Algebraic Curves

Jonny Evans

March 4, 2024

0 Overview

The origins of the study of geometry are lost in the mists of time. As long ago as the third century BC, Apollonius wrote a tract on the theory of conic sections, building on centuries of accumulated geometrical knowledge. In the 17th century, the idea of writing down equations to describe geometrical objects developed in the work of Descartes and Fermat; this "algebraic" way of handling geometry is now known as *algebraic geometry*, and allowed for a flourishing interaction with analysis, number theory, and algebra which continues to this day.

By the early 20th century, an influential Italian school of algebraic geometers were making enormous progress, but as their arguments became more and more unwieldy and based on intuition, cracks started to appear: in 1946, Severi gave a proof that a projective sextic surface can have at most 52 nodes, but you only have to look at the Barth sextic, with its 65 nodes, to see that something had gone wrong. In the mid-to-late 20th century, algebraic geometry underwent a reformation, and the foundations of the subject were rewritten carefully using the language of commutative algebra, and the notion of a "sheaf" of functions. This foundational rigour has unfortunately raised the conceptual bar of entry to the subject, and the conventional view is that the modern working algebraic geometer needs to master a monolithic amount of material and ideas before they get to study the kinds of fun things that the Italians were playing with over a century ago (one should minimally "read Hartshorne and do all the exercises").

This makes it hard to teach a first course in algebraic geometry, because there is a balance to be struck between giving an idea of what the subject is actually about (why its questions are natural, or where they come from) and giving the acolyte a sufficient grounding in the basics that they could go on and become an algebraic geometer. When I took a first course in algebraic geometry, it emphasised the latter: I learned very little.

What follows is a zeroth course in algebraic geometry. I will focus on communicating the basic ideas and goals of the subject, with an emphasis on examples and calculation. Later we will make contact with some of the ideas from algebra which form the foundations of the subject (rings, ideals, local rings, etc), but if you are planning on becoming a card-carrying[1] algebraic geometer, you will still need to do a first course. My hope is that this zeroth course will help you to get the most out of that.

[1] Sadly, they don't actually give out cards for algebraic geometers.

1 Algebraic varieties

Let[2] k be a field[3]. Mostly, we will be concerned with the cases where k is one of $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, but any field will do. We will write $k[x_1, \ldots, x_n]$ for the ring of polynomials in the variables $x_1, \ldots, x_n$ with coefficients in k. For example:

$$\pi x^2 - iy \in \mathbb{C}[x, y], \quad \frac{2}{3}x^3 y + xy^2 + 27z^2 + 1 \in \mathbb{Q}[x, y, z].$$

Definition 1.1. Let $f_1, \ldots, f_m \in k[x_1, \ldots, x_n]$ be polynomials. Define

$$\mathbb{V}_k(f_1, \ldots, f_m) := \{(a_1, \ldots, a_n) \in k^n \,:\, f_1(a_1, \ldots, a_n) = \cdots = f_m(a_1, \ldots, a_n) = 0\}.$$

This is called the *affine algebraic set cut out* (or *defined*) by these polynomials. We will often write $\{f = 0\}$ or $\mathbb{V}(f)$ instead of $\mathbb{V}_k(f)$.

Remark 1.2. I will sometimes use the word *variety* instead of affine algebraic set. In most modern treatments, varieties are a special kind of affine algebraic set (having one irreducible component). I feel that this abuse of language is not too bad if you're aware of it: you should imagine that I'm giving you a course on primates and calling them all monkeys. Once we've discussed irreducibility, you can go back through the notes and identify everywhere I wrongly labelled something a variety.

Example 1.3. Perhaps the simplest algebraic set is the *affine space*

$$\mathbb{A}^n(k) := \mathbb{V}_k(0).$$

In other words, this is the set cut out by the zero polynomial! Since the zero polynomial vanishes everywhere, $\mathbb{A}^n(k)$ is a fancy name for k^n.

Example 1.4. The unit circle $S \subset \mathbb{R}^2$ is a variety: it is the set of points (x, y) satisfying $x^2 + y^2 = 1$, so

$$S = \mathbb{V}_{\mathbb{R}}(x^2 + y^2 - 1).$$

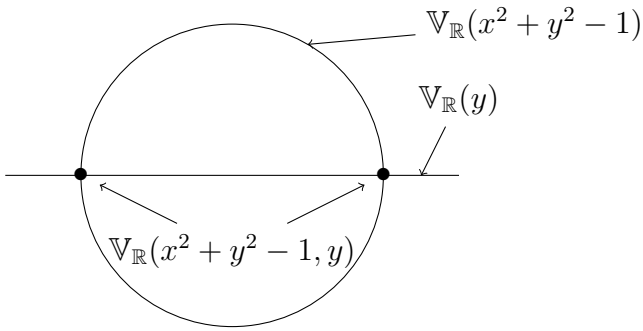
This example is a *plane algebraic curve* (*curve* for short), that is an affine algebraic set cut out by a single equation in two variables (where the equation is not just $0 = 0$). We will focus almost exclusively on plane algebraic curves in this course because there is already a lot to say about them, but we will necessarily come across other affine algebraic sets too, when we take intersections between curves.

Example 1.5. Let $f_1(x, y) = x^2 + y^2 - 1$ and $f_2(x, y) = y$. We've seen that $\mathbb{V}_{\mathbb{R}}(f_1)$ is the unit circle, and $\mathbb{V}_{\mathbb{R}}(f_2) = \{(x, 0) \,:\, x \in \mathbb{R}\}$ is the x-axis. The algebraic variety $\mathbb{V}_{\mathbb{R}}(f_1, f_2)$ comes from imposing both constraints $f_1 = f_2 = 0$, i.e. it consists of points which lie on the intersection of the unit circle and the x-axis:

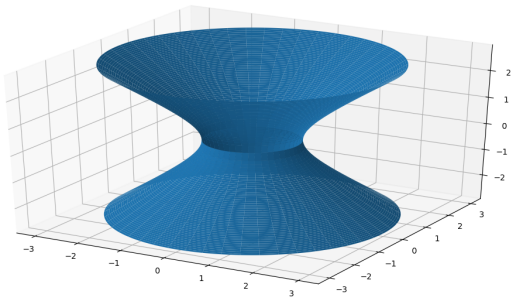
$$\mathbb{V}_{\mathbb{R}}(f_1, f_2) = \{(-1, 0),\, (1, 0)\}.$$

[2]The letter k here stands for "Körper", the German word for field. Much of the foundational work on algebra in the early twentieth century was done by German-speakers like Hilbert and Noether.

[3]In case you don't remember: a field is a system of numbers in which you can add, subtract, multiply and divide, as long as you don't divide by zero. All the usual "laws" are assumed work, like distributivity, associativity, commutativity...

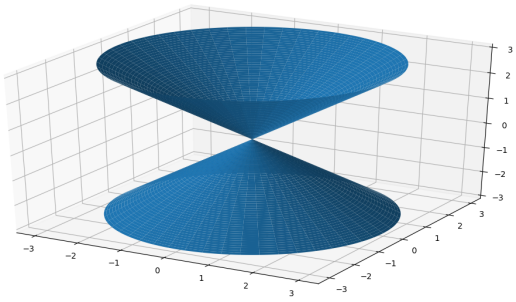


Example 1.6. Let $f(x,y,z) = x^2+y^2-z^2-1$. The algebraic variety $\mathbb{V}_{\mathbb{R}}(f)$ is an *algebraic surface* in $\mathbb{R}^3$



We can think of this as a family of curves (parametrised by z). At a given height z, we have the circle $x^2+y^2 = 1+z^2$ of radius $\sqrt{1+z^2}$.

Example 1.7. Let $f(x,y,z) = x^2+y^2-z^2$. The algebraic variety $\mathbb{V}_{\mathbb{R}}(f)$ is another algebraic surface in $\mathbb{R}^3$; it consists of two pieces which are the graphs of $z = \pm\sqrt{x^2+y^2}$:



These two pieces meet at a *singular point* (the origin). We can "smooth" this surface by deforming the polynomial slightly, for example $\mathbb{V}_{\mathbb{R}}(x^2+y^2-z^2-\epsilon)$. The picture for $\epsilon = 1$ was given in the previous example: it looks a little like a wormhole. If we decrease ϵ then the neck of the wormhole shrinks and eventually pinches to become a singularity when $\epsilon = 0$.

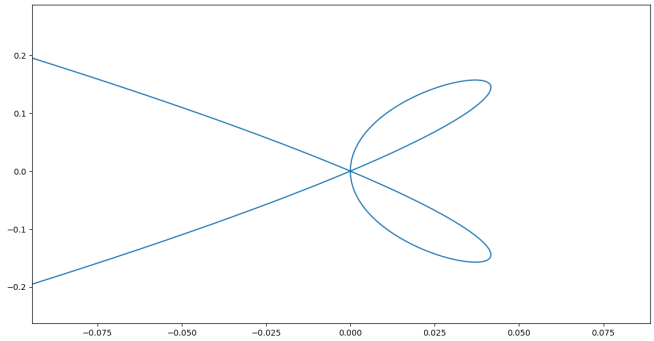
These pictures were produced by a computer, but it is not hard to sketch these surfaces by hand. In both cases, the equation is unchanged if you rotate around the z-axis because x and y enter in the equation only through the combination x^2+y^2, which is the squared

radius in the xy-plane. So these are surfaces of revolution: they are obtained by taking a curve in the xz-plane and rotating it around the z-axis. The curve we rotate is cut out by the equation we get by setting $y = 0$. In the first case, the curve we rotate is $z^2 = x^2 + 1$, which is a hyperbola; in the second case, it is $z^2 = x^2$ (i.e. $z = \pm x$, or two lines with slopes ± 1).

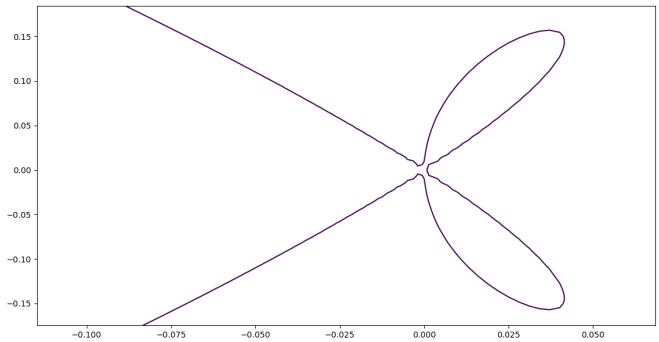
In general, if someone just hands you a system of polynomials, it can be extremely difficult to draw the algebraic set they define, even using a computer[4]. This is because the set is being given *implicitly*: it is easy to check if a given point is in the algebraic set by substituting its coordinates into the equation and checking that the equation is satisfied, but it is hard to find *all* the points that satisfy the equation.

Example 1.8. Consider the curve in $\mathbb{R}^2$ cut out by the polynomial

$$xy^2 - 6x^3 = y^4.$$



If I just naively ask my computer to plot this directly from the equation, I get the following:



In Section 3, we will see how I managed to get the nicer picture (by finding a rational parametrisation of the curve) but for now, let's try and read off some of the coarse features of the curve from the equation. For example:

- The equation is unchanged if we replace y by $-y$, so the curve will be symmetric under reflection in the x-axis.

[4]For example, computers struggle to plot surfaces near singular points.

- If we think of $y^4 - xy^2 + 6x^3 = 0$ as a quadratic equation in y^2 then we can solve it as $y^2 = \frac{x \pm \sqrt{x^2 - 24x^3}}{2} = \frac{x}{2}(1 \pm \sqrt{1 - 24x})$. This has real solutions if and only if $x \leq 1/24 \approx 0.41667$, which is why the curve lies to the left of this value.
- By a more careful analysis, we can see that each x-value has 2, 4 or 0 possible y-values according to whether $x < 0$, $x \in (0, 1/24)$, $x > 1/24$. Something funny is happening at $x = 0$, where there is only one possible y-value ($y = 0$).
- If x is large and negative then $(x/2)(1 - \sqrt{1 - 24x}) \approx |x|^{3/2}\sqrt{6}$, so the two unbounded ends of the curve are asymptotic to $y = \pm\sqrt[4]{6}|x|^{3/4}$.

Most algebraic varieties are not amenable to this kind of detailed analysis: the equations are just too complicated. We will develop techniques that will help us to:

- Identify and understand singular points.
- Study the "geometry at infinity" (like the asymptotes in this example).
- Prove that certain configurations of intersections or singularities are impossible.

Remark 1.9. Most of the time, we will work over the complex numbers. This will make it much harder to draw accurate pictures, but it will have the advantage that we don't need to indulge in the kind of case analysis we made for Example 1.8, worrying about whether a quantity is positive before we take its square root. If you are really interested in $\mathbb{V}_{\mathbb{R}}(f_1, \ldots, f_m)$, one approach is to first study the corresponding variety over $\mathbb{C}$, that is $\mathbb{V}_{\mathbb{C}}(f_1, \ldots, f_m)$, and then observe that this complex variety has a complex conjugation map whose fixed point set is $\mathbb{V}_{\mathbb{R}}(f_1, \ldots, f_n)$. More generally, if k is a field, $\bar{k}$ is an algebraic closure of k, and $f_1, \ldots, f_m \in k[x_1, \ldots, x_n]$ are polynomials, then $\mathbb{V}_k(f_1, \ldots, f_m)$ is the fixed locus for the action of the Galois group $Gal(\bar{k}, k)$ on $\mathbb{V}_{\bar{k}}(f_1, \ldots, f_m)$, and the latter variety is easier to understand (even if the Galois action is tricky). We will occasionally discuss non-algebraically closed fields like $\mathbb{R}$ (so we can draw pictures) and $\mathbb{Q}$ (to illustrate how all of this connects with number theory).

Definition 1.10 (K-points)**.** If k is a subfield of K and $f_1, \ldots, f_m \in k[x_1, \ldots, x_n]$ then a K-point of $\mathbb{V}_k(f_1, \ldots, f_m)$ is a point of $\mathbb{V}_K(f_1, \ldots, f_m)$, and a k-point of $\mathbb{V}_K(f_1, \ldots, f_m)$ is a point of $\mathbb{V}_k(f_1, \ldots, f_m)$.

Example 1.11. i is a $\mathbb{C}$-point of $\{x^2 + 1 = 0\}$: this variety has no $\mathbb{R}$-points, even though it's defined over $\mathbb{R}$.

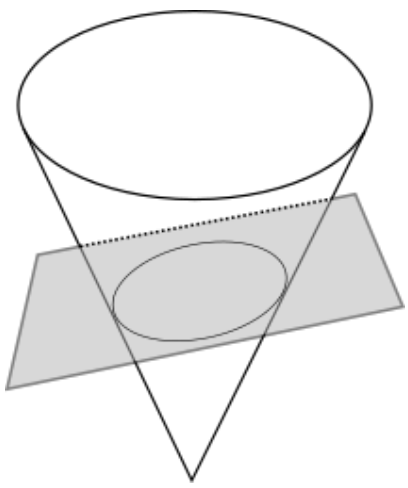
Example 1.12. $(1, 0)$ is a $\mathbb{Q}$-point of $x^2 + y^2 = 1$: are there any other $\mathbb{Q}$-points? How many? We will return to these questions next week.

2 Curves, degree-by-degree

2.1 Conic sections

As I mentioned, *conic sections* are curves which have been studied since antiquity. Nowadays, we think of them as plane curves cut out by a quadratic equation, but how did Apollonius think about them before the idea of coordinates/variables/equations had been developed?

You can describe a conic purely geometrically in the following way. Let S be a circle in the plane, and translate the plane vertically upwards so that it sits at height 1. Let C be the *cone* on this circle: that is the set traced out by the lines connecting the origin to points of the circle. A *section* of C is obtained by taking a plane that misses the origin and seeing where it slices C. For example, if you take the plane $z = 1$ then you recover the original circle. However, if you take the plane $x = 1$ then you get a hyperbola, and if you take the plane $z = y + 1$ then you get a parabola. The figure below shows how you could get an ellipse.

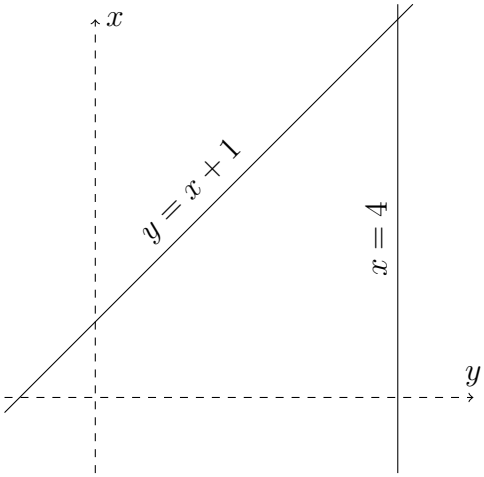


To see how to recast this in terms of equations, we first observe that C is cut out by the equation $x^2 + y^2 = z^2$. You should think of this equation as saying that the intersection of C with $z = r$ is a circle of radius r. Now if we further impose the equation of the plane (e.g. $x = 1$) we get a quadratic equation in two variables (e.g. $1 + y^2 = z^2$) which then defines the conic section (in this example a hyperbola).

We will return to this picture of slicing cones later: it will lead us to the theory of *projective varieties*, and a rigorous way of handling "points at infinity". But for now, let's work degree-by-degree and see what kinds of curves we can get.

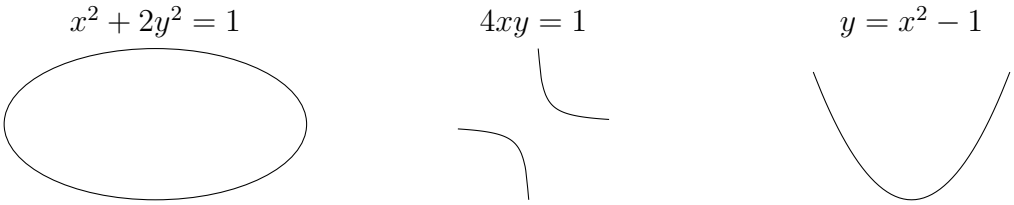
2.2 Degree 1

Recall that the *degree* of a polynomial is the total exponent of the highest order term. For example: $x^2 + y - 1$ has degree 2, $xy^4 - xyz$ has degree 5, and 67 has degree zero. The most general polynomial of degree 1 in two variables is therefore $ax + by + c$.



If our curve is cut out by an equation of degree 1 then it is a line. This is why equations of degree 1 are called "linear"[5]. For example, if the equation is $y = x + 1$ then we get the line with slope 1 passing through $(0, 1)$. If the equation is $x = 4$ then the line is vertical and hits the x-axis at $(4, 0)$.

2.3 Degree 2



If our curve is cut out by an equation of degree 2 then it is a conic section of the sort we have already mentioned, or possibly a "degenerate conic", like $xy = 0$, $x^2 = 1$, $x^2 = 0$ which represent a pair of intersecting lines, a pair of parallel lines, and a "double line", respectively. Because we're working over the real numbers, you can also get some very odd "curves" like $x^2 + y^2 = 0$ (a single point) or $x^2 + y^2 = -1$ (the empty set). These become a lot less pathological when you think of the set of complex solutions: in the first case you get $(x + iy)(x - iy) = 0$, which defines a pair of complex lines ($y = \pm ix$) meeting at the origin, and in the second you get $x + iy = -1/(x - iy)$ which is a complex hyperbola.

