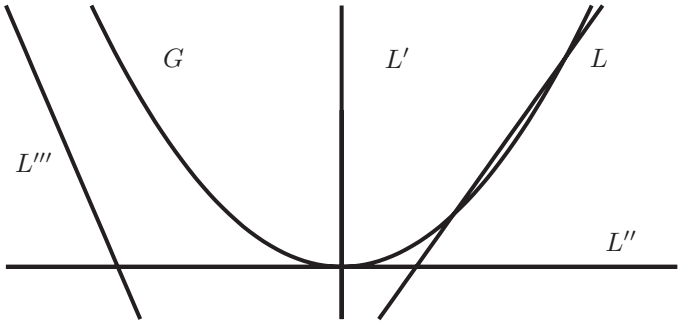

Figure 1.6. Four lines and a parabola

If we expand using the binomial theorem and regroup terms, we obtain the equation

$$a^3t^3 + (3a^2b - c^2)t^2 + (3b^2a - a - 2ec)t + b^3 - b - e^2 = 0$$

We see that the equation will have degree *at most* 3, no matter what a, b, c, and e are. Therefore, it will have at most 3 roots. So at most 3 values of t can yield intersection points of E and L.

Now this lack of definiteness as to the number of intersection points leads to a serious problem with our probing line definition of degree. Let us suppose we have a curve C and we choose a probing line L and we get 5 points in the intersection of C and L. How do we know 5 is the maximum we can get? Maybe a different line L' will yield 6 or more points in the intersection of C and L'. How will we know when to stop probing?

In fact, we can become greedy. We could hope to redefine the geometric degree of C to be the number of points in the intersection of C and L *no matter what line L we pick*! This may sound like a tall order, but if we could do it, we'd have a beautiful definition of degree. It wouldn't matter what probe we choose. Pick any one and count the intersection points.

It may seem like this is hopeless. But let's look at an example. Although it is very simple, this example will show all the problems in our greedy approach to the concept of degree that we will solve in the following three chapters. When we solve them, by redefining the concept of "intersection point" in an algebraically cogent way, our hope will have come true!

Here is the example, which is illustrated in figure 1.6. Let G be the graph of the parabola $y = x^2$. We consider various different probing lines. For

example, let L be the line given parametrically by $x = t$, $y = 3t - 2$. The intersection of L and G will occur at points on the line with parameter value t exactly when $3t - 2 = t^2$. This yields the quadratic equation $t^2 - 3t + 2 = 0$, which has the two solutions $t = 1$ and $t = 2$. The two points of the intersection are $(1, 3 \cdot 1 - 2) = (1, 1)$ and $(2, 3 \cdot 2 - 2) = (2, 4)$. This is the optimal case: We get 2 points of intersection, the most possible for an equation of degree 2.

Now look at the probing line L', given parametrically by $x = 0$, $y = t$. This is the vertical line otherwise known as the y-axis. When we plug these values into the equation for the parabola we get $t = 0^2$, which has only one solution: $t = 0$, corresponding to the single point of intersection $(0, 0)$.

The horizontal probing line L'' given by $x = t$, $y = 0$ doesn't fare any better: Plugging in we get $0 = t^2$, which again has only one solution: $t = 0$, corresponding to the single point of intersection $(0, 0)$.

Finally, look at the probing line L''', given parametrically by $x = t$, $y = -5t - 10$. When we plug these values into the equation for the parabola we get $-5t - 10 = t^2$ or equivalently, $t^2 + 5t + 10 = 0$. Using the quadratic formula, we see that since the discriminant $25 - 40 < 0$, there are *no* solutions for t and hence no points of intersection.

Thus, in the simple case of a parabola, we have some probing lines that meet the parabola in 2 points, others that meet the parabola in only 1 point, and others that don't intersect it at all. Yet the equation of a parabola has degree 2. After we finish the next 3 chapters, we will be able to say that *any* probing line intersects the parabola in 2 points, after we have suitably *redefined* the concept of "intersection."

The constructions we will have to make to find a suitable redefinition of "intersection" will be crucial later for our understanding of elliptic curves and so of the Birch–Swinnerton-Dyer Conjecture.