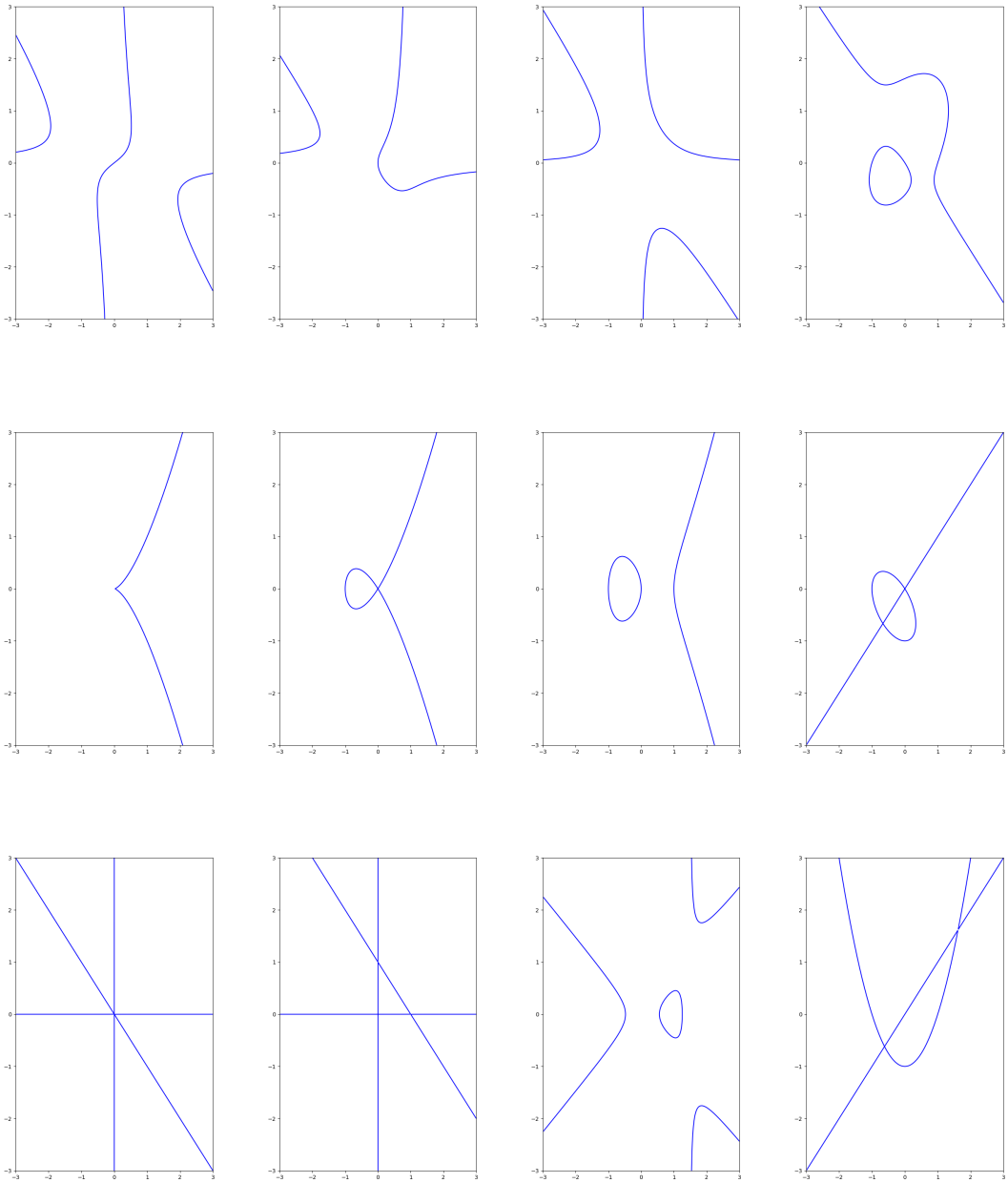
We will completely classify all the degree 2 curves (up to change of coordinates) in this week's worksheet.

2.4 Degree 3

Cubic curves are much harder to study. Newton was the first to attempt to classify all the cubic curves; he found 72 types (though there were gaps in his classification: there should be 78). The pictures below show a small selection of cubic curves, and some of

[5]Technically, these equations are "linear plus a constant", i.e. *affine linear*. I am likely to make this abuse of language repeatedly, so please get used to it now.

the interesting shapes you get.



Here are some observations which look reasonable from the pictures, and which we will come back to later:

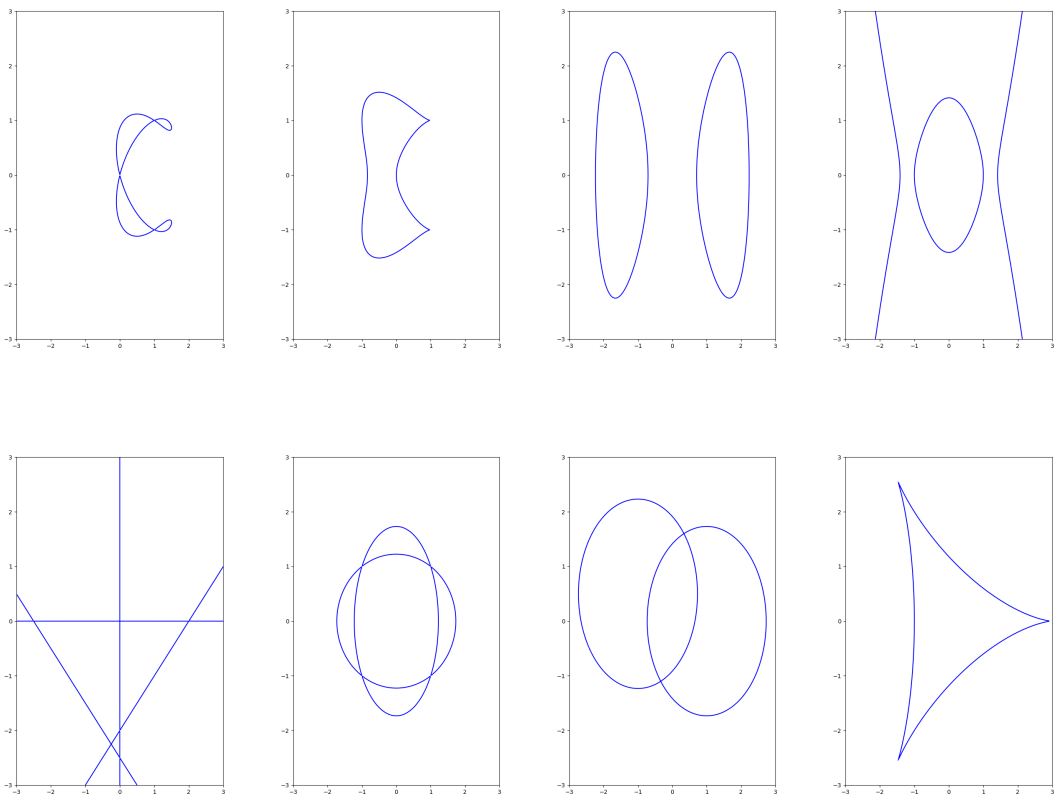
- a cubic curve has at most three asymptotes,
- a cubic curve has at most one "oval",
- a cubic curve can have at most three singularities.

Roughly speaking, a singularity is a point where the curve doesn't look like a straight line no matter how much you zoom in. For example, three of the curves on the bottom row have singularities: the left-most has a "cusp", the next right has a "node", and the right-most has two nodes.

The arithmetic of cubic curves (i.e. the question of when they have $\mathbb{Q}$-points and how many) is an extremely rich and highly-developed part of number theory. They are also known as *elliptic curves*. Elliptic curves have played an important role in pure mathematics (e.g. the Frey curve which was key in proving Fermat's Last Theorem), and they also appear in cryptography through Lenstra's prime factorisation algorithm and many cryptographic protocols whose names begin with EC (like ECDH or ECDSA). Don't get them confused with ellipses (which are much simpler degree 2 curves). The reason for the name is the appearance of *elliptic functions* which give natural parametrisations of elliptic curves. We will discuss this next week.

2.5 Degree 4

Quartic curves are harder again, and we just provide some pictures.



Here are some general facts that we will be able to prove by the end of the course:

- quartic curves have at most 4 asymptotes,
- quartic curves have at most 4 ovals,
- quartic curves have at most 6 singularities.

2.6 Outlook

You can ask "How many types of plane curves of a given degree are there?". Depending on how coarse a notion of "type" you use you will get different answers. Depending on how high the degree is, you may not even get an answer. For example, at the International Congress of Mathematicians in 1900, Hilbert posed 24 challenging problems to inspire mathematicians for the next hundred years. His sixteenth problem was:

Question 2.1. What "configurations of ovals" are possible for plane curves of degree d?

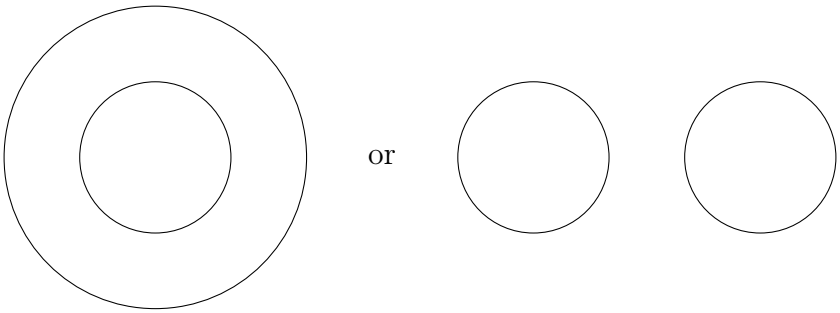
This is still unsolved for $d \geq 8$. To understand what Hilbert was asking, let's think about an example. Imagine a pair of circles (radii r_1, r_2 and centred at $(x_1, y_1), (x_2, y_2)$ respectively). The equation for the ith circle is

$$(x - x_i)^2 + (y - y_i)^2 - r_i^2 = 0, \quad i = 1, 2.$$

The union of the circles is defined by the product of these equations[6]:

$$P = ((x - x_1)^2 + (y - y_1)^2 - r_1^2)((x - x_2)^2 + (y - y_2)^2 - r_2^2) = 0.$$

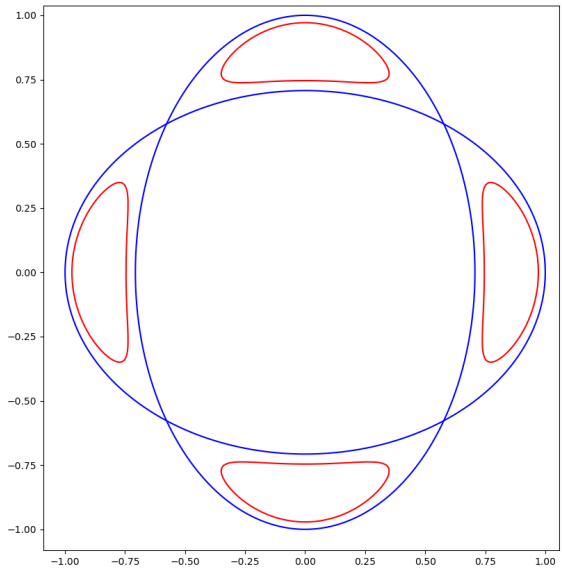
This is an equation of degree 4. So you can have quartic curves which look like a pair of circles. These could be nested, or just sitting next to one another:



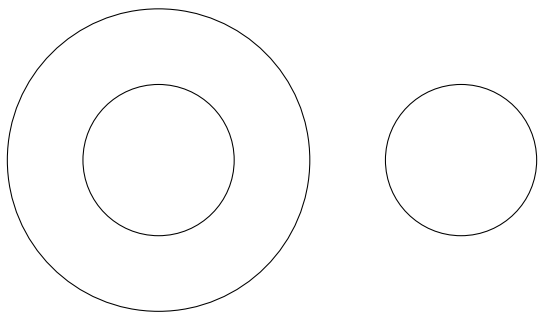
Another example is shown in red below: it comprises two intersecting ellipses, $x^2+2y^2 = 1$ and $2x^2+y^2 = 1$. If we perturb the equation a little bit we can "smooth" the intersections, to obtain the red quartic as shown. This particular quartic has equation

$$(x^2 + 2y^2 - 1)(2x^2 + y^2 - 1) = 0.05.$$

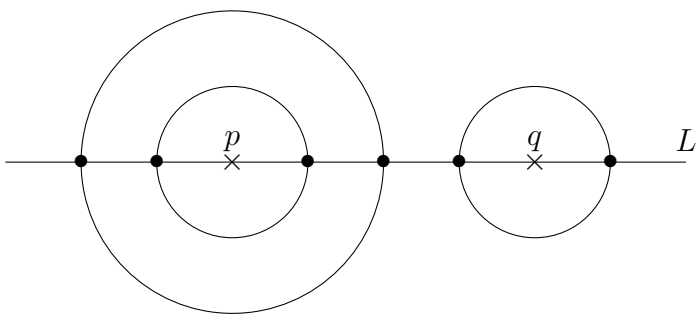
[6]because if the product vanishes then one of the two factors vanishes.



So we have seen three of the different "configurations of ovals" possible for quartic curves. Here is a configuration which cannot occur[7]:



To see that it cannot occur, suppose it was defined by a quartic equation $f(x, y) = 0$. Pick a line L passing through points p and q as shown:



The line L intersects the curve in at least six points, as shown. But we will see (Bézout's theorem, later) that a line intersects a quartic in at most four points. For example $\{f(x, y) = 0\}$ intersects the line $L = \{y = 0\}$ at the points $(x, 0)$ where x is a root of the degree 4 polynomial $f(x, 0)$: this polynomial has at most four roots.

It turns out there are six possible configurations of ovals for a quartic (what are the other

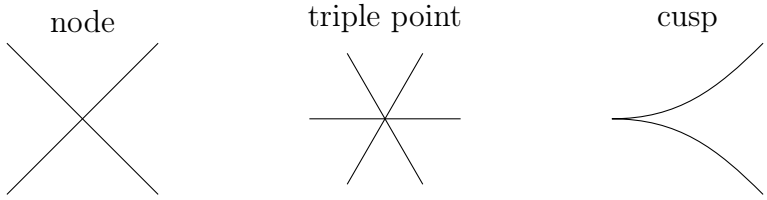
[7]The ovals don't need to be honest circles like in the picture.

three?). It is a theorem of Harnack (which we will prove later) that a plane curve of degree d can have at most $1+(d-1)(d-2)/2$ ovals, so for each d there is a finite number of ways these could be arranged/nested. Hilbert was asking for a complete list of these arrangements for each d.

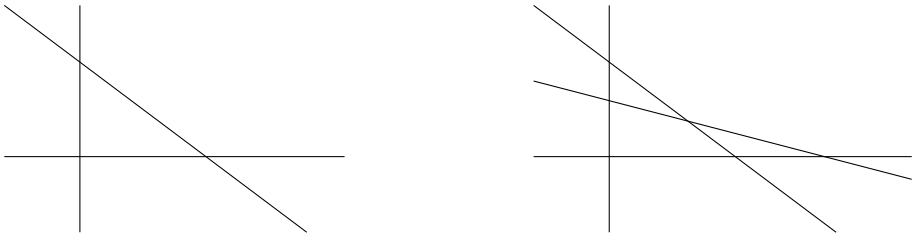
Here are some more questions with a similar flavour:

Question 2.2. How many singularities can a plane curve of degree d have? What do they look like, and how can they be arranged?

Here are some examples of curve singularities:



We will encounter more examples later; for now, let's just focus on *nodes*. A node is a singularity where two strands of the curve cross transversely (i.e. they are not tangent where they meet). We will see that a curve of degree d can have at most $d(d-1)/2$ nodes: in fact, this happens for a generic configuration of d lines, as we illustrate for $d = 3, 4$ below.



These examples are *reducible*: they are given by polynomials which can be factored (into the equations for d lines). If we focus on the more interesting case of *irreducible* varieties (whose defining polynomials do not factor) then we can get at most $(d-1)(d-2)/2$ nodes. So, for example, an irreducible quartic can have at most three nodes. But it turns out they cannot all lie on a single straight line!

The analogous questions for higher-dimensional varieties are still the subject of intense research by algebraic geometers today. Although I'm not an algebraic geometer, these are some of the questions I think about when I'm not teaching you guys.

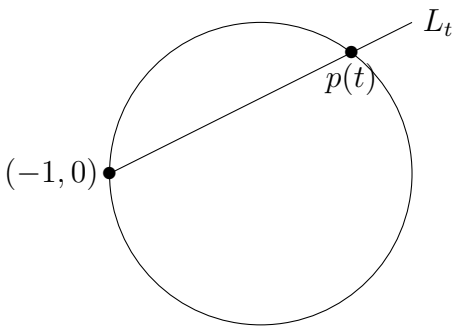
3 Parametrisations, I: Rational parametrisations

One way to get a better understanding of a curve is to find a *parametrisation*. For example, every point on the unit circle can be written as $x = \cos(\theta)$, $y = \sin(\theta)$ for some angle θ, so $\theta \mapsto (\cos\theta, \sin\theta)$ is a parametrisation of the circle.

Trigonometric functions are quite complicated; for example can you evaluate $\cos(1)$ in any meaningful way without the aid of a calculator? Here is an example of a question which is hard to answer from this point of view: how many points on the unit circle have both coordinates equal to rational numbers[8]? We know a few, e.g. $(1, 0)$, $(0, 1)$, $(-1, 0)$ and $(0, -1)$. We also know that if a, b, c form a Pythagorean triple (i.e. $a^2 + b^2 = c^2$) then $(a/c, b/c)$ lies on the unit circle and has rational coordinates; for example $(3/5, 4/5)$ works. In fact, the converse is also true: if $(x = \alpha/\beta, y = \gamma/\delta)$ is a rational point on the unit circle with $\alpha, \beta, \gamma, \delta$ all integers then $\alpha^2\delta^2 + \beta^2\gamma^2 = \beta^2\delta^2$, so $(\alpha\delta, \beta\gamma, \beta\delta)$ is a Pythagorean triple.

So this question is really: *how many Pythagorean triples are there?*

We will look for a simpler parametrisation of the circle, which will give us a formula for generating all the Pythagorean triples. Let L_t be the straight line with slope t passing through the point $(-1, 0)$. This intersects the circle at two points: at $(-1, 0)$ and at one more point, say $p(t) = (x(t), y(t))$.



We can calculate the coordinates of $p(t)$ as follows. First note that

$$(x(t), y(t)) = (-1, 0) + s(1, t) = (s - 1, st)$$

for some $s \in \mathbb{R}$ (because $p(t)$ lies on the line L_t) and $x(t)^2 + y(t)^2 = 1$ (because $p(t)$ also lies on the unit circle). Therefore

$$(-1 + s)^2 + (st)^2 = 1$$

or

$$(1 + t^2)s^2 - 2s = 0.$$

This is a quadratic equation in s with roots $s = 0$ (corresponding to the point $(-1, 0)$) and $s = 2/(1 + t^2)$ (corresponding to $p(t)$). Substituting back into $x(t) = s - 1$ and $y(t) = st$, we get

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

[8] Recall from Definition 1.10 that we call these $\mathbb{Q}$-points.

This is a different parametrisation of the circle by *rational functions*[9].

Remark 3.1. These formulae may look familiar from A-level maths. Using some trigonometric magic you find that if $x(t) = \cos(\theta)$ and $y(t) = \sin(\theta)$ then $t = \tan(\theta/2)$.

Corollary 3.2. *For any rational number t,*

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

is a point on the circle both of whose coordinates are rational. As a consequence, there are infinitely many points on the circle whose coordinates are both rational.

If $t = m/n$ then this gives us the Pythagorean triple $m^2 - n^2, 2mn, m^2 + n^2$. For example $t = 1/2$ corresponds to $3, 4, 5$.

Definition 3.3. Let $p \in k[x, y]$. We say that the algebraic curve $\mathbb{V}_k(p)$ is *rational* over k if there are polynomials $\alpha(t), \beta(t), \gamma(t), \delta(t) \in k[t]$ such that $\mathbb{V}_k(p)$ is the image of $t \mapsto (\alpha(t)/\beta(t), \gamma(t)/\delta(t))$ (where the domain of the parametrisation is the complement of the set of zeros of β and δ).

Example 3.4. We have just seen that the unit circle is rational over $\mathbb{Q}$.

Exercise 3.5. By looking at the line $x = ty$ with slope $1/t$ through the origin, find a rational (in fact, polynomial) parametrisation of $xy^2 - 6x^3 = y^4$. Parametric curves are much easier to plot using a computer; this is how I obtained the nice picture of this curve earlier.

Lemma 3.6. *Let k be an infinite field. If $\mathbb{V}_k(p)$ is rational over k then it is nonempty.*

Proof. There are at most finitely many points where the rational parametrisation is ill-defined (where $x(t)$ or $y(t)$ have poles) so there exists a point $(x(t), y(t)) \in \mathbb{V}_k(p)$. □

Example 3.7. Consider the curve $x^2 + y^2 = -1$. This is rational over $\mathbb{C}$: for example, we can write it as $x - iy = -1/(x + iy)$, so if we take $t = x + iy$ we get $x - iy = -1/t$ and so $x = (t - 1/t)/2$ and $y = (t + 1/t)/(2i)$. Therefore

$$(x(t), y(t)) = \left(\frac{t^2-1}{2t}, \frac{t^2+1}{2it}\right)$$

gives a rational parametrisation with $t \in \mathbb{C} \setminus \{0\}$. However, it is not rational over $\mathbb{R}$, because it doesn't have any real points.

Example 3.8. Consider the curve $x^2 + y^2 = 3$. This is rational over $\mathbb{R}$ (exercise using slopes!) but it is not rational over $\mathbb{Q}$. Again, this follows from the fact that $\mathbb{V}_{\mathbb{Q}}(x^2+y^2-3)$ is empty. To see this, suppose there were a $\mathbb{Q}$-point $x = \alpha/\beta$, $y = \gamma/\delta$. Then

$$\alpha^2\delta^2 + \beta^2\gamma^2 = 3\beta^2\delta^2.$$

Setting $A = \alpha\delta$, $B = \beta\gamma$, $C = \beta\delta$ (all integers), we get $A^2 + B^2 = 3C^2$. Suppose that A, B, C is an integer solution of this equation with $|A|$ minimal. Note that $|A| \neq 0$ because $\sqrt{3}$ is irrational. Reduce modulo 4. If C is odd then $3C^2 = 3 \mod 4$. If C is

[9] A function is called *rational* if it is a ratio of two polynomials. Warning: the word "rational" will have several meanings in this course. In fact, it is one the most overused words in all of mathematics.

even then $3C^2 = 0 \mod 4$. Moreover, $A^2 + B^2$ can only take on the values $0, 1, 2 \mod 4$, because the only squares modulo 4 are

$$0^2 = 2^2 = 0 \mod 4, \quad 1^2 = 3^2 = 1 \mod 4,$$

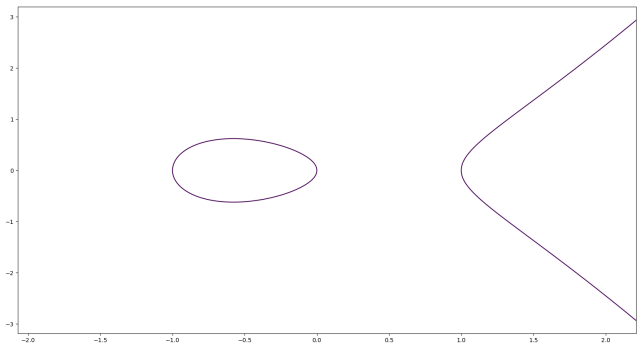
so the only sums of two squares are

$$0 + 0 = 0, \quad 1 + 0 = 1, \quad 1 + 1 = 2 \mod 4.$$

The only possibility consistent with both of these constraints is that A, B and C are all even, but then $A/2, B/2, C/2$ gives another solution with $|A|$ strictly smaller, which contradicts our minimality assumption.

Remark 3.9. The above argument proves $x^2 + y^2 = p$ is not rational over $\mathbb{Q}$ for any prime $p = 3 \mod 4$. By contrast, if $p = 1 \mod 4$, there is a $\mathbb{Q}$-point on this curve (in fact you can find a point with $x, y \in \mathbb{Z}$): an odd prime p can be written as a sum of squares if and only if $p = 1 \mod 4$.

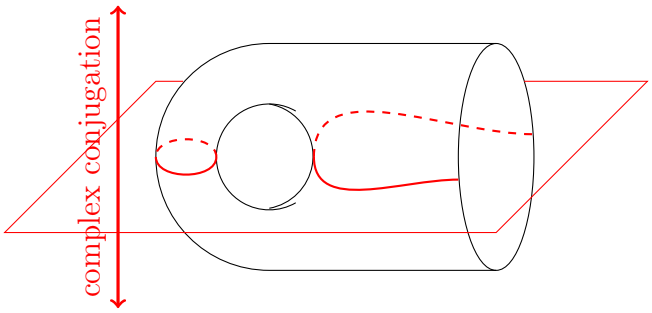
Example 3.10. Consider the curve $y^2 = x^3 - x$. We can sketch the real points of this curve. It can be written as $y = \sqrt{x^3 - x}$. The polynomial $x^3 - x$ has roots at $-1, 0, 1$; it is negative if $x < -1$ or $x \in (0, 1)$ and positive otherwise. Over $x \in [-1, 0] \cup [1, \infty)$ we get two possible y values (the square roots, symmetric about the x-axis).



This is called an *elliptic curve*, and is one of the simplest examples of a curve which admits *no* rational parametrisation. Instead, we can parametrise it using *elliptic functions*. As a result, the set of rational points of a cubic (elliptic) curve like this is usually very tricky to understand. In this particular case, there are only three: $(-1, 0)$, $(0, 0)$ and $(1, 0)$. We will see later that if you add in a "point at infinity", the set of rational points on an elliptic curve is an abelian group in a natural way, and the Mordell-Weil theorem tells us it is finitely generated[10]. One of the biggest open problems in mathematics (the Birch–Swinnerton-Dyer conjecture) asks how to calculate the rank of that group for an arbitrary elliptic curve over $\mathbb{Q}$.

Remark 3.11. The set of complex points of this curve is a 2-dimensional surface called a punctured torus; the picture below shows how you can imagine the real points (red) sitting inside it. The real locus cuts the complex curve in two pieces, which are interchanged by complex conjugation.

[10]This particular curve gives us the group $\mathbb{Z}/2 \times \mathbb{Z}/2$.


complex conjugation

4 Parametrisations, II: Elliptic functions

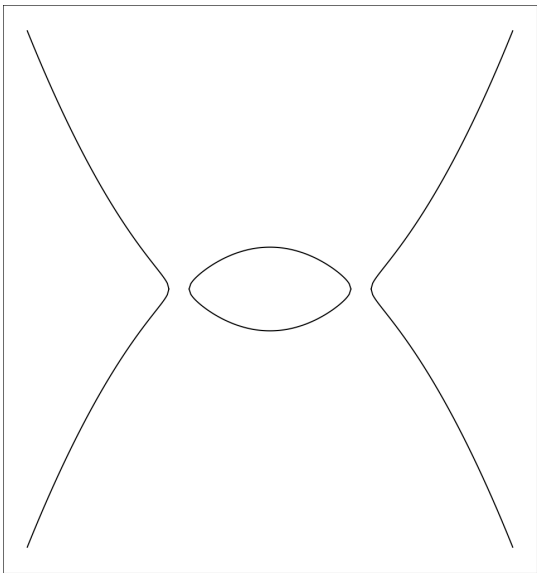
4.1 Goal

Our goal in this section is to write down a parametrisation of a specific non-rational curve: the *Jacobi quartic*

$$y^2 = (1 - x^2)(1 - k^2x^2).$$

Here, k is a constant; for now we will take $k \in [0, 1)$–it will have the geometric interpretation of the *eccentricity* of an ellipse–but later it will just be any complex number.

The curve looks like this (for $k = 0.8$):



which you can see by plotting $y = \pm\sqrt{(1 - x^2)(1 - k^2x^2)}$ in the three ranges where it takes on real values. The x-intercepts are at $-1/k, -1, 1, 1/k$, and the curve is symmetric under reflection in the x- and y-axes.

As mentioned, the Jacobi quartic is not a rational curve unless $k = 0$ (in which case it's a circle) so we will need to parametrise it using something more complicated than rational functions. When $k = 0$ we can parametrise it using $x(t) = \cos(t)$ and $y(t) = \sin(t)$. We will now introduce the *Jacobi elliptic functions*, which are deformations of the usual trigonometric functions, and which allow us to parametrise the Jacobi quartic.

4.2 Jacobi elliptic functions

Fix a number $b \geq 1$. Consider the ellipse $E = \{p^2 + q^2/b^2 = 1\} \subset \mathbb{R}^2$. The *eccentricity* of E is defined to be $k := \sqrt{1 - 1/b^2}$.

Given a point $(p, q) \in E$ let r be the radius and θ the angle such that

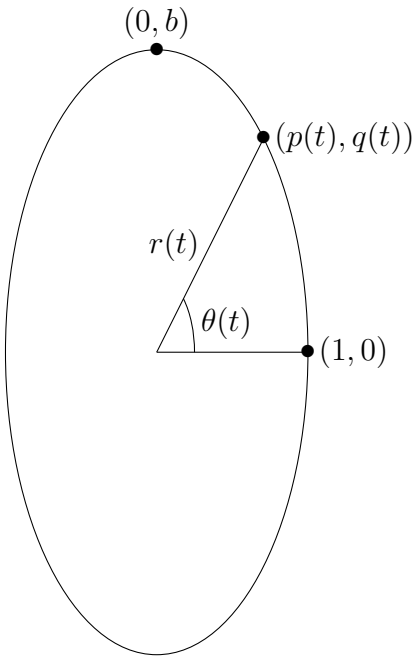
$$p = r\cos\theta, \quad q = r\sin\theta.$$

We will parametrise the ellipse in a slightly odd way. We could use the angle θ. Or we could use the arc-length (which is equivalent to θ if E is the unit circle). Instead, we will

use the parameter

$$t = \int_0^\theta r \, d\theta.$$

Since $r > 0$, this quantity increases as we go around the ellipse; if E were a circle with $r \equiv 1$, t would be precisely θ. The reason for this choice of parametrisation is that it will make the algebra simpler later.



Definition 4.1. We define:

- Jacobi's *amplitudinis* function $\mathrm{am}(t; k^2) := \theta(t)$,
- Jacobi's *sinus amplitudinis* function $\mathrm{sn}(t; k^2) := p(t)/r(t) = \sin \mathrm{am}(t; k^2)$,
- Jacobi's *cosinus amplitudinis* function $\mathrm{cn}(t; k^2) := q(t)/r(t) = \cos \mathrm{am}(t; k^2)$,
- Jacobi's *delta amplitudinis* function, $\mathrm{dn}(t; k^2) := 1/r(t)$.

In the case $k = 0$ (when our ellipse becomes the unit circle) these reduce to $\mathrm{am}(t; 0) = t$, $\mathrm{dn}(t; 0) = 1$, $\mathrm{sn}(t; 0) = \sin(t)$, $\mathrm{cn}(t; 0) = \cos(t)$. Most of the time, we suppose that we're working with a fixed ellipse and omit k from the notation.

Lemma 4.2. *We have*

$$\begin{aligned} \mathrm{cn}^2(t) + \mathrm{sn}^2(t) &= 1, & \mathrm{cn}^2(t) + \mathrm{sn}^2(t)/b^2 &= \mathrm{dn}^2(t), \\ \mathrm{dn}^2(t) + k^2 \mathrm{sn}^2(t) &= 1, & \mathrm{dn}^2(t) - k^2 \mathrm{cn}^2(t) &= 1 - k^2, \end{aligned}$$

Proof. Since $\mathrm{cn}(t) = \cos \mathrm{am}(t)$ and $\mathrm{sn}\,\mathrm{am}(t)$ the formula $\mathrm{cn}^2 + \mathrm{sn}^2 = 1$ is immediate.

Since $p(t)^2 + q(t)^2/b^2 = 1$ we have

$$r^2(t)\,\mathrm{cn}^2(t) + r^2(t)\,\mathrm{sn}^2(t)/b^2 = 1,$$

which implies $\mathrm{cn}^2(t) + \mathrm{sn}^2(t)/b^2 = \mathrm{dn}^2(t)$.

Replacing $\mathrm{cn}^2 = 1 - \mathrm{sn}^2$ and dividing by r^2, we get

$$\mathrm{sn}^2(t)(1 - 1/b^2) + \mathrm{dn}(t)^2 = 1.$$

Since $k^2 = 1 - 1/b^2$ this gives $\mathrm{dn}^2 + k^2 \,\mathrm{sn}^2 = 1$.

Similarly we can replace $\mathrm{sn}^2 = 1 - \mathrm{cn}^2$ and get $\mathrm{dn}^2 - k^2 \,\mathrm{cn}^2 = 1/b^2 = 1 - k^2$. □

Lemma 4.3. *We have*

$$\frac{d}{dt}\,\mathrm{sn}(t) = \mathrm{cn}(t)\,\mathrm{dn}(t), \quad \frac{d}{dt}\,\mathrm{cn}(t) = -\,\mathrm{sn}(t)\,\mathrm{dn}(t), \quad \frac{d}{dt}\,\mathrm{dn}(t) = -k^2\,\mathrm{cn}(t)\,\mathrm{sn}(t).$$

Proof. In what follows, we write $df/dt = f'$ for derivatives. By definition, we have $r = 1/\,\mathrm{dn}$ and $t = \int_0^{\theta(t)} r\,d\theta := \int_0^t r(\tau)\theta'(\tau)\,d\tau$. Differentiating the formula for t with respect to t using the fundamental theorem of calculus gives $1 = \theta'(t)/\,\mathrm{dn}(t)$, or $\mathrm{dn}(t) = \theta'(t)$. Now the chain rule tells us that

$$\begin{aligned}
\mathrm{sn}'(t) &= \frac{d}{dt}\sin\theta(t) = \theta'(t)\cos\theta(t) = \mathrm{cn}(t)\,\mathrm{dn}(t) \\
\mathrm{cn}'(t) &= \frac{d}{dt}\cos\theta(t) = -\theta'(t)\sin\theta(t) = -\,\mathrm{sn}(t)\,\mathrm{dn}(t)
\end{aligned}$$

Finally, differentiating $\mathrm{dn}^2 + k^2\,\mathrm{sn}^2 = 1$ gives

$$\mathrm{dn}\,\mathrm{dn}' + k^2\,\mathrm{sn}\,\mathrm{sn}' = 0$$

which means

$$\mathrm{dn}' = -k^2\,\mathrm{sn}\,\mathrm{sn}'\,/\,\mathrm{dn} = -k^2\,\mathrm{cn}\,\mathrm{sn}\,.$$ □

4.3 Differential equation and parametrisation

Corollary 4.4. *We have*

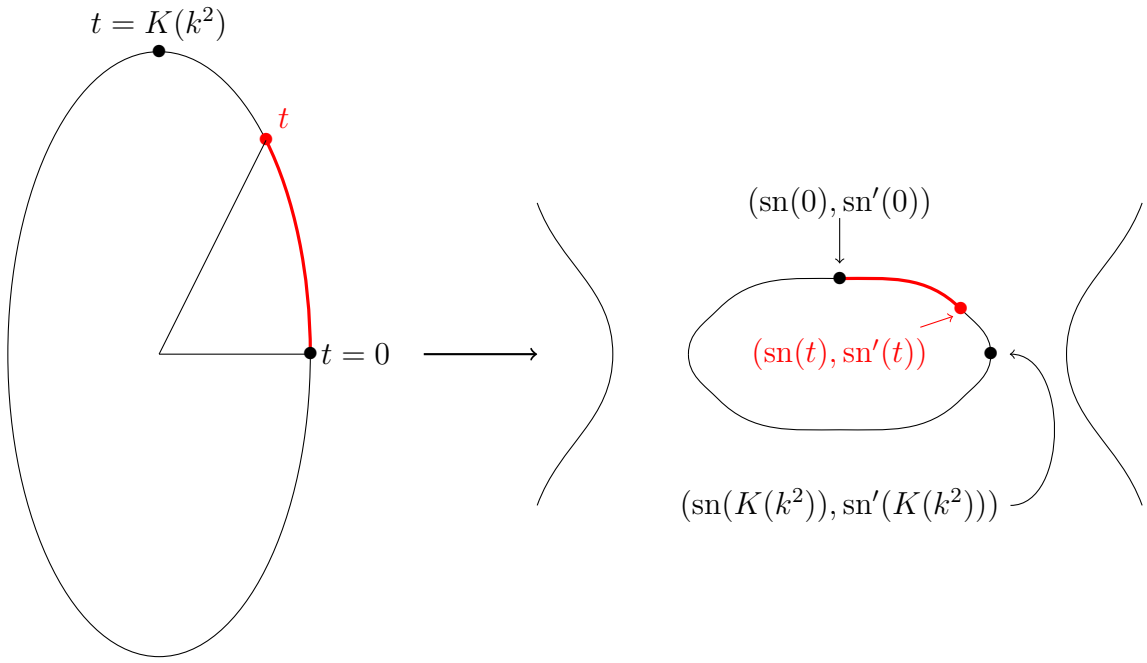
$$\left(\frac{d\,\mathrm{sn}(t)}{dt}\right)^2 = (1 - \mathrm{sn}^2(t))(1 - k^2\,\mathrm{sn}^2(t)).$$

In other words, $z(t) = \mathrm{sn}(t; k^2)$ *solves the differential equation* $z' = \sqrt{(1 - z^2)(1 - k^2 z^2)}$.

Proof. We know that $\mathrm{sn}' = \mathrm{dn}\,\mathrm{cn}$, $\mathrm{cn} = \sqrt{1 - \mathrm{sn}^2}$ and $\mathrm{dn} = \sqrt{1 - k^2\,\mathrm{sn}^2}$, so squaring and substituting gives the result. □

Corollary 4.5. *The real parametric curve* $x(t) = \mathrm{sn}(t; k^2)$, $y(t) = \mathrm{sn}'(t; k^2)$ *is the Jacobi quartic* $y^2 = (1 - x^2)(1 - k^2 x^2)$.

Let's think about this for a moment. When $t = 0$ we have $\mathrm{sn}(0) = 0$ and $\mathrm{sn}'(0) = \mathrm{cn}(0)\,\mathrm{dn}(0) = 1$, so this parameter value corresponds to the point $(0, 1)$. As t increases, $\mathrm{sn}(t; k^2)$ increases, whilst $\mathrm{cn}(t; k^2)$ and $\mathrm{dn}(t; k^2)$ decreases (hence $\mathrm{sn}'(t; k^2)$ decreases), so the parametric curve moves to the right and down. We see that this is tracing out the loop in the middle of the quartic. Let $K(k^2) = \int_0^{\pi/2} r\,d\theta$ denote the value of t a quarter of the way around the ellipse. We see that $\mathrm{sn}(K(k^2); k^2) = 1$, $\mathrm{cn}(K(k^2); k^2) = 0$ and $\mathrm{dn}(K(k^2); k^2) = b$, so this value of the parameter corresponds to the point $(\mathrm{sn}(K(k^2); k^2), \mathrm{sn}'(K(k^2); k^2)) = (1, 0)$ on the quartic. Once we have reached $t = 4K(k^2)$, we should return to $(0, 1)$, having traced out the whole loop.



Lemma 4.6. *We have*

$$K(k^2) = \int_0^1 \frac{dz}{\sqrt{(1-z^2)(1-k^2z^2)}}.$$

More generally, the function

$$\mathrm{arcsn}(x; k^2) = \int_0^x \frac{dz}{\sqrt{(1-z^2)(1-k^2z^2)}}$$

defines a function inverse to sn.

Proof. We prove the more general formula; the formula for $K(k^2)$ follows because $K(k^2) = \mathrm{arcsn}(1)$ by definition. Let's write t as a function of x, that is $t(x)$ is the parameter value such that $\mathrm{sn}(t(x); k^2) = x$. By the fundamental theorem of calculus, we know that

$$t(x) = \int_0^x \frac{dt}{dx} dx = \int_0^x \frac{dx}{y} = \int_0^x \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}},$$

where we are sloppily using the same name x for the limit and for the free variable x in the integral. If you are careful to distinguish, e.g. writing z for the free variable inside the integral, then you get the stated formula. □

Remark 4.7. This is an example of an *elliptic integral* (more precisely, an incomplete elliptic integral of the first kind–the specific value $\mathrm{arcsn}(1) = K(k^2)$ is a *complete* elliptic integral). Compare this with the more familiar formula (the special case $k = 0$)

$$\arcsin(x) = \int_0^x \frac{dz}{\sqrt{1-z^2}}.$$

Remark 4.8. Integrals of the form $\int dx/y$ taken along a loop in an algebraic curve are called *period integrals* because the same elliptic integral arises when computing the period of a simple pendulum. Note that the proof of the lemma only depended on having a parametrisation $(x(t), y(t))$ with $y(t) = dx(t)/dt$, which is what our differential equation guarantees.

4.4 The rest of the curve

So far, our parametrisation only covers the central loop in the Jacobi quartic; what about the two unbounded branches? It turns out that we can use the same parametrisation, but we need to allow t to take on complex values. The easiest way to proceed is to work with the elliptic integral.

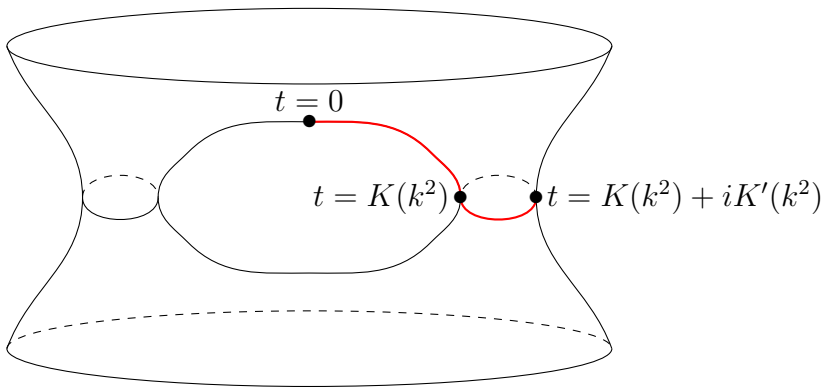
Think about it this way. To find the parameter t associated to the point $(x(t), y(t))$ we perform the integral $\int_0^x \frac{dx}{y}$ along the path $y = \sqrt{(1-x^2)(1-k^2x^2)}$, $x \in [0, x(t)]$. At $(1, 0)$, we got $t = K(k^2)$ by following the loop inside the quartic for a quarter of its length. Now let's pick a path in the quartic from $(1, 0)$ to $(1/k, 0)$. There is no real path, because the points $(1, 0)$ and $(1/k, 0)$ belong to different connected components of the real quartic. But there is a path of complex points! Namely $(x, \sqrt{(1-x^2)(1-k^2x^2)})$, $x \in [1, 1/k]$ (the y values are now imaginary). If we integrate dx/y along this path, we should get the difference between the parameter values at $(1, 0)$ and $(1/k, 0)$, namely we get $iK'(k^2)$ where:

$$K'(k^2) = \int_1^{1/k^2} \frac{dx}{\sqrt{(x^2-1)(1-k^2x^2)}}.$$

The picture below shows a cartoon of the set

$$C = \{\mathbb{V}_{\mathbb{C}}(y^2 - (1-x^2)(1-k^2x^2))$$

of complex points (which really live in the 4-dimensional space $\mathbb{C}^2$, so cannot be represented accurately in 3-d) to give you a sense of what is going on; the red path is the path we have chosen to integrate dx/y along to find t.



Remarkably, we have:

Lemma 4.9.

$$K'(k^2) = K(1-k^2).$$

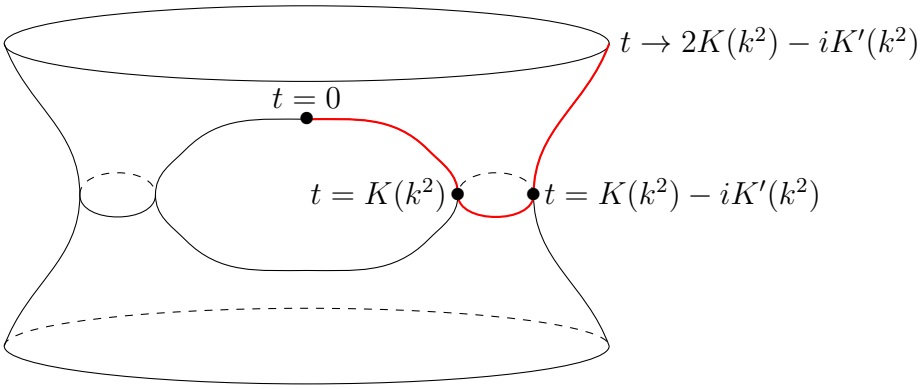
Proof. This is an exercise in substitution: use the substitution $\xi = \frac{1}{\sqrt{1-(1-k^2)x^2}}$ to convert the integral $K(1-k^2)$ into $K'(k^2)$. □

Corollary 4.10. *The parameter value $t = K(k^2) + iK'(k^2)$ maps to the point $(1/k, 0)$ under $(\mathrm{sn}(t; k^2), \mathrm{sn}'(t; k^2))$.*

What happens as we move out along the unbounded branch towards $x = \infty$? We actually

arrive there after a finite parameter:

$$\int_1^{1/k} \frac{dx}{\sqrt{(x^2-1)(k^2x^2-1)}} = K(k^2).$$

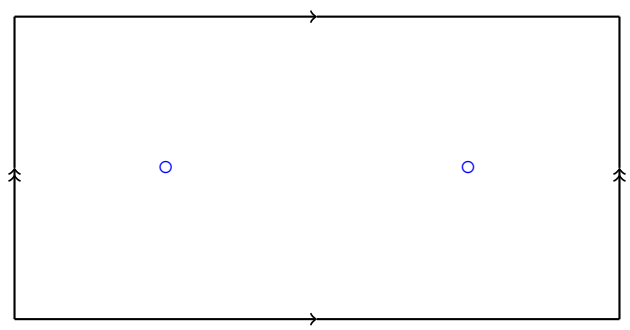


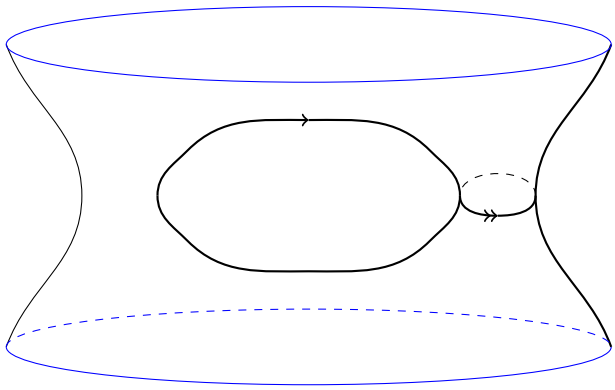
So we are trying to construct the value of the parameter t at $q \in C$ by integrating dx/y along a path from $p = (0,1) \in C$ to q. Why does the result not depend on the path of integration we chose? It is a consequence of Stokes's theorem that the value of the integral does not change if we wiggle the path a little. However, if we chose two paths from p to q which are not connected by a family of paths from p to q then the integral might give different answers. So t is really a *multivalued function* on the set of complex points of the Jacobi quartic (just as arcsin and arccos are multivalued functions on the circle).

An easier way to imagine this is that the function sn extends to a complex-valued function on $\mathbb{C}$, with infinitely many poles at the set of points $P = \{2\alpha K(k^2) + i(1+2\beta)K'(k^2) : \alpha\,\beta \in \mathbb{Z}^2\}$, and $(\mathrm{sn}, \mathrm{sn}')\colon \mathbb{C}\setminus P \to C$ gives a parametrisation which is infinitely redundant: each point is covered by infinitely many parameter values. The set of preimages of a given point $p \in C$ actually forms a lattice, because sn has the following periodicities:

$$\mathrm{sn}(u + 4K(k^2); k^2) = \mathrm{sn}(u; k^2), \qquad \mathrm{sn}(u + 2iK'(k^2); k^2) = \mathrm{sn}(u; k^2) \text{ for all } u \in \mathbb{C}.$$

If we restrict attention to the rectangle $[-K, 3K] \times [0, 2K']$ then opposite points on the boundary map to the same point in C, and the points $iK'(k^2)$ and $2K(k^2) + iK'(k^2)$ are the only poles in this region. A rectangle with its opposite sides identified is topologically a torus, so we deduce that C is topologically a torus with two punctures.





5 Algebra and geometry, I

5.1 Ideal of a subset

What are the equations of the origin in $\mathbb{R}^2$? You could say "$x = y = 0$". Or you could say "$x + y = x - y = 0$". Or $x = y = x + y = 0$. There are many possibilities.

Definition 5.1. Given a subset $X \subset k^n$, define *the ideal of* X

$$\mathbb{I}(X) = \{f \in k[x_1, \ldots, x_n] \,:\, f(p) = 0 \,\forall\, p \in X\}.$$

That is $\mathbb{I}(X)$ consists of all polynomials which vanish on X (they are also allowed to vanish elsewhere).

Recall that an ideal in a ring R is a subset I which is closed under addition and under multiplication by any element of the ring. That is, $a, b \in I$ implies $a + b \in I$ and $a \in I$, $r \in R$ implies $ra \in I$.

Lemma 5.2. *$\mathbb{I}(X)$ is an ideal.*

Proof. Exercise! □

Example 5.3. If $X = \emptyset$ is the empty set then $\mathbb{I}(X) = k[x_1, \ldots, x_n]$. This is because the condition that "f vanishes at every point of the empty set" is vacuously satisfied by every polynomial f.

Example 5.4. If k is infinite and $X = k^n$, we have $\mathbb{I}(X) = 0$. If $n = 1$, this is easy to see: a nonconstant polynomial can only have a finite number of roots, so if $f(x)$ vanishes for all the infinitely many points $x \in k$ then f must be constant and equal to zero. It's a bit harder to see for higher n (you can prove it by induction–I omitted the proof in lectures).

Remark 5.5. If k is finite then there are nontrivial polynomials which nonetheless vanish everywhere. For example, if $k = \mathbb{Z}/2$ then $x^2 - x$ vanishes at $x = 0$ and $x = 1$, which exhausts the possible points in k.

Example 5.6. Consider the subset $\{0\} \subset k$. A polynomial $f(x)$ vanishes at $x = 0$ if and only if its constant term is zero, so $\mathbb{I}(\{0\})$ consists of all polynomials with zero constant term.

This gives us a dictionary between geometry and algebra: a subset X gives us an ideal $\mathbb{I}(X)$. The bigger the subset, the smaller the ideal (for example, the empty set corresponds to the bigger ideal, while the zero ideal corresponds to the set of all points).

5.2 The algebraic set of an ideal

In the other direction, given an ideal I, you can define a set

$$\mathbb{V}(I) = \{x \in k^n \,:\, f(x) = 0 \,\forall\, f \in I\}.$$

We will expand our definition of algebraic sets to include such sets[11].

Example 5.7. We have $\mathbb{V}(0) = k^n$ because the zero polynomial vanishes everywhere. Algebraic geometers often write $\mathbb{A}^n(k)$ or just $\mathbb{A}^n$ for k^n: it is called the "affine n-space over k".

Example 5.8. We have $\mathbb{V}(k[x_1, \ldots, x_n]) = \emptyset$ because the ideal $k[x_1, \ldots, x_n]$ contains the constant polynomial 1, which vanishes nowhere.

Lemma 5.9. *If $X = \mathbb{V}(I)$ and $Y = \mathbb{V}(J)$ then*

$$X \cap Y = \mathbb{V}(I + J),$$

where $I + J$ is the ideal

$$I + J = \{a + b \,:\, a \in I,\, b \in J\}.$$

Proof. Suppose $a \in I$, $b \in J$ and $z \in X \cap Y$. Since $z \in X$ we have $a(z)$ and since $z \in Y$ we have $b(z) = 0$. Therefore $a(z) + b(z) = 0$ for all $a \in I$ and $b \in J$, so $z \in \mathbb{V}(I + J)$.

Suppose $z \in \mathbb{V}(I + J)$. Taking $b = 0 \in J$, we have $a(z) + 0 = 0$ for all $a \in I$, so $z \in X$. Taking $a = 0 \in I$, we have $0 + b(z) = 0$ for all $b \in J$, so $z \in Y$. Therefore $z \in X \cap Y$. □

Lemma 5.10. *If $X = \mathbb{V}(I)$ and $Y = \mathbb{V}(J)$ then*

$$X \cup Y = \mathbb{V}(IJ),$$

where IJ is the ideal

$$IJ = \{a_1 b_1 + \cdots + a_m b_m \,:\, a_1, \ldots, a_m \in I,\ b_1, \ldots, b_m \in J\}.$$

Proof. Pick $a_1, \ldots, a_m \in I$ and $b_1, \ldots, b_m \in J$. If $z \in X \cup Y$ then either:

- $z \in X$, in which case $a_i(z) = 0$ for all i, so $a_1(z)b_1(z) + \cdots + a_m(z)b_m(z) = 0$ and $z \in \mathbb{V}(IJ)$, or
- $z \in Y$, in which case $b_i(z) = 0$ for all i, so $a_1(z)b_1(z) + \cdots + a_m(z)b_m(z) = 0$ and $z \in \mathbb{V}(IJ)$.

This shows that $z \in X \cup Y$ implies $z \in \mathbb{V}(IJ)$.

Conversely, suppose $z \in \mathbb{V}(IJ)$ but $z \notin X$. Then there exists $a \in I$ with $a(z) \neq 0$ (otherwise z would be in $X = \mathbb{V}(I)$). Since $z \in \mathbb{V}(IJ)$, $a(z)b(z) = 0$ for all $b \in J$. Since $a(z) \neq 0$, this implies $b(z) = 0$ for all $b \in J$, therefore $z \in \mathbb{V}(J) = Y$. So $z \in \mathbb{V}(IJ)$ implies $z \in X \cup Y$ □

Example 5.11. Given polynomials $f_1, \ldots, f_m$, define $(f_1, \ldots, f_m)$ ("the ideal generated by $f_1, \ldots, f_m$") to be the ideal

$$\{\alpha_1 f_1 + \cdots + \alpha_m f_m \,:\, \alpha_1, \ldots, \alpha_m \in k[x_1, \ldots, x_n]\}.$$

In our notation from earlier, $\mathbb{V}(f_1, \ldots, f_m) = \mathbb{V}((f_1, \ldots, f_m))$.

Example 5.12. For any polynomial g, the ideal (g) consists (by definition) of polynomials

[11] In fact, it is a theorem of Hilbert that any ideal in a polynomial ring is "generated" by finitely many polynomials, so $\mathbb{V}(I)$ is actually $\mathbb{V}(f_1, \ldots, f_m)$ for some finite set $f_1, \ldots, f_m$ of polynomials.

divisible by g. So $f \in (g)$ implies $f = gh$ for some polynomial h.

In fact, any ideal in $k[x_1, \ldots, x_n]$ is generated by some finite set of polynomials. This important fact is called *Hilbert's basis theorem.* We will not prove it in this course.

5.3 Nullstellensatz

Start with an ideal I. Look at the set of points $\mathbb{V}(I)$ where everything in I vanishes. Look at the ideal $\mathbb{I}(\mathbb{V}(I))$ of all polynomials which vanish at all these points. By definition, any polynomial $f \in I$ vanishes on $\mathbb{V}(I)$, so $I \subset \mathbb{I}(\mathbb{V}(I))$. But sometimes we can discover *more* polynomials that vanish on $\mathbb{V}(I)$.

Example 5.13. If $I = (x^2)$ then $\mathbb{V}(I) = \{0\}$. But $\mathbb{I}(\{0\}) = (x)$. Since x is not divisible by x^2, $x \notin (x^2)$.

In fact, this is the only way we can get more polynomials: by taking roots of polynomials in I. More precisely:

Theorem 5.14 (Hilbert's Nullstellensatz)**.** *Suppose that k is an algebraically closed field. Then*

$$\mathbb{I}(\mathbb{V}(I)) = \{f \in k[x_1, \ldots, x_n] \,:\, f^r \in I \text{ for some } r\}.$$

This ideal has a name: it is called the *radical* of I, often written $\sqrt{I}$ or $\mathrm{rad}(I)$. The name comes from the Latin *radix* meaning root: we have to add in any roots (square or otherwise) of the polynomials in our defining ideal.

The Nullstellensatz looks like a weird technical statement, but it will turn out to be the magical ingredient which will allow us to prove many other facts that are intuitive but otherwise hard to get a grasp on. For example:

Theorem 5.15 (Weak Nullstellensatz)**.** *Let k be an algebraically closed field. The only ideal I which gives an empty variety $\mathbb{V}_k(I) = \emptyset$ is the full ideal $I = k[x_1, \ldots, x_n]$.*

Proof. The constant function 1 vanishes on $\emptyset$, so the Nullstellensatz implies $1^r \in I$. Since $1^r = 1$, this implies $1 \in I$, and hence $I = k[x_1, \ldots, x_n]$. □

Now how do we prove the Nullstellensatz? In Chapter 8, we will give a different (direct) proof of the weak Nullstellensatz. In Sheet 2 there is a question which shows that the weak Nullstellensatz implies the Nullstellensatz.

Remark 5.16. The fact we're working over an algebraically closed field is very important: otherwise the Nullstellensatz fails miserably. For example, the variety $\mathbb{V}_{\mathbb{R}}(x^2+1)$ is empty, even though the ideal $(x^2 + 1) \subset \mathbb{R}[x]$ is proper.

5.4 Irreducibility

In this section we will give another application of the Nullstellensatz. But first some background. Recall that the curve $\{gh = 0\}$ consists of two pieces: $\{g = 0\}$ and $\{h = 0\}$. We call these *components* of the curve.

Definition 5.17. A polynomial f is called *reducible* if it admits a factorisation into two nonconstant polynomials $f = gh$. Otherwise it is called *irreducible*.

Definition 5.18. A curve C is called *reducible* if it can be written as a union of two curves $C' \cup C''$ with $C \neq C'$ and $C \neq C''$. Otherwise it is called *irreducible*.

There is a closely-related notion of *prime element*:

Definition 5.19. A polynomial f is called prime if, whenever f divides gh, f divides either g or h.

Remark 5.20 (Remarks on irreducible components). In a general ring, prime implies irreducible, but not vice versa. However, in a polynomial ring, irreducible implies prime (MATH322). Just like integers, polynomials can be factored in a unique way (up to overall scale factors and permutation of factors) into irreducibles. Another way to say this is that the polynomial ring $k[x_1, \ldots, x_n]$ is a *unique factorisation domain* (this, too, will be proved in MATH322). Geometrically, this means that any affine algebraic set can be decomposed in a unique way into a finite number of irreducible subsets, called the *irreducible components*. It also means that we can make sense of the gcd of two polynomials. We won't focus much on factorisation and irreducibility, as this would overlap too much with MATH322.

Theorem 5.21. *If f is irreducible and k is algebraically closed then $\{f = 0\}$ is irreducible.*

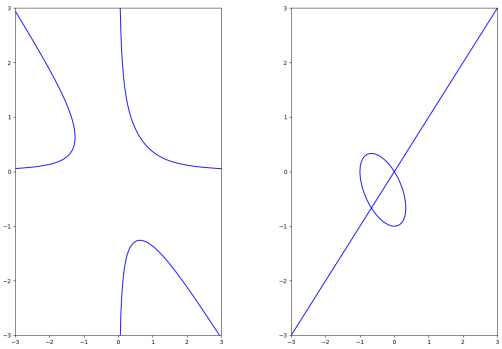
Proof. We'll prove the contrapositive statement: if $\{f = 0\}$ is a reducible curve C then f is reducible.

Since C is reducible, we can decompose C as $C = C' \cup C''$ with $C' = \{g = 0\} \neq C$ and $C'' = \{h = 0\} \neq C$. We observe that:

- f does not divide g because there is a point $p \in C \setminus C'$ where $f(p) = 0$ but $g(p) \neq 0$.
- Similarly, f does not divide h.
- gh vanishes on C.

Since gh vanishes on C, we have $gh \in \mathbb{I}(C)$, so $gh \in \mathrm{rad}(f)$ by the Nullstellensatz. Therefore, for some r, $(gh)^r$ is divisible by f. If f is irreducible then it is prime so must divide either g or h. But f divides neither g nor h, so f cannot be irreducible. $\square$

Remark 5.22. What do irreducible components actually look like? For example, take the curves $C = \{x^2y + xy^2 = 1/2\}$ and $C' = \{y^2 - x^3 - x^2 + y^3 = 0\}$:



The curve C is made up of three unbounded arcs. Are these the irreducible components? In fact, this polynomial is irreducible, so these arcs are not irreducible components. Indeed, if we looked at the set of complex points of this curve, it would be connected. Irreducibility means that if a polynomial vanishes on one of these arcs, it necessarily vanishes on the other two. By contrast, the curve C' is reducible. The defining polynomial factors as $(y-x)(x^2+y^2+x+y)$, so the irreducible factors are the line $y=x$ and the ellipse $x^2+y^2+x+y=0$.

5.5 Appendix: Recap from MATH225

Definition 5.23. A ring is a set R together with operations $+, \cdot$ satisfying the usual distributivity/associativity rules and containing a zero element 0 such that $0+x = x+0 = x$ and $0 \cdot x = x \cdot 0 = 0$ for all $x \in R$. We will focus exclusively on commutative rings (for which $a \cdot b = b \cdot a$ for all $a, b \in R$) with an identity element 1, i.e. an element 1 such that $1 \cdot x = x \cdot 1 = x$ for all $x \in R$. We usually omit the $\cdot$ for multiplication.

Definition 5.24. A ring is called a *field* if every nonzero element admits a multiplicative inverse, i.e. an element x^{-1} with $x^{-1}x = xx^{-1} = 1$.

Definition 5.25. A morphism of rings from R to S is a map $f\colon R \to S$ such that $f(xy) = f(x)f(y)$ and $f(x+y) = f(x)+f(y)$ for all $x, y \in R$, and $f(1) = 1$.

Definition 5.26. If k is a field, a k-algebra is a ring R together with an injective morphism $k \to R$. In other words, it is a ring containing "scalars"; for example $R = k[x_1, \ldots, x_n]$ has k embedded inside as the constant polynomials.

Definition 5.27. Given a ring R, an *ideal* is a subset $I \subset R$ such that $0 \in I$, $x+y \in I$ for all $x, y \in I$, and $xz \in I$ for all $x \in I$, $z \in R$.

Example 5.28. The subset $\{0\} \subset R$ is an ideal (the *zero ideal* 0). The subset $R \subset R$ is an ideal. Any ideal other than these two is called *proper*.

Lemma 5.29. *Let I be an ideal of R. We have $I = R$ if and only if $1 \in I$.*

Proof. If $I = R$ then $1 \in I$. Conversely, if $1 \in I$ then $1 \cdot x = x \in I$ for all $x \in R$, so $I = R$. $\square$

Lemma 5.30. *If $\varphi\colon R \to S$ is a morphism then $\ker \varphi = \{f \in R : \varphi(f) = 0\}$ is an ideal.*

Proof. If $f, g \in \ker \varphi$ then $\varphi(f) = \varphi(g) = 0$, so $\varphi(f+g) = \varphi(f) + \varphi(g) = 0$, which means $f + g \in \ker \varphi$. If $f \in \ker \varphi$ and $h \in R$ then $\varphi(fh) = \varphi(f)\varphi(h) = 0$. Together, these facts show that $\ker \varphi$ is an ideal. $\square$

Lemma 5.31. *Given any ideal $I \subset R$ there is a* quotient ring *R/I and a surjective morphism $\varphi\colon R \to R/I$ ("reduction modulo I") with* $\ker \varphi = I$.

Proof. The elements of R/I are equivalence classes of elements in R under the equivalence relation $f \sim g$ if and only if $f - g \in I$. Sum and product are defined in the unique way so as to make φ a morphism (i.e. $\varphi(f) + \varphi(g) = \varphi(f+g)$ and $\varphi(f)\varphi(g) = \varphi(fg)$). It is an easy exercise to check that the resulting structure is well-defined and a ring. $\square$

Theorem 5.32 (First isomorphism theorem for rings)**.** *If $\varphi\colon R \to S$ is a morphism then* $\mathrm{im}(\varphi) \cong R/\ker \varphi$.

6 Algebra and geometry, II

Intersections between curves consist of finite sets of points. It will therefore be important for us to understand which ideals define finite algebraic sets.

6.1 Maximal ideals

Example 6.1. Take $I = (x_1, \ldots, x_n)$. Then $\mathbb{V}(I) = \{(0, \ldots, 0)\}$, as the origin is the only place where all the coordinates vanish. More generally, $\mathbb{V}(x_1 - a_1, \ldots, x_n - a_n) = \{(a_1, \ldots, a_n)\}$.

These algebraic sets correspond to imposing as many constraints as possible whilst still admitting a solution; in other words $\mathbb{I}(\{\boldsymbol{a}\})$ feels like it should be a *maximal ideal*.

Definition 6.2. An ideal $I \subset R$ is called *maximal* if $I \neq R$ and there is no ideal $J \subset R$ with $I \subsetneq J \subsetneq R$.

In fact, points correspond to maximal ideals only if we work over an algebraically closed field (like $\mathbb{C}$). For example, $(x^2 + 1) \subset \mathbb{R}[x]$ is a maximal ideal but does not correspond to a real point: $x^2 + 1 = 0$ has no real solutions. To show that maximal ideals correspond to points when we work over an algebraically closed field, we will prove a sequence of lemmas.

Lemma 6.3. *We have $f(\boldsymbol{a}) = 0$ if and only if $f \in (x_1 - a_1, \ldots, x_n - a_n)$.*

Proof. By changing variables to $x_1' = x_1 - a_1, \ldots, x_n' = x_n - a_n$, we can reduce to the case $\boldsymbol{a} = (0, \ldots, 0)$. This reduces the problem to showing that $f \in (x_1', \ldots, x_n')$ if and only if $f(0, \ldots, 0) = 0$. But $f(0, \ldots, 0)$ is the constant term in f. The constant term of f vanishes if and only if every monomial in f is divisible by at least one of $x_1', \ldots, x_n'$, which happens if and only if $f \in (x_1', \ldots, x_n')$. □

Lemma 6.4. *An ideal $I \subset R$ is maximal if and only if the quotient ring R/I is a field.*

Proof. We separate this into some subclaims.

Claim 1: A ring S is a field if and only if it has no proper ideals.

Claim 2: The ideals of R/I correspond bijectively with the ideals of R containing I.

From these two claims, the lemma will follow: $S = R/I$ is a field if and only if it has no proper ideals, if and only if there are no ideals of R containing I other than R and I.

Proof of Claim 1: Suppose that S is a field and $J \subset S$ is a nonzero ideal then it contains some nonzero element f. Since f has an inverse, $1 = f^{-1}f \in J$, and hence $J = S$. So J is either 0 or S. Conversely, if S has no proper ideals and $f \in S$ is nonzero then $(f) \subset S$ is a nonzero ideal, hence $(f) = S$ and hence $1 \in (f)$. This means $1 = gf$ for some $g \in S$, and hence f is invertible.

Proof of Claim 2: Let $q \colon R \to R/I$ be the quotient map. The correspondence sends an intermediate ideal $I \subset J \subset R$ to $q(J) \subset R/I$. It is an exercise to show that this gives a bijection, as claimed. □

Lemma 6.5. *For any $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{A}^n(k)$, the ideal*

$$(x_1 - a_1, \ldots, x_n - a_n) \subset k[x_1, \ldots, x_n]$$

is maximal.

Proof. Write $R = k[x_1, \ldots, x_n]$ and I for the ideal $(x_1 - a_1, \ldots, x_n - a_n)$, and consider the quotient R/I. By Lemma 6.4, it suffices to show that R/I is a field. Consider the map $\mathrm{ev}_{\boldsymbol{a}} \colon R \to k$ given by evaluating polynomials at $\boldsymbol{a}$, that is $\mathrm{ev}_{\boldsymbol{a}}(f) = f(\boldsymbol{a})$. This is a ring homomorphism:

$$\mathrm{ev}_{\boldsymbol{a}}(f + g) = \mathrm{ev}_{\boldsymbol{a}}(f) + \mathrm{ev}_{\boldsymbol{a}}(g), \qquad \mathrm{ev}_{\boldsymbol{a}}(fg) = \mathrm{ev}_{\boldsymbol{a}}(f)\mathrm{ev}_{\boldsymbol{a}}(g).$$

It is surjective: for any $b \in k$, the constant polynomial b evaluates to b at $\boldsymbol{a}$. The kernel consists of those polynomials which evaluate to 0 at $\boldsymbol{a}$, which is precisely I by Lemma 6.3. Now the first isomorphism theorem for rings tells us that $R/I \cong k$. □

The converse to this lemma holds if k is algebraically closed.

Theorem 6.6. *If k is algebraically closed then any maximal ideal of $k[x_1, \ldots, x_n]$ has the form $(x_1 - a_1, \ldots, x_n - a_n)$.*

Proof. We will prove it only for $k = \mathbb{C}$, or more generally for uncountable fields, where we can get away with a minimum of technology from the theory of field extensions.

If $I \subset R$ is maximal then R/I is a field by Lemma 6.4. Since R/I is generated by $x_1, \ldots, x_n \mod I$, it is spanned (as a vector space over k) by the countable set of monomials $x_1^{m_1} \cdots x_n^{m_n}$. So R/I is countable-dimensional as a k-vector space.

For each $x \in R/I$, the set of elements $\{\frac{1}{x-c} : c \in k\}$ is uncountable[12]. Since R/I is countable-dimensional, they must be linearly dependent, so there is a linear dependence

$$\frac{a_1}{x - c_1} + \cdots + \frac{a_n}{x - c_n} = 0$$

for some $a_i, c_i \in k$. Adding fractions, we rewrite this as

$$\frac{p(x)}{(x - c_1) \cdots (x - c_n)} = 0,$$

for some polynomial p. This implies $p(x) = 0$, so x is "algebraic" over k (i.e. satisfies a polynomial equation). Since k is algebraically closed, this implies $x \in k$. Therefore $R/I = k$.

This means that each $x_i \in R$ is equivalent modulo I to some element $a_i \in k$, that is

$$x_i - a_i = 0 \mod I, \quad i = 1, 2, \ldots, n.$$

This implies that $x_i - a_i \in I$ for $i = 1, \ldots, n$, which implies that I contains the maximal ideal $(x_1 - a_1, \ldots, x_n - a_n)$. Since $(x_1 - a_1, \ldots, x_n - a_n)$ is maximal, this implies that $I = (x_1 - a_1, \ldots, x_n - a_n)$ and we are done. □

[12] This is where we're using the additional assumption that k is uncountable.

This result actually implies the weak Nullstellensatz.

Proof of the weak Nullstellensatz. If $I \neq R$ then I is contained in *some* maximal ideal[13], and hence there exists $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{A}^n(k)$ such that $I \subset (x_1 - a_1, \ldots, x_n - a_n)$. In other words, even after adding extra constraints, $\boldsymbol{a}$ is still a solution of our system of equations. In particular, $\boldsymbol{a}$ was a solution of our original system, i.e. $\boldsymbol{a} \in \mathbb{V}_k(I)$, so $\mathbb{V}_k(I) \neq \emptyset$. □

Remark 6.7. In fact, the weak Nullstellensatz also implies Theorem 6.6: it will be an exercise to show this.

6.2 Finite sets of points

We have now seen that $\mathbb{I}(\{p\})$ is a maximal ideal; by induction, we can prove that $\mathbb{I}(\{p_1, \ldots, p_N\})$ is an intersection of N maximal ideals. We would now like to study the converse problem: for which ideals I is $\mathbb{V}(I)$ a finite set of points? This will become important when we study intersections of curves, because these are exactly the ideals that arise when your curves have a finite number of intersection points.

Example 6.8. Let $C = \{y - x^2\}$ and $C' = \{y - 2x^2\}$ be two parabolas which intersect precisely at the origin.

The ideal $I = (y - x^2, y - 2x^2)$ certainly has $\mathbb{V}(I) = \{(0,0)\}$, but it does not coincide with the maximal ideal (x, y). For example, the polynomial x is not in I, only its square. In fact, the polynomials 1 and x form a basis for the quotient ring $k[x, y]/I$.

We will prove that $\mathbb{V}(I)$ is finite if and only if $k[x, y]/I$ is finite-dimensional as a vector space over k. We start with a preparatory lemma.

Lemma 6.9. *Given a finite set of points $p_1, \ldots, p_N \in \mathbb{A}^n(k)$ there exists a polynomial g such that $g(p_1) = 1$ and $g(p_2) = \cdots = g(p_N) = 0$.*

Proof. Exercise! □

Theorem 6.10. *Let $R = k[x_1, \ldots, x_n]$ and $I \subset R$ be an ideal. Then $\mathbb{V}(I)$ is finite if and only if R/I is finite-dimensional over k.*

Proof. $\Leftarrow$: We will prove the stronger claim that the number of points in $\mathbb{V}(I)$ is bounded

[13] The set-theoretically heavy-handed way of proving this is as follows: if I is not itself maximal, it is contained in a bigger ideal. If the bigger ideal is not maximal, it is contained in a bigger ideal, etc. The union of all these ideals is still an ideal; if that is not maximal, it is contained in a bigger ideal... *ad transfinitum.* The fact that this outputs a maximal ideal is basically an axiom of set theory, with the suitably froody name of *Zorn's lemma.* Since we are working with something as mundane and finitary as polynomial rings, it should not be surprising that, with care, one can avoid transfinite induction; see for example the short paper Bernstein, J. (2012) *Elementary proof of the Nullstellensatz*, `http://www.math.tau.ac.il/~bernstei/Unpublished_texts/unpublished_texts/Elementary-proof-of-Nullstellensatz.pdf` – thanks to Y. Lekili for drawing it to my attention.

above by $\dim_k(R/I)$. If $p_1, \ldots, p_N \in \mathbb{V}(I)$ are distinct points then, by Lemma 6.9, we can find polynomials $g_1, \ldots, g_N$ with $g_i(p_j) = \delta_{ij}$. These define linearly independent classes in R/I: if $\sum_{i=1}^N \alpha_i g_i = 0 \mod I$ then $\sum_{i=1}^N \alpha_i g_i$ vanishes on $\mathbb{V}(I)$, so $0 = \sum_{i=1}^N \alpha_i g_i(p_j) = \alpha_j$ for all j. This gives N linearly independent elements of R/I, so $N \leq \dim_k(R/I)$ as required.

$\Rightarrow$: Suppose $\mathbb{V}(I) = \{p_1, \ldots, p_N\}$. We will write

$$p_i = (a_{i1}, \ldots, a_{in}).$$

Then

$$\begin{aligned} f_1 &:= (x_1 - a_{11}) \cdots (x_1 - a_{N1}) = x_1^N + \cdots \\ f_2 &:= (x_2 - a_{12}) \cdots (x_2 - a_{N2}) = x_2^N + \cdots \\ &\vdots \\ f_n &:= (x_n - a_{1n}) \cdots (x_n - a_{Nn}) = x_n^N + \cdots \end{aligned}$$

all vanish on $\mathbb{V}(I)$. By the Nullstellensatz, for each i, there exists r_i such that $f_i^{r_i} \in I$. This means that, modulo I, the monomial $x_i^{Nr_i}$ can be written in terms of lower powers of x_i. Therefore R/I is generated over k by the monomials $x_1^{i_1} \cdots x_n^{i_n}$ with

$$0 \leq i_1 \leq Nr_1 - 1, \qquad \cdots, \qquad 0 \leq i_n \leq Nr_n - 1,$$

so $\dim_k(R/I)$ is finite. $\square$

7 Singularities of affine curves

7.1 Singularities

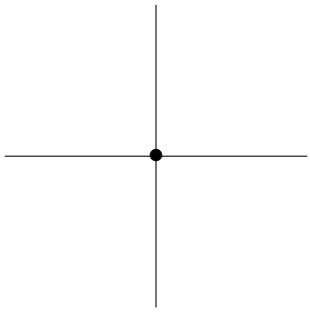
The next things we focus on are the singular points; these are easy to identify from the equation of the curve.

Definition 7.1. Let $C = \{f = 0\}$ be an algebraic curve. A point $p \in C$ is said to be *singular* or *multiple* if $\frac{\partial f}{\partial x}(p) = \frac{\partial f}{\partial y}(p) = 0$. Otherwise, $p \in C$ is said to be a *smooth* point[14].

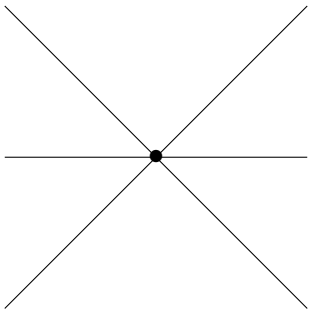
Example 7.2 (Nodal singularity)**.** Let $f(x, y) = xy$. The point $(0, 0)$ is a singularity of $\{f = 0\}$. This is because

$$\frac{\partial f}{\partial x} = y, \qquad \frac{\partial f}{\partial y} = x,$$

both of which vanish at $(0, 0)$. In fact, this is the only point where both partial derivatives vanish, so all the other points are smooth. This is called a *nodal* or A_1 singularity.



Example 7.3 (D_4 singularity)**.** Let $f(x, y) = y^3 - x^2y$. Again, the origin is the only singular point; we call this a D_4 singularity. The real locus of the variety looks like this:



Example 7.4. Consider $\mathbb{V}_k((x^2 - 1)(x - 1)^2 + (y^2 - 1)^2)$. We have

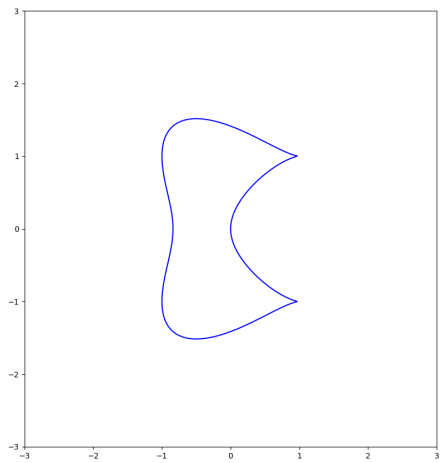
$$f = (x^2-1)(x-1)^2+(y^2-1)^2, \quad \partial f/\partial x = 2x(x-1)^2+2(x^2-1)(x-1), \quad \partial f/\partial y = 4y(y^2-1).$$

The $\partial f/\partial x = 0$ equation is:

$$(2x^2 - x - 1)(x - 1)$$

which has roots at $x = -1/2, 1$. The $\partial f/\partial y$ equation has roots at $y = -1, 0, 1$. This gives six possible singular points (x, y), but we still need to impose the equation $f = 0$. The only two which satisfy this final equation are $(1, -1)$ and $(1, 1)$. This is called the *bicuspid curve*.

[14]The words *simple*, *regular* and *nonsingular* are also used for this.



Remark 7.5. You might ask why Definition 7.1 is the right one to make: how does it capture our intuition that a smooth point is one where the curve looks locally like a straight line? This is because of a result called the *implicit function theorem*. We will illustrate the theorem in the following special case. Suppose that $f(x, y)$ is a polynomial with $f(0, 0) = 0$ and $\frac{\partial f}{\partial y}(0, 0) = \beta \neq 0$. Then there exists an analytic function $y(x)$ such that $x \mapsto (x, y(x))$ is a parametrisation of the curve $\{f = 0\}$ in a neighbourhood of $(0, 0)$. Since analytic functions are differentiable, the curve is well-approximated by its tangent line at this point.

Rather than proving the implicit function theorem in full gory detail, I will just explain how to extract the power series expansion $y(x) = a_1 x + a_2 x^2 + \cdots$. Substituting into $f(x, y(x))$, the first order part of f is

$$\alpha x + \beta a_1 x, \qquad \alpha = \frac{\partial f}{\partial x}(0, 0), \beta = \frac{\partial f}{\partial y}(0, 0),$$

which should vanish for all x, so we take $a_1 = -\alpha/\beta$ (here we are using $\beta \neq 0$). Now we take the second order term:

$$\beta a_2 x^2 + \gamma x^2 + \delta a_1 x^2 + \epsilon a_1^2 x^2,$$

where γ, δ, ϵ are the coefficients of x^2, xy, y^2 in f. This tells us to take

$$a_2 = -\frac{1}{\beta}(\gamma + a_1 \delta + a_1^2 \epsilon).$$

This gives us a recursive way to compute the coefficients a_i. Each time, we just need to divide by β, which we're assuming is nonzero.

7.2 Multiplicity

In some sense, the D_4-singularity looks worse than the A_1-singularity, because there are three lines meeting at the singular point. We can make precise sense of this through the notion of *multiplicity*.

Definition 7.6. The *multiplicity* of a point $p \in C$ is the maximal nonnegative integer k such that all partial derivatives of f of order strictly less than k vanish. Here, f is considered as the zeroth derivative of f. We call a point with multiplicity 2 a *double point*, multiplicity 3 a *triple point*, etc. (Hence the term "multiple point" for singularity).

Example 7.7. This means

- p has multiplicity 0 if and only if $p \notin C$ (because $f(p) \neq 0$)
- p has multiplicity 1 if and only if it is a smooth point of C ($f(p) = 0$ but the first derivatives do not all vanish).
- p is singular if and only if its multiplicity is 2 or more.
- The nodal singularity $p = (0, 0)$ of $xy = 0$ has multiplicity 2 because, although the first derivatives both vanish at p, the mixed second derivative $\partial^2 f/\partial x \partial y = 1$ does not.
- The singularity $(0, 0)$ of $\{y^3 - x^2 y = 0\}$ has multiplicity 3, because the derivatives up to second order are:

$$f = y^3 - x^2 y, \qquad \frac{\partial f}{\partial x} = -2xy, \qquad \frac{\partial f}{\partial y} = 3y^2 - x^2,$$
$$\frac{\partial^2 f}{\partial x^2} = -2y \qquad \frac{\partial^2 f}{\partial x \partial y} = -2x, \qquad \frac{\partial^2 f}{\partial y^2} = 6y,$$

  all of which vanish at $(0, 0)$, but the third derivative $\partial^3 f/\partial y^3 = 6$ does not.

7.3 Tangencies

Let $C = \{f = 0\}$ be an algebraic curve and let $p \in C$. We can change coordinates by translation to put p at the origin which amounts to replacing f by its Taylor series (which is again a polynomial). If p has multiplicity m then the Taylor series of f at p starts with terms of degree m. Let us group these terms and write $f(x, y) = f_m(x, y) + \cdots$ where $\cdots$ stands for higher order terms.

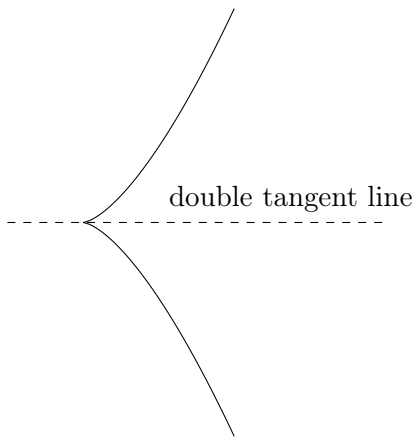
Definition 7.8. A straight line through p is called a *tangent* to C at p if f_m vanishes on this line. If the line has slope λ (so it is given by $y = \lambda x$ in coordinates centred at p) then this means $f_m(x, \lambda x) = 0$ for all x; the only other line is $x = 0$ (corresponding to $\lambda = \infty$) which gives $f_m(0, y) = 0$.

In our earlier examples, the tangent lines are the obvious lines in the pictures.

Example 7.9. When $f(x, y) = xy$, we have $m = 2$ and $f_m(x, y) = f(x, y) = xy$. This vanishes on $y = \lambda x$ if and only if $\lambda x^2 = 0$ for all x, i.e. $\lambda = 0$. It also vanishes on $(0, y)$. The lines with slope 0 and ∞ are precisely the x and y axes.

Example 7.10. When $f(x, y) = y^3 - x^2 y$ we have $m = 3$ and $f_m(x, y) = f(x, y)$. This vanishes on the three lines $y = x$, $y = -x$ and $y = 0$.

Example 7.11. A more interesting example is $f(x, y) = y^2 - x^3$. This has $m = 2$ (because the second derivative $\partial^2 f/\partial y^2 = 2$ does not vanish at the singularity) and $f_2(x, y) = y^2$. This vanishes on the line $y = 0$. In fact, it vanishes with multiplicity 2.



Theorem 7.12. *If $m_p(C) = m$ then there are m tangent lines to C at p, when counted with multiplicities (working over an algebraically closed field k).*

Proof. In the course of the proof, we will explain why we mean by "counted with multiplicities". Write $f_m(x, y) = a_0x^m + a_1x^{m-1}y + \cdots + a_my^m$. Assume first that $a_m \neq 0$; this ensures that $x = 0$ is not a tangent line because $f_m(0, y) = a_my^m$ which is not identically zero. Therefore the possible tangent lines are $y = \lambda x$, $\lambda \in k$. Substituting, we get $f_m(x, \lambda x) = x^m(a_0 + a_1\lambda + \cdots + a_m\lambda^m)$. This is identically zero if and only if λ is a root of the polynomial $a_m\lambda^m + \cdots + a_1\lambda + a_0 = 0$. This polynomial has m roots, counted with multiplicities, since it is a polynomial of degree m (here we are using the fact that k is algebraically closed).

If $a_m = 0$ then suppose a_j is the nonvanishing coefficient with j maximal, so $f_m(x, y) = x^{m-j}(a_0x^j + a_1x^{j-1}y + \cdots + a_jy^j)$. We say that $x = 0$ is a tangent line with multiplicity $m - j$, and as before we get a total of j possible slopes λ (counted with multiplicity as a root of $a_0 + a_1\lambda + \cdots + a_j\lambda^j$) for which $y = \lambda x$ is a tangent line. □

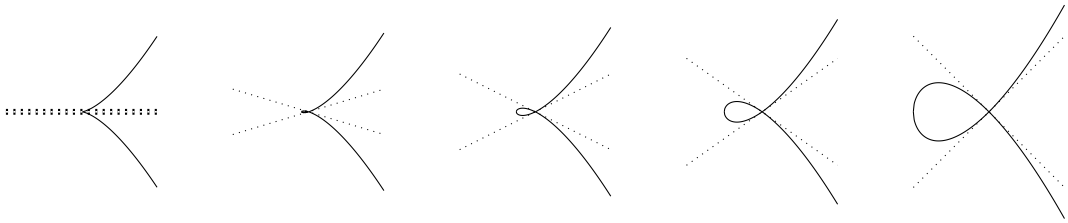
Example 7.13. Consider the family of curves $y^2 = x^3 + tx^2$, $t \in [0, 1]$. The first derivatives of the defining equation $F_t(x, y) = y^2 - x^3 - tx^2 = 0$ are

$$-3x^2 - 2tx, \qquad 2y,$$

which vanish at $(0, 0)$ and $(-2t/3, 0)$. The function F_t only vanishes at the first of these, so for all t the curve $\{F_t = 0\}$ has precisely one singularity (at the origin). This has multiplicity 2 because the second derivative $\partial^2F_t/\partial y^2 = 2$ is never zero. The Taylor expansion of F_t at the origin is $F_{t,2} + F_{t,3}$ where

$$F_{t,2} = y^2 - tx^2, \quad F_{t,3} = -x^3.$$

The tangent lines occur where $F_{t,2}$ vanishes, i.e. $y = \pm x\sqrt{t}$. This means there are two distinct tangent lines when $t \neq 0$ and one tangent line with multiplicity 2 when $t = 0$.

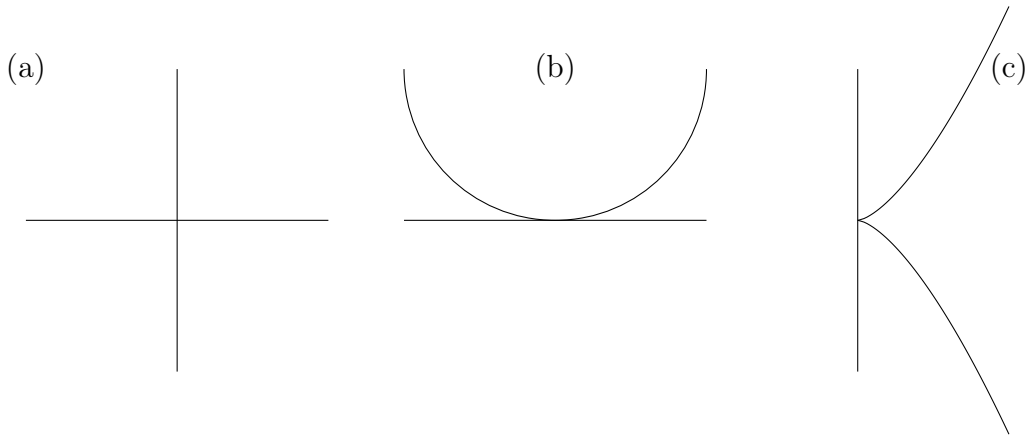


8 Intersection theory, I

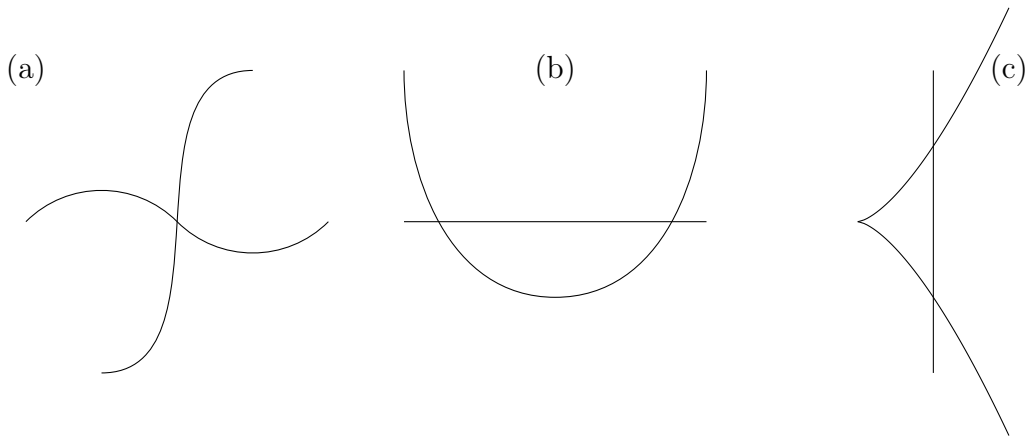
8.1 Transversality

Definition 8.1. If $p \in C$ is a smooth (simple) point then C has a unique tangent line at p. We write T_pC for this line and call it *the tangent space of C at p.*

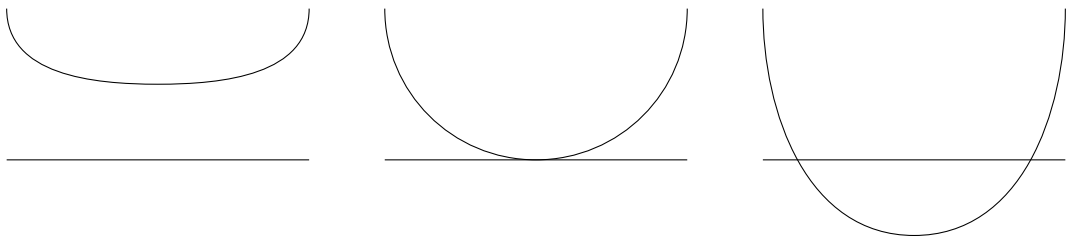
Definition 8.2. We say that two curves C and C' intersect transversely at a point p if p is a smooth point of both curves, and the tangent lines T_pC and T_pC' are different. In the figure below, the intersection depicted in (a) is transverse, while the intersections depicted in (b) and (c) are not.



It is intuitively clear that if you have a pair of curves intersecting transversely, when you deform the curves by a small amount then this intersection point stays transverse. By contrast, when there is a nontransverse intersection, it looks like one can wiggle the curves around so that they intersect transversely in a bunch of points.



In cases (b) and (c), we perturbed a single nontransverse intersection and obtained two transverse intersections. You need to be careful with these real pictures: there could be intersections hiding at complex points. Let's examine this example in more detail. Suppose the curves in case (b) are $C_t = \{y = x^2 + t\}$ and $C' = \{y = 0\}$. These intersect at the points $(\pm\sqrt{-t}, 0)$. When $t = 0$ we have the single nontransverse intersection $(0, 0)$: both curves are tangent to the x-axis. When $t < 0$ there are two transverse intersections at points on the real x-axis. When $t > 0$ the two intersections are at $(\pm i\sqrt{t}, 0)$, so do not show up in the real picture.



If we are given a pair of curves which intersect nontransversely at a point p, we would like to be able to count how many transverse intersections should "appear" when we perturb the equations. This is called the *local intersection multiplicity*, $i_p(C, C')$ of C and C' at p. There are different approaches to defining this. We will define $i_p(C, C')$ as the dimension of a certain vector space called the *local ring* of $C \cap C'$ at p. This is traditionally defined using ideas from commutative algebra (localisation) but we will take a slightly different route, and define it using generalised eigenspaces of certain linear maps. We will see (Proposition 10.10) that nontransverse intersections always have $i_p(C, C') > 1$:

Proposition 8.3 (Propositions 10.10 and 10.11)**.** *Let $C = \{f = 0\}$ and $C' = \{g = 0\}$ be curves and $p \in C \cap C'$. If p is not transverse then $i_p(C, C') > 1$. In fact, when p is a singularity of C or C', we have the bound $i_p(C, C') \geq m_p(C)m_p(C')$.*

8.2 Interlude

The definition of $i_p(C, C')$ and the proofs of its basic properties will be somewhat technical, so, to keep you motivated throughout the next few sessions while we develop this theory, I want to start by giving you a quick sketch of what we will be able to do once we have developed enough intersection theory.

The first important result we will need (and one of the last we'll prove) is[15]:

Theorem 8.4 (Bézout's theorem)**.** *If $C = \{f = 0\}$ and $C' = \{g = 0\}$ are curves of degree m and n respectively, with $\gcd(f, g) = 1$, then*

$$\sum_{p \in C \cap C'} i_p(C, C') \leq mn.$$

Without knowing anything more than these properties of $i_p(C, C')$, we can prove some fun things.

Theorem 8.5. *An irreducible cubic curve C can have at most one double point, and no points with higher multiplicity.*

Proof. Suppose $p \in C$ is a point with multiplicity m and $q \in C$ is a point with multiplicity m'. Let L be the line through p and q (considered as a curve of degree 1). Then Bézout's theorem tells us that

$$3 \times 1 \geq \sum_{r \in C \cap C'} i_p(C, C') \geq i_p(C, C') + i_q(C, C') = m + m'.$$

[15] You may know another result called Bézout's theorem: namely that in a Euclidean domain the greatest common divisor of two elements can be written as a linear combination of those elements. That is a completely unrelated (and much easier) result, which we will henceforth relegate to the status of *Bézout's lemma*.

Since $m, m' \geq 1$, the only possibilities are

$$m = m' = 1, \qquad m = 1, m' = 2 \qquad m' = 2, m = 1. \qquad \square$$

We were able to apply Bézout's theorem because irreducibility of the cubic means that it does not contain L as an irreducible component, that is the defining cubic is not divisible by a linear factor.

Theorem 8.6. *Suppose that C is an irreducible quartic curve. Then one of the following holds:*

- *C is smooth;*
- *C has one triple point and no other singularities,*
- *C up to three double points and no other multiple points.*

If C has three double points then they cannot all lie on a line.

Proof. If $p, q \in C$ have multiplicities m, m', then intersecting with the line through p, q and applying Bézout's theorem gives us $4 \geq m+m'$. Since $m, m' \geq 1$, the only possibilities (assuming without loss of generality that $m \leq m'$) are:

$$m = m' = 1, \qquad m = 1, m' = 2, \qquad m = 1, m' = 3, \qquad m = 2, m' = 2.$$

Thus:

- Any point has multiplicity ≤ 3.
- If q is a triple point any other point p must be smooth.

If p and q are double points ($m = m' = 2$) then no other point of C can lie on L; in particular three double points cannot lie on a line.

The final thing to show is that there are at most three double points. Since these do not need to (indeed, cannot) lie in a line, we need to use a different kind of auxiliary curve. We will see later that through any choice of five points there exists a conic curve. If there were four double points $q_1, \ldots, q_4$ on C, we could pick a conic passing through these q_i and a fifth point (multiplicity at least 1). Bézout's theorem would give

$$8 = 2 \times 4 \geq 2 + 2 + 2 + 2 + 1 = 9,$$

which is impossible. $\square$

You can imagine that these are just the first in an infinite sequence of ever-more-complicated constraints on configurations of singularities for curves of higher and higher degree which follow from Bézout's theorem (and the existence of higher degree auxiliary curves). One of the MATH323 projects will be to explore these constraints (for quintics, sextics, and to see what you can say for arbitrary degree).

There are many more applications of Bézout's theorem, but I will now save them until after we have developed all the necessary theory and proved it.

9 Intersection theory, II

9.1 Basic idea and an example

Let $C = \{f = 0\}$ and $C' = \{g = 0\}$ be two curves and suppose they do intersect at a finite number of points[16]. How do we find their intersections? We can reduce this problem to finding the eigenvalues of a certain matrix.

Definition 9.1. Let $R = k[x, y]$ and $I \subset R$ be an ideal (for example $I = (f, g)$ for f and g defining two curves). Let $S = R/I$ be the quotient ring obtained by identifying polynomials which agree modulo I. Each element $f \in R$ defines a linear map

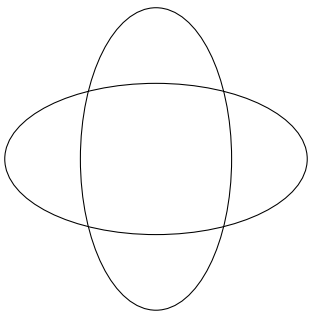
$$\hat{f} \colon S \to S, \qquad \hat{f}(g) = fg.$$

Remark 9.2. This linear map is particularly useful if S is finite-dimensional, which we saw in Theorem 6.10 happens if and only if $\mathbb{V}(I)$ is a finite set of points.

Theorem 9.3. *If $p = (a, b) \in \mathbb{V}(I)$ then a is an eigenvalue of $\hat{x}$ and b is an eigenvalue of $\hat{y}$. In fact, the a, b are simultaneous eigenvalues for this pair of matrices, that is the generalised a-eigenspace of $\hat{x}$ and the generalised b-eigenspace of $\hat{y}$ intersect in a nonzero subspace. Conversely, any pair of simultaneous eigenvalues tell us the coordinate of a point in $\mathbb{V}(I)$.*

We will recap generalised eigenspaces in a moment. When we prove the theorem, we will do so more generally for $I \subset k[x_1, \ldots, x_n]$. Before proving the theorem, we illustrate it with an example.

Example 9.4. Consider the ellipses $C = \{x^2 + 2y^2 = 1\}$ and $C' = \{2x^2 + y^2 = 1\}$.



We can find their intersections in the traditional way as follows: if (x, y) is a point satisfying both equations then subtracting one equation from the other we obtain

$$x^2 - y^2 = 0$$

so $x = y$ or $x = -y$. Substituting $x^2 = y^2$ in $x^2 + 2y^2 = 1$ we get $3x^2 = 1$, which gives $x = \pm 1/\sqrt{3}$ so overall we have four intersection points $(\pm 1/\sqrt{3}, \pm 1/\sqrt{3})$.

Now let's do it again using eigenvalues. Let $I = (x^2 + 2y^2 - 1, 2x^2 + y^2 - 1)$ be the ideal generated by the equations of C and C' and let S be the quotient ring $k[x, y]/I$ (i.e. the ring of polynomials considered up to equivalence modulo I: we will write $=$ for equality modulo I). In this quotient ring, we know that $x^2 = y^2 = 1/3$ from the manipulations

[16] We will see on Sheet 5 that the curves $\{f = 0\}$ and $\{g = 0\}$ intersect in a finite set of points if $\gcd(f, g) = 1$.

above. Therefore any monomial $x^m y^n$ is equivalent modulo I to a power of $1/3$ times one of $1, x, y, xy$ (if, m and n are, respectively, even/even, odd/even, even/odd, odd/odd). In fact, $1, x, y, xy$ form a basis for S as a vector space over k. This means that each polynomial acts as a 4-by-4 matrix on S with respect to this basis.

We calculate $\hat{x}$:

$$\hat{x}1 = x, \quad \hat{x}x = x^2 = 1/3, \quad \hat{x}y = xy, \quad \hat{x}xy = x^2y = y/3$$

which corresponds to the matrix

$$\hat{x} = \begin{pmatrix} 0 & 1/3 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Similarly, we get

$$\hat{y} = \begin{pmatrix} 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

The characteristic polynomial of $\hat{x}$ is

$$(\lambda^2 - 1/3)^2$$

so we see that the possible values of x on the points of $C \cap C'$ are $\pm 1/\sqrt{3}$. Similarly we find the possible values of y to be $\pm 1/\sqrt{3}$.

Remark 9.5. In this example, we were lucky that every possible combination (λ, μ) with λ an eigenvalue of $\hat{x}$ and μ an eigenvalue of $\hat{y}$ actually corresponded to a solution of our system. This is not always the case, as the following modification of the previous example shows.

Example 9.6. We saw in the last example that the triple intersection points between the ellipses C, C' and the line $C'' = \{x = y\}$ are $(-1/\sqrt{3}, -1/\sqrt{3})$ and $(1/\sqrt{3}, 1/\sqrt{3}$. If we work with the bigger ideal J generated by all three equations then the quotient $S = k[x, y]/J$ is 2-dimensional with basis $1, x$: the monomial y is now equivalent to x and the monomial xy is equivalent to x^2 and hence to $1/3$. We calculate

$$\hat{x} = \hat{y} = \begin{pmatrix} 0 & 1/3 \\ 1 & 0 \end{pmatrix},$$

and both matrices still have the eigenvalues $\pm 1/\sqrt{3}$. But this calculation doesn't rule out the possibility of $(1/\sqrt{3}, -1/\sqrt{3})$ being an intersection point.

To rule this out, we need to look at the *simultaneous eigenspaces* (more precisely the *simultaneous generalised eigenspaces*) of $\hat{x}$ and $\hat{y}$. For example, the vector $(1, \sqrt{3})$ (corresponding to the polynomial $1 + x/\sqrt{3}$) is an eigenvector of both $\hat{x}$ and $\hat{y}$ with eigenvalues $1/\sqrt{3}$ and $1/\sqrt{3}$; this corresponds to the intersection point $(1/\sqrt{3}, 1/\sqrt{3})$. The vector $(1, -\sqrt{3})$ is an eigenvector of both $\hat{x}$ and $\hat{y}$ with eigenvalues $-1/\sqrt{3}$ and $-1/3\sqrt{3}$, corresponding to the intersection point $(-1/\sqrt{3}, -1/\sqrt{3})$. It is this which picks out the points $(1/\sqrt{3}, 1/\sqrt{3})$ and $(-1/\sqrt{3}, -1/\sqrt{3})$ rather than $(1/\sqrt{3}, -1/\sqrt{3})$ or $(-1/\sqrt{3}, 1/\sqrt{3})$.

9.2 Recap on generalised eigenspaces

Definition 9.7. Recall that a vector $v \neq 0$ is called a *generalised λ-eigenvector* of a linear map M if $(M - \lambda\,\mathrm{id})^n v = 0$ for some n. The generalised λ-eigenspace of M is the space of all generalised λ-eigenvectors of M (together with zero).

We call $v \neq 0$ an *honest eigenvector* (with eigenvalue λ) if $Mv = \lambda v$. Certainly an honest eigenvector is a generalised eigenvector, but there may be more: for example, if $M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ then the first basis vector e_1 is an honest eigenvector with eigenvalue 0, while e_2 and e_3 are generalised eigenvectors with eigenvalue 0. We start with a simple lemma:

Lemma 9.8. *Any generalised eigenspace contains at least one honest eigenvector.*

Proof. This will be an exercise! □

Theorem 9.9. *Let V be a vector space over an algebraically closed field k. If $M\colon V \to V$ is a k-linear map then there is a finite subset* $\mathrm{spec}(M) \subset k$ *(the* spectrum *of M) such that*

$$V = \bigoplus_{\lambda \in \mathrm{spec}(M)} V_\lambda,$$

where V_λ is the generalised λ-eigenspace of M.

Proof. This follows from the fact that M can be put into Jordan normal form: if $e_1, \ldots, e_n$ is a basis with respect to which M is in Jordan normal form then the generalised eigenspace with eigenvalue λ is just the span of all basis vectors e_i for which the corresponding matrix entry M_{ii} equals λ. Since you never actually saw a proof of this in MATH220, I give an alternative proof of the theorem in the appendix. □

Lemma 9.10. *If $M_1, M_2\colon V \to V$ are two commuting linear maps and V_λ is the generalised λ-eigenspace of M_1 then V_λ is invariant under M_2 (i.e. $v \in V_\lambda$ implies $M_2 v \in V_\lambda$).*

Proof. Suppose $(M_1 - \lambda\,\mathrm{id})^n v = 0$. Multiplying on the left by M_2, we get $0 = M_2(M_1 - \lambda id)^n v$. Since M_2 commutes with M_1 (by assumption) and id (always) this becomes

$$(M_1 - \lambda\,\mathrm{id})^n M_2 v = 0,$$

so we see that $M_2 v$ is another generalised λ-eigenvector of M_1, as required. □

Corollary 9.11. *If $M_1, M_2, \ldots, M_n\colon V \to V$ are pairwise commuting linear maps then there is a finite set S of n-tuples $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ such that $V = \bigoplus_{\boldsymbol{\lambda} \in S} V_{\boldsymbol{\lambda}}$, where*

$$V_{\boldsymbol{\lambda}} = \{v \in V \,:\, \exists\, (\ell_1, \ldots, \ell_n) \;\; s.t. \;\; (M_i - \lambda_i\,\mathrm{id})^{\ell_i} v = 0 \;\forall i\}.$$

Proof. Apply Theorem 9.9 to M_1 to split $V = \bigoplus_{\lambda_1 \in \mathrm{spec}(M_1)} V_{\lambda_1}$ as a direct sum of generalised eigenspaces of M_1. Now each M_i, $i \geq 2$, preserves each generalised eigenspace V_{λ_1} by Lemma 9.10. For each $\lambda_1 \in \mathrm{spec}(M_1)$, apply Theorem 9.9 to $M_2|_{V_{\lambda_1}} : V_{\lambda_1} \to V_{\lambda_1}$, to split V_{λ_1} into simultaneous generalised eigenspaces of M_1 and M_2. If $n = 2$, we get precisely the statement of the corollary; otherwise, proceed in the same manner with M_3, M_4, etc. □

Lemma 9.12. *Each simultaneous generalised eigenspace contains an honest simultaneous eigenvector, i.e. a vector v such that $M_1 v = \lambda_1 v, \ldots, M_n v = \lambda_n v$.*

Proof. Exercise! □

This will be useful because generalised eigenvectors can be more of a pain to work with.

Definition 9.13. We call S the *set of simultaneous eigenvalues* or *joint spectrum* of $M_1, \ldots, M_n$. We will write $\operatorname{spec}(M_1, \ldots, M_n)$ for the joint spectrum. We call the subspaces $V_{\boldsymbol{\lambda}}$ the *generalised simultaneous eigenspaces.*

9.3 The proof

Throughout this section, we write $R = k[x_1, \ldots, x_n]$, fix an ideal $I \subset R$ and write S for the quotient ring R/I. We will assume that S and all the vector spaces that get mentioned along the way are finite-dimensional. We will prove the following theorem, which implies Theorem 9.3 upon taking $n = 2$.

Theorem 9.14. *We have* $\operatorname{spec}(\hat{x}_1, \ldots, \hat{x}_n) = \mathbb{V}(I)$ *and* $S = \bigoplus_{\boldsymbol{a} \in \mathbb{V}(I)} S_{\boldsymbol{a}}$ *where* $S_{\boldsymbol{a}}$ *for* $\boldsymbol{a} = (a_1, \ldots, a_n)$ *is the generalised simultaneous eigenspace of* $\hat{x}_1, \ldots, \hat{x}_n$ *with eigenvalues* $a_1, \ldots, a_n$.

The key thing to prove is $\operatorname{spec}(\hat{x}_1, \ldots, \hat{x}_n) = \mathbb{V}(I)$, which we will conclude in Corollary 9.18 below; the rest follows from Corollary 9.11. We now prove a sequence of lemmas, culminating in Corollary 9.18.

Given $f \in R$, we write $\hat{f} \colon S \to S$ for the linear map $\hat{f} g = fg$. The following properties are easy to check.

Lemma 9.15. *Given two polynomials $f_1, f_2 \in k[x_1, \ldots, x_n]$, we have the following properties:*

- *if* $f_1 = f_2 \mod I$ *then* $\hat{f}_1 = \hat{f}_2$.
- $\widehat{f_1 f_2} = \hat{f}_1 \hat{f}_2$.
- $\widehat{f_1 + f_2} = \hat{f}_1 + \hat{f}_2$.
- $\hat{f}_1 \hat{f}_2 = \hat{f}_2 \hat{f}_1$
- $\hat{f}_1 = f_1(\hat{x}, \hat{y})$.

In particular, the operators $\hat{x}_i$ commute with one another, so we can apply Corollary 9.11 with $M_1 = \hat{x}_1, \ldots, M_n = \hat{x}_n$. We write $\operatorname{Spec}(S)$ for the joint spectrum $\operatorname{spec}(\hat{x}_1, \ldots, \hat{x}_n)$ and decompose the ring

$$S = \bigoplus_{\boldsymbol{a} \in \operatorname{Spec}(S)} S_{\boldsymbol{a}}$$

into generalised eigenspaces for these matrices.

Lemma 9.16. *If $I \subset R$ is an ideal such that $S = R/I$ is finite-dimensional over k then*

$$\operatorname{Spec}(S) \subset \mathbb{V}(I).$$

Proof. Suppose that $\boldsymbol{a} = (a_1, \ldots, a_n) \in \operatorname{Spec}(S)$. By Lemma 9.12, this means that there exists a simultaneous $\boldsymbol{a}$-eigenvector $g \in S$:

$$\hat{x}_1 g = a_1 g, \ \ldots, \ \hat{x}_n g = a_n g.$$

Given any $f \in I$, we know that $\hat{f} g = fg = 0 \mod I$. We also know by Lemma 9.15 that $\hat{f} g = f(\hat{x}_1, \ldots, \hat{x}_n) g$. Since $\hat{x}_i g = a_i g$ for all i, we also deduce that $h(\hat{x}_1, \ldots, \hat{x}_n) g = h(a_1, \ldots, a_n) g$ for any polynomial $h(x_1, \ldots, x_n)$. Therefore we have

$$0 = \hat{f} g = f(\hat{x}_1, \ldots, \hat{x}_n g) = f(\boldsymbol{a}).$$

This shows $\boldsymbol{a} \in \mathbb{V}(I)$.

□

Proposition 9.17. *In the setting of Lemma 9.16, we also have*

$$\mathbb{V}(I) \subset \operatorname{Spec}(S).$$

Proof. By assumption, $\mathbb{V}(I)$ is finite. By Lemma 6.9, for each $\boldsymbol{a} \in \mathbb{V}(I)$, there exists a polynomial $g_{\boldsymbol{a}}$ such that

$$g_{\boldsymbol{a}}(\boldsymbol{a}) = 1, \qquad g_{\boldsymbol{a}}(\boldsymbol{b}) = 0 \text{ for all } \boldsymbol{b} \in \mathbb{V}(I) \setminus \{\boldsymbol{a}\}.$$

Then $(x_i - a_i) g_{\boldsymbol{a}}$ vanishes at all points of $\mathbb{V}(I)$, so by the Nullstellensatz

$$(x_i - a_i)^r g_{\boldsymbol{a}}^{r_i} = 0 \text{ for some } r_i.$$

Since $g_{\boldsymbol{a}}^{r_i}(\boldsymbol{a}) = 1$, $g_{\boldsymbol{a}} \neq 0$, so $g_{\boldsymbol{a}}^{r_i}$ is a nonzero generalised a_i-eigenvector of $\hat{x}_i$. The same argument works for each i. Let $r = \max(r_1, \ldots, r_n)$. Then $g_{\boldsymbol{a}}^r \in S$ is nonzero and annihilated by all of the linear maps $(\hat{x}_i - a_i)^{r_i}$. Therefore is is a nonzero generalised simultaneous $\boldsymbol{a}$-eigenvector, which shows that $\boldsymbol{a} \in \operatorname{Spec}(S)$. □

Corollary 9.18. *In the setting of Lemma 9.16,* $\operatorname{Spec}(S) = \mathbb{V}(I)$.

9.4 Appendix: Proof of Theorem 9.9

As you never actually saw a proof of Theorem 9.9 in MATH220, we include one here for completeness. In what follows, V will be a vector space over an algebraically closed field k, and $M \colon V \to V$ will be a linear map.

By a λ-*block*, we will mean a subspace $W \subset V$ invariant under M together with a basis $e_1, \ldots, e_m$ such that $M|_W$ is upper triangular with respect to this basis, with diagonal entries λ. Note that any λ-block consists of generalised λ-eigenvectors: $M - \lambda \operatorname{id}$ restricted to W is strictly upper triangular (zeros on the diagonal) so $(M - \lambda \operatorname{id})|_W^n = 0$ for some n.

We will show that V splits as a direct sum of λ-blocks for distinct λs. Our proof will be by induction on the dimension of V. When $\dim V = 1$ there is nothing to check. Suppose we have proved that for any vector space W of dimension at most n and any endomorphism of W, there is a splitting of W into λ-blocks. Suppose V has dimension $n+1$ and $M \colon V \to V$ is an endomorphism.

Pick an eigenvector v of M with some eigenvalue λ. Pick a subspace $W \subset V$ which is complementary to the span $\langle v \rangle$ of the chosen eigenvector. With respect to $\langle v \rangle \oplus W$, M is block upper triangular:

$$M = \begin{pmatrix} \lambda & b \\ 0 & M' \end{pmatrix}$$

where $M' \colon W \to W$ and $b \colon W \to \langle v \rangle$ are linear maps.

By our inductive hypothesis, W splits into μ-blocks W_μ under the action of M'. We analyse the cases $\lambda = \mu$ and $\lambda \neq \mu$ separately.

If $\lambda = \mu$ and the basis of the μ-block is $e_1, \ldots, e_m$ then $\langle v \rangle \oplus W_\mu$ is a λ-block with basis $v, e_1, \ldots, e_m$.

Suppose that $(W_\lambda, (e_1, \ldots, e_m))$ is a λ-block with $\lambda \neq \mu$. We will modify the vectors $e_1, \ldots, e_m$ to obtain a new λ-block with the property that

$$\operatorname{span}(v, e_1, \ldots, e_m) = \operatorname{span}(v, e_1', \ldots, e_m').$$

We have $Me_1 = \mu e_1 + b_1 v$ for some b_1. Set $e_1' = e_1 + b_1 v/(\mu - \lambda)$. We have

$$Me_1' = \mu e_1 + b_1 v + b_1 \lambda v/(\mu - \lambda) = \mu e_1'.$$

Now we have $Me_2 = \mu e_2 + \sigma_2 + b_2 v$ where σ_2 is a multiple of e_1'. Set $e_2' = e_2 + b_2 v/(\mu - \lambda)$. This satisfies $Me_2' = \mu e_2' + \sigma$. Now we have $Me_3 = \mu e_3 + \sigma_3 + b_3 v$ where σ_3 is a linear combination of e_1' and e_2'. Set $e_3' = e_3 + b_3 v/(\mu - \lambda)$. Continue in this manner all the way up to e_m'.

The result is a splitting of V into blocks, as required. Because the matrix of M is now block-diagonal, with each block having the form $\lambda \operatorname{id} +$strictly upper triangular, we see that:

- the blocks are precisely the generalised eigenspaces
- the eigenvalues are precisely the roots of the characteristic polynomial, and the dimension of the λ generalised eigenspace equals the multiplicity of λ as a root in the characteristic polynomial.

10 Intersection theory, III

10.1 Decomposition into local rings

Let $R = k[x_1, \ldots, x_n]$ and $I \subset R$ be an ideal such that $S = R/I$ is finite-dimensional. Theorem 9.14 asserts that S decomposes as a direct sum $S = \bigoplus_{\boldsymbol{a} \in \mathbb{V}(I)} S_{\boldsymbol{a}}$ of simultaneous generalised eigenspaces for $\hat{x}_1, \ldots, \hat{x}_n$.

Definition 10.1. We call $S_{\boldsymbol{a}}$ the *local ring* of $\mathbb{V}(I)$ at $\boldsymbol{a}$. The dimension of $S_{\boldsymbol{a}}$ as a k-vector space is called the *multiplicity* of $\mathbb{V}(I)$ at $\boldsymbol{a}$. If I represents the intersection of two curves C and C', we call this the *local intersection multiplicity* of C and C' at $\boldsymbol{a}$, and write it as $i_{\boldsymbol{a}}(C, C')$.

The local ring at $\boldsymbol{a}$ will be important in what follows, so we establish some of its basic properties. First, we observe that it is a ring:

Lemma 10.2. *Each subspace $S_{\boldsymbol{a}} \subset S$ is a subring.*

Proof. If $g_1, g_2 \in S_{\boldsymbol{a}}$ then there exist $\kappa_1, \ldots, \kappa_n$ and $\ell_1, \ldots, \ell_n$ such that

$$(\hat{x}_i - a_i)^{\kappa_i} g_1 = 0, \qquad (\hat{x}_i - a_i)^{\ell_i} g_2 = 0$$

for all $i = 1, \ldots, n$. Let $M_i = \max(\kappa_i, \ell_i)$ and $m_i = \min(\kappa_i, \ell_i)$. We have

$$(\hat{x}_i - a_i)^M (g_1 + g_2) = 0, \qquad (\hat{x}_i - a_i)^m (g_1 g_2) = 0$$

for all i, so $g_1 + g_2 \in S_{\boldsymbol{a}}$ and $g_1 g_2 \in S_{\boldsymbol{a}}$. □

Lemma 10.3. *The subrings $S_{\boldsymbol{a}}$ and $S_{\boldsymbol{b}}$ are orthogonal for $\boldsymbol{a} \neq \boldsymbol{b}$ in the sense that $g_1 \in S_{\boldsymbol{a}}$ and $g_2 \in S_{\boldsymbol{b}}$ satisfy $g_1 g_2 = 0$.*

Proof. Exercise. □

Definition 10.4. Since $S = \bigoplus_{\boldsymbol{a} \in \mathbb{V}(I)} S_{\boldsymbol{a}}$, we can write the identity as $1 = \sum_{\boldsymbol{a} \in \mathbb{V}(I)} e_{\boldsymbol{a}}$ for some uniquely determined elements $e_{\boldsymbol{a}} \in S_{\boldsymbol{a}}$. We call $e_{\boldsymbol{a}}$ the *idempotent for $S_{\boldsymbol{a}}$*.

Lemma 10.5. $e_{\boldsymbol{a}}^2 = e_{\boldsymbol{a}}$ *(this is what "idempotent" means).*

Proof. $1^2 = 1$ and $e_{\boldsymbol{a}} e_{\boldsymbol{b}} = 0$ for $\boldsymbol{a} \neq \boldsymbol{b}$, so

$$\sum_{\boldsymbol{a} \in \mathbb{V}(I)} e_{\boldsymbol{a}} = 1 = 1^2 = \left(\sum_{\boldsymbol{a}} e_{\boldsymbol{a}} \right)^2 = \sum_{\boldsymbol{a} \in \mathbb{V}(I)} e_{\boldsymbol{a}}^2,$$

therefore $e_{\boldsymbol{a}} = e_{\boldsymbol{a}}^2$ by comparing terms. □

Lemma 10.6. *The idempotent $e_{\boldsymbol{a}}$ is an identity element of the subring $S_{\boldsymbol{a}}$.*

Proof. If $g \in S_{\boldsymbol{a}}$ then $g = 1g = \sum e_{\boldsymbol{b}} g = e_{\boldsymbol{a}} g$ since g is orthogonal to all the other idempotent elements $e_{\boldsymbol{b}}$ for $\boldsymbol{b} \neq \boldsymbol{a}$. □

The idempotent $e_{\boldsymbol{a}}$ evaluates to 1 at $\boldsymbol{a}$.

Lemma 10.7. *The ring $S_{\boldsymbol{a}}$ consists of functions which vanish at every $\boldsymbol{b} \in \mathbb{V}(I) \setminus \{\boldsymbol{a}\}$.*

Proof. If $g \in S_{\boldsymbol{a}}$ then $(x_i - a_i)^{\ell_i} g = 0 \mod I$ for some ℓ_i. Evaluating this at $\boldsymbol{b}$ gives

$$(b_i - a_i)^{\ell_i} g(\boldsymbol{b}) = 0.$$

If $\boldsymbol{b} \neq \boldsymbol{a}$ then at least one of the terms $b_i - a_i$ is nonvanishing in the field k, so we can divide by it to obtain $g(\boldsymbol{b}) = 0$. □

Lemma 10.8. *The ring $S_{\boldsymbol{a}}$ has a unique maximal ideal $\mathfrak{n}$, namely the ideal of polynomials which vanish at $\boldsymbol{a}$. All elements of this maximal ideal are nilpotent, i.e. $g \in \mathfrak{n}$ implies $g^\ell = 0$ for some ℓ.*

Proof. Consider the evaluation homomorphism $\mathrm{ev}_{\boldsymbol{a}} \colon S_{\boldsymbol{a}} \to k$, $\mathrm{ev}_{\boldsymbol{a}}(g) = g(\boldsymbol{a})$. The kernel of this homomorphism is precisely the ideal $\mathfrak{n}$ of polynomials in $S_{\boldsymbol{a}}$ which vanish at $\boldsymbol{a}$. Applying the first isomorphism theorem to $\mathrm{ev}_{\boldsymbol{a}}$, we see that $S_{\boldsymbol{a}}/\mathfrak{n} \cong k$, therefore $\mathfrak{n}$ is a maximal ideal by Lemma 6.4.

If $g \in \mathfrak{n}$ then g vanishes at $\boldsymbol{a}$ and, by Lemma 10.7, at all other points of $\mathbb{V}(I)$. By the Nullstellensatz, $g^\ell = 0$ for some ℓ, proving that elements of $\mathfrak{n}$ are nilpotent.

If $g \in S_{\boldsymbol{a}} \setminus \mathfrak{n}$ then $g(\boldsymbol{a}) \neq 0$. The polynomial $N := e_{\boldsymbol{a}} - \frac{g}{g(\boldsymbol{a})}$ does vanish at $\boldsymbol{a}$ and hence belongs to $\mathfrak{n}$, hence is nilpotent. We have $g/g(\boldsymbol{a}) = e_{\boldsymbol{a}} - N$, which is invertible with inverse $e_{\boldsymbol{a}} + N + N^2 + \cdots$ where the sum is finite because N is nilpotent. This implies that g itself is invertible. Therefore any ideal which contains an element of $S_{\boldsymbol{a}} \setminus \mathfrak{n}$ is necessarily the whole of $S_{\boldsymbol{a}}$. This implies that $\mathfrak{n}$ is the only maximal ideal. □

10.2 Bounds on intersection multiplicity

Suppose that p is an intersection point of two curves $C = \{f = 0\}$ and $C' = \{g = 0\}$ for some polynomials $f, g \in R = k[x, y]$. To get a better understanding of $i_p(C, C')$, we use a clever trick. Let $\mathfrak{m} \subset R$ be the ideal of polynomials vanishing at p and fix a number $d \geq 1$. Let $I = (f, g)$, $J = I + \mathfrak{m}^d$, $S = k[x, y]/(f, g)$ and $T = k[x, y]/J$. Imposing the extra equations $\mathfrak{m}^d$ just leaves us with the point p, that is $\mathbb{V}(J) = \mathrm{Spec}(T) = \{p\}$. This means $T = T_p$, i.e. T is a local ring; let $\mathfrak{n} \subset T$ be its unique maximal ideal. Recall that $\mathfrak{n}$ consists of polynomials which vanish at p.

Lemma 10.9. *T is a quotient of S_p. In particular, $\dim(T)$ is a lower bound for $i_p(C, C')$.*

Proof. Since $I \subset J$, we have a surjective quotient map $\varphi \colon S \to T$. If $q \in \mathrm{Spec}(S)$ is a point different from p, we will show that $\varphi(S_q) = 0$. This implies the lemma because then $T = \varphi(S_p) = S_p/\ker(\varphi)$.

To prove the claim, suppose that $g \in S_q$ for some $q \neq p$. Then $\varphi(g) \in \mathfrak{n}$ because $g(p) = 0$. In particular, $\varphi(e_q) \in \mathfrak{n}$ and hence $\varphi(e_q)$ is nilpotent, say $\varphi(e_q)^N = 0$. Since e_q is an identity for S_q, this means that for any $g \in S_q$, we have

$$\varphi(g) = \varphi(e_q^N g) = \varphi(e_q)^N \varphi(g) = 0,$$

as required. □

We start with a simple application which we will generalise.

Proposition 10.10. *If $i_p(C, C') = 1$ then p is a smooth point of C and of C', and C and C' intersect transversely at p, i.e. the tangents T_pC and T_pC' are distinct.*

Proof. We will assume $p = (0, 0)$ for simplicity, so $\mathfrak{m} = (x, y)$.

Take $d = 2$ and let $T = k[x, y]/(\mathfrak{m}^2, f, g)$. There is a (surjective) quotient map

$$\psi\colon k[x, y]/\mathfrak{m}^2 \to T,$$

so $\dim(T) = \operatorname{rank}(\psi) = \dim(k[x, y]/\mathfrak{m}^2) - \operatorname{null}(\psi)$, by the rank-nullity theorem. The space $k[x, y]/\mathfrak{m}^2$ consists of polynomials $a + bx + cy$ modulo terms of higher order, so it has dimension 3 over k. A polynomial $a + bx + cy \mod \mathfrak{m}^2$ is in the kernel of ψ if and only if it is the lowest-order part of an element of (f, g).

Let f_1 and g_1 be the first order parts of the Taylor expansions of f and g at 0. If p fails to be a smooth point on C (respectively C') then $f_1 = 0$ (respectively $g_1 = 0$). If p is a smooth point on both but $T_pC = T_pC'$ then f_1 and g_1 are nonzero but linearly dependent. In any of these cases, the subspace of $k[x, y]/\mathfrak{m}^2$ consisting of lowest order parts of polynomials in (f, g) is at most 1-dimensional. Thus in any of these cases, $\operatorname{null}(\psi) \leq 1$ and $\dim(T) \geq 3 - 1 = 2$. The proposition now follows from Lemma 10.9. □

Proposition 10.11. *Let $C = \{f = 0\}$ and $C' = \{g = 0\}$ be curves and $p \in C \cap C'$. Let $c = m_p(C)$ and $c' = m_p(C')$. Then $i_p(C, C') \geq cc'$.*

Proof. Take $d = c + c'$ and let $T = k[x, y]/(\mathfrak{m}^d, f, g)$. By Lemma 10.9, it suffices to show

$$\dim_k(T) \geq cc'.$$

Consider the reduction map modulo (f, g):

$$k[x, y]/\mathfrak{m}^{c+c'} \xrightarrow{\varphi} k[x, y]/(\mathfrak{m}^{c+c'}, f, g).$$

This is surjective, so $k[x, y]/(\mathfrak{m}^{c+c'}, f, g) \cong \left(k[x, y]/\mathfrak{m}^{c+c'}\right) / \ker\varphi$.

If $h \in \ker\varphi$ then $h = \alpha f + \beta g + \mathfrak{m}^{c+c'}$. Since $f \in \mathfrak{m}^c$ and $g \in \mathfrak{m}^{c'}$, we have $h = \bar{\alpha} f + \bar{\beta} g + \mathfrak{m}^{c+c'}$ where $\bar{\alpha} = \alpha \mod \mathfrak{m}^{c'}$ and $\bar{\beta} = \beta \mod \mathfrak{m}^c$. The space of possible $\bar{\alpha}$s (respectively $\bar{\beta}$s) is $k[x, y]/\mathfrak{m}^{c'}$ (respectively $k[x, y]/\mathfrak{m}^c$). This means that

$$\dim\left(\left(k[x, y]/\mathfrak{m}^{c+c'}\right) / \ker\varphi\right) \geq \dim k[x, y]/\mathfrak{m}^{c+c'} - \dim k[x, y]/\mathfrak{m}^c - \dim k[x, y]/\mathfrak{m}^{c'}.$$

Since $\dim k[x, y]/\mathfrak{m}^\ell = \ell(\ell + 1)/2$, the lower bound is

$$\frac{(c + c')(c + c' + 1)}{2} - \frac{c(c + 1)}{2} - \frac{c'(c' + 1)}{2} = cc',$$

so $i_p(C, C') \geq \dim k[x, y]/(\mathfrak{m}^{c+c'}, f, g) = \dim\left(\left(k[x, y]/\mathfrak{m}^{c+c'}\right) / \ker\varphi\right) \geq cc'$. □

11 Bézout's theorem

11.1 Statement and proof

Theorem 11.1. *If $C = \{f = 0\}$ and $C' = \{g = 0\}$ are affine curves with no common component (i.e. $\gcd(f, g) = 1$) then the number of intersection points $C \cap C'$ is less than mn where $m = \deg(f)$ and $n = \deg(g)$. In fact,*

$$\sum_{p \in C \cap C'} i_p(C, C') \leq mn.$$

Proof. Recall that $i_p(C, C') = \dim_k S_p$ where $S = k[x, y]/(f, g)$ and S_p is the local ring at p, that is the generalised simultaneous p-eigenspace of S (i.e. the biggest subspaces of S on which $\hat{x} - x(p)$ and $\hat{y} - y(p)$ are both nilpotent). By Theorem 9.14, $S = \bigoplus_{p \in C \cap C'} S_p$, so

$$\dim_k(S) = \sum_{p \in C \cap C'} i_p(C, C'),$$

and we just need to show that $\dim_k(S) \leq mn$.

Since $C \cap C'$ is finite, $\dim_k(R)$ is finite, and hence $S = k[x, y]/(f, g)$ is generated by a finite set of polynomials. Let's pick a generating set, and choose a number D bigger than the degrees of all these generators.

Write $P_{\leq D}$ for the space of all polynomials in $k[x, y]$ having degree $\leq D$; this is a k-vector space of dimension $(D + 1)(D + 2)/2$. Let

$$\psi \colon P_{\leq D} \to S = k[x, y]/(f, g)$$

be the restriction of the quotient map $k[x, y] \to S$ to $P_{\leq D}$. This is surjective by our choice of D, so $\dim(S) = \operatorname{rank}(\psi)$. By the rank-nullity theorem,

$$\operatorname{rank}(\psi) = \dim P_{\leq D} - \operatorname{null}(\psi),$$

so finding an upper bound on the rank means finding...

A lower bound on $\operatorname{null}(\psi)$**:** Any polynomial of the form $\alpha f + \beta g$ with $\alpha \in P_{\leq D-m}$ and $\beta \in P_{\leq D-n}$ necessarily lives in $\ker(\psi)$ (it lives in $P_{\leq D}$ and also in (f, g)). So $\operatorname{im}(\phi) \subset \ker(\psi)$ where

$$\phi \colon P_{\leq D-m} \times P_{\leq D-n} \to P_{\leq D}$$

is the map $\phi(\alpha, \beta) = \alpha f + \beta g$. So $\operatorname{null}(\psi) \geq \operatorname{rank}(\phi)$. But

$$\operatorname{rank}(\phi) = \dim P_{\leq D-m} + \dim P_{\leq D-n} - \operatorname{null}(\phi).$$

Putting this together, we have

$$\dim(S) \leq \dim P_{\leq D} - \dim P_{\leq D-m} - \dim P_{\leq D-n} + \operatorname{null}(\phi).$$

So we need to find...

An upper bound for $\operatorname{null}(\phi)$**:** If $(\alpha, \beta) \in \ker(\phi)$ then $\alpha f + \beta g = 0$. Since $\gcd(f, g) = 1$,

this implies $\alpha = qg$ and $\beta = -qf$ for some q. Since $\deg(\alpha) \leq D - m$, $\deg(g) = n$, we need $\deg(q) \leq D - m - n$. Therefore[17] $\ker(\phi) \subset \mathrm{im}(\theta)$, where

$$\theta \colon P_{\leq D-m-n} \to P_{D-m} \times P_{D-n}$$

is given by $\theta(q) = (qg, -qf)$. This gives[18]

$$\mathrm{null}(\phi) \leq \dim P_{\leq D-m-n}.$$

Putting it all together, and using the formula $\dim P_{\leq d} = (d-1)(d-2)/2$, we get

$$\dim(S) \leq \dim P_{\leq D} - \dim P_{\leq D-m} - \dim P_{\leq D-n} + \dim P_{\leq D-m-n} = mn,$$

as required. □

Remark 11.2. For the connaisseur, this argument has the flavour of cohomology: we are studying the sequence

$$0 \to P_{\leq D-m-n} \xrightarrow{\theta} P_{\leq D-m} \times P_{\leq D-n} \xrightarrow{\phi} P_{\leq D} \xrightarrow{\psi} S \to 0$$

of maps. We have shown that this is a *chain complex*, which means that the image of each map is in the kernel of the next):

$$\mathrm{im}(\theta) \subset \ker(\phi), \qquad \mathrm{im}(\phi) \subset \ker(\psi), \qquad R = \mathrm{im}(\psi),$$

and we have calculated some (but not all) of the *cohomology groups*:

$$\ker(\theta) = 0, \qquad \ker(\phi)/\mathrm{im}(\theta) = 0, \qquad \ker(\psi)/\mathrm{im}(\phi) = \mathbf{?}, \qquad S/\mathrm{im}(\psi) = 0.$$

Now take the Euler characteristic (i.e. the alternating sum of the dimensions of the vector spaces in the sequence):

$$-\dim(S) + \dim P_{\leq D} - (\dim P_{\leq D-m} + \dim P_{\leq D-n}) + \dim P_{\leq D-m-n} = mn - \dim(S).$$

This should be the same as the alternating sum of the dimensions of the cohomology groups:

$$0 - 0 + \mathbf{?} - 0 \geq 0$$

which gives

$$mn - \dim(S) \geq 0.$$

More sophisticated versions of Bézout's theorem can be understood in terms of homological intersection numbers.

Remark 11.3. This also lets us see precisely what the obstruction is to equality in Bézout's theorem: we get $\dim(S) = mn$ if and only if $\ker(\psi) = \mathrm{im}(\phi)$.

[17] In fact, $\phi(\theta(q)) = qgf - qfg = 0$, so $\mathrm{im}(\theta) \subset \ker(\phi)$, so we really have $\ker(\phi) = \mathrm{im}(\theta)$.

[18] In fact, since θ is injective, we get $\ker(\phi) = \mathrm{im}(\theta) \cong P_{\leq D-m-n}$, so $\mathrm{null}(\phi) = \dim P_{\leq D-m-n}$.

11.2 Examples

There are two ways that Bézout's theorem can become a strict inequality:

- if the field k is not algebraically closed then some of the intersection points might only exist over an extension field.
- some of the intersection points could be "hiding at infinity".

Here are some examples to illustrate this.

Example 11.4. The circle $C = \{x^2 + y^2 = 1\}$ and the hyperbola $C' = \{xy = 2\}$ do not intersect in $\mathbb{R}^2$ (Bézout's theorem gives a bound of $|C \cap C'| \leq 4$). Of course they do intersect in $\mathbb{C}^2$: substituting $y = 2/x$ in the equation for the circle gives

$$x^2 + 4/x^2 = 1,$$

or $(x^2)^2 - x^2 + 4 = 0$, which means $x^2 = \frac{1 \pm i\sqrt{3}}{2}$ giving four points of intersection

$$\left\{ (x, 2/x) \,:\, x = \pm \frac{1 \pm i\sqrt{3}}{2} \right\}.$$

Example 11.5. Take $C = \{x = 0\}$ and $C' = \{x = 1\}$. These vertical lines are parallel and do not intersect. There is, in some sense, an intersection hiding at infinity, which reveals itself at a finite height if you tilt one of the lines slightly.

Example 11.6. Consider the parabolas $\{y = x^2\}$ and $\{y = 2x^2\}$. These curves have degree 2, so Bézout gives an upper bound on 4 for their intersection. They intersect with multiplicity 2 at the origin. There is an additional multiplicity 2 intersection point hiding at infinity.

We can compute the intersection multiplicity at the origin as follows. Let $I = (y - x^2, y - 2x^2)$ and $S = k[x, y]/I$. This ring is 2-dimensional over k: the monomials 1 and x form a basis (for example, you can see that $x^2 = (y - x^2) - (y - 2x^2) \in I$ and $y = 2(y - x^2) - (y - 2x^2) \in I$). Since the origin is the only intersection point, $\mathrm{Spec}(S) = \{(0, 0)\}$ and $S = S_0$. Thus

$$i_0(C, C') = \dim S_0 = 2.$$

11.3 Applications of Bézout, I: Singularity bounds

We already saw how Bézout's theorem can be applied in Section 8.2. Here are some more applications with a similar flavour.

Theorem 11.7. *Let C be an irreducible curve of degree $d \geq 1$. A point $p \in C$ can have multiplicity at most $d - 1$, and if p has multiplicity $d - 1$ then p is the only singular point of C.*

Proof. Given another point $q \in C$, pick the unique line L connecting p and q. Since C is irreducible of degree at least 1 and L is of degree 1, the curves C and L share no common component, so Bézout's theorem applies to their intersection. By Bézout's theorem, we

have

$$m_p(C) + m_q(C) \leq \deg(L)\deg(C) = d.$$

Since $m_q(C) \geq 1$ we have $m_p(C) \leq d-1$. If q is singular then $m_q(C) \geq 2$ so $m_p(C) \leq d-2$. □

Example 11.8. An irreducible cubic curve can have either no singularities or one double point.

There are innumerable further results one could prove. For example, what is the maximum number of double points on an irreducible quintic curve? Rather than dwelling on this here, I leave it as a project for you to explore using the techniques we have now developed. Instead, we will discuss what happens when we drop the irreducibility assumption.

Remark 11.9. If we allow C to be reducible then the theorem breaks down. For example:

- the cubic $\{xy(x+y) = 0\}$ has a triple point at the origin
- the cubic $\{xy(x+y-1) = 0\}$ has three nodes.

In each of these cases, if you try and run the proof of Theorem 12.7, you find that the line L is contained as an irreducible component in C, so Bézout's theorem cannot be applied (the cubic and the equation of the line have a common factor).

We can constrain the singularities of reducible curves using Bézout's theorem by a similar method: intersecting C with an auxiliary curve C' so that all the intersections lie in $C \cap C'$. But we have to take more care to ensure that C' and C share no common irreducible components. Here is an example of such an argument.

Theorem 11.10. *Suppose k is an infinite field. A curve $C \subset \mathbb{A}^2(k)$ of degree $d > 0$ with no multiple components*[19] *can have at most $d(d-1)/2$ singular points.*

Proof. Suppose that $C = \{f = 0\}$. Recall that a singularity of C is a point where all of the following vanish:

$$f, \qquad \partial f/\partial x, \qquad \partial f/\partial y.$$

Given two numbers α, β, consider the differential operator

$$D := \alpha\frac{\partial}{\partial x} + \beta\frac{\partial}{\partial y}.$$

We will specify how to choose α and β later. Note that if p is singular then $Df(P) = 0$. In particular,

$$\mathrm{Sing}(C) \subset \{f = 0\} \cap \{Df = 0\}.$$

The degree of f is d and the degree of Df is at most $d-1$. First of all, we will pick α and β so that Df is not identically zero, so that $\{Df = 0\}$ is a curve. Providing $\{Df = 0\}$ and $\{f = 0\}$ have no common components, Bézout's theorem implies that the number of singularities is at most $d(d-1)$. The remainder of the proof is therefore dedicated to choosing α and β to ensure that $\{f = 0\}$ and $\{Df = 0\}$ have no common components.

[19] i.e. $C = \{f = 0\}$ where f has no repeated irreducible factors. If f does have repeated factors you can always just throw them away to get a polynomial of lower degree without changing C.

Suppose that f factors into distinct irreducible polynomials g_i of degree $d_i > 0$:

$$f = g_1 g_2 \cdots g_K.$$

We need pick α and β that none of the irreducible factors of f divide Df. By Leibniz's rule:

$$Df = (Dg_1)g_2 \cdots g_K \; + \; g_1(Dg_2)g_3 \cdots g_K \; + \cdots + \; g_1 \cdots g_{K-1}(Dg_K).$$

Suppose without generality that g_i divides Df. Then

$$(Dg_i) \prod_{j \neq i} g_j = g_i \left(\frac{Df}{g_i} - \sum_{j \neq i} (Dg_j) \prod_{k \neq i,j} g_k \right)$$

which means that g_i divides $(Dg_i) \prod_{j \neq i} g_j$. Since the irreducible factors are distinct and prime, this means that g_i divides Dg_i. If Dg_i is not identically zero then this is not possible, since it is a polynomial of lower degree than g_i. So we need to choose α and β to ensure Dg_i is not identically zero for all the irreducible factors g_i of f.

For each factor g_i, since it is a nonconstant polynomial, there is a point p_i where at least one of its partial derivatives is nonzero. Therefore

$$L_i := \left\{ (\alpha, \beta) \; : \; \alpha \frac{\partial g_i}{\partial x}(p_i) + \beta \frac{\partial g_i}{\partial y}(p_i) = 0 \right\}$$

is a line in the space of possible (α, β). We just need to pick a point which does not lie on any of these lines. As long as k is infinite, there are infinitely many points which do not lie on a finite union of lines. □

Remark 11.11. We really do need to be careful; for example if we carelessly take $f(x, y) = xy$ and $D = \partial/\partial x$ then $Df(x, y) = y$ which shares the component $\{y = 0\}$ with $\{f = 0\}$. Similarly if we try $D = \partial/\partial y$. But there are choices of D which work, for example $D = \partial/\partial x + \partial/\partial y$ gives $Df = x + y$ which has no common factor with $f = xy$.

A more careful analysis reveals:

Theorem 11.12. *If C is a curve of degree d having singularities $p_1, \ldots, p_K$ with multiplicities $m_1, \ldots, m_K$ respectively then*

$$\sum_{i=1}^{K} m_i(m_i - 1) \leq d(d - 1).$$

Proof. Recall that p_i has multiplicity m_i if and only if all derivatives of f up to order $m_i - 1$ vanish at p_i. This implies that the derivatives of Df at p_i vanish up to order $m_i - 2$ (as Df is a linear combination of the first derivatives of f). In particular, $\{Df = 0\}$ has p_i as a singularity of multiplicity $m_i - 1$. Therefore

$$i_{p_i}(\{f = 0\}, \{Df = 0\}) \geq m_i(m_i - 1).$$

By Bézout's theorem, we get

$$\sum_{i=1}^{K} m_i(m_i - 1) \leq d(d - 1)$$

as before. □

11.4 Applications of Bézout, II: Harnack's theorem

In this section, we focus on real plane curves in $\mathbb{R}^2$, and to avoid some annoying boundary cases we assume they have no isolated points (so nothing like $\{x^2+y^2=0\}$). By an *oval*, we mean a closed loop in the curve. The Jordan curve theorem (which we are assuming here) asserts that each loop divides the plane into a bounded region ("inside") and an unbounded region ("outside"). We will not prove this, hoping that it is intuitive enough for you to take on faith, and will focus instead on our main goal:

Theorem 11.13 (Harnack's theorem)**.** *Let C be an irreducible plane curve of degree d and let $M = 1+\frac{1}{2}(d-1)(d-2)$. If C has M ovals then it cannot have any other connected components (oval or not).*

Remark 11.14. Note that curves of odd degree necessarily have at least one asymptote; we will prove this when we discuss projective geometry later. This means an odd-degree curve can have at most $M-1$ ovals.

We will prove the theorem for cubics and quartics, then leave the case of general d as an exercise. We start with a lemma.

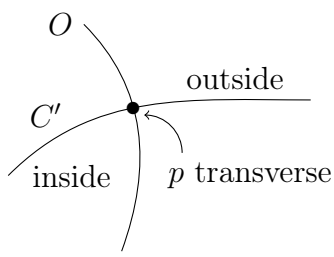
Lemma 11.15. *Let C and C' be curves and suppose that O is an oval of C. If C' intersects O at a point p then at least one of the following occurs:*

(a) C' intersects O at another point q,

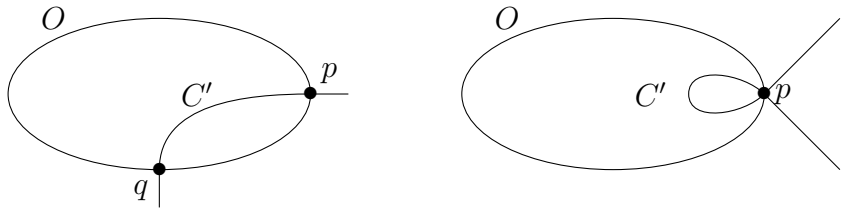
(b) C' intersects O with multiplicity ≥ 2 at p.

Proof. There are finitely many intersections between C' and O (at most $\deg(C)\deg(C')$ by Bézout's theorem) so $C' \setminus (C' \cap O)$ consists of finitely many segments (and possibly ovals or isolated points, but we will ignore these). We will call these segments "inside" or "outside" according to whether they are contained in the bounded or unbounded region of $\mathbb{R}^2 \setminus O$.

If $i_p(C, C') \geq 2$ then we are in case (b) and there is nothing to prove, so assume that $i_p(C, C') = 1$. In this case, the intersection at p is transverse by Proposition 10.10, so that some of the points of C' lie inside O and some lie outside O:

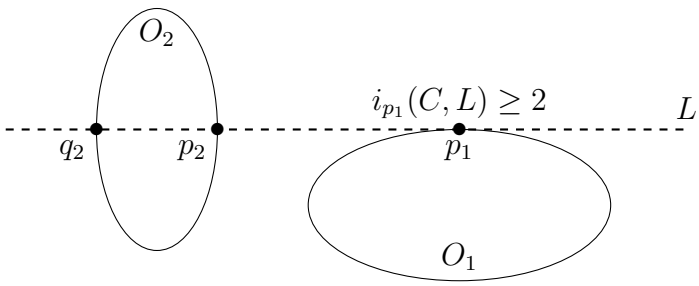


But each "inside" segment has another endpoint, q. If $p \neq q$ then we are in case (a). If $p = q$ then the multiplicity $m_p(C)$ is at least 2, so $i_p(C, C') \geq 2$ and we are in case (b).



□

Proof of Harnack for cubics. In this case, the theorem asserts there should be at most $1 + (3-1)(3-2)/2 = 2$ ovals. In fact, we can do better and prove there is at most one oval. Suppose there are two ovals O_1 and O_2, and let $p_1 \in O_1$ and $p_2 \in O_2$ be points on these ovals. Let L be the straight line joining p_1 to p_2. By Lemma 12.15, either $i_{p_k}(L, C) \geq 2$ or there is an additional point $q_k \in L \cap O_k$. This gives us a lower bound of 4 on the intersection number between L and C, but Bézout's theorem tells us that this intersection number is at most $\deg(L)\deg(C) = 3$, giving a contradiction. (We can apply Bézout because C is irreducible, so does not have L as a component.)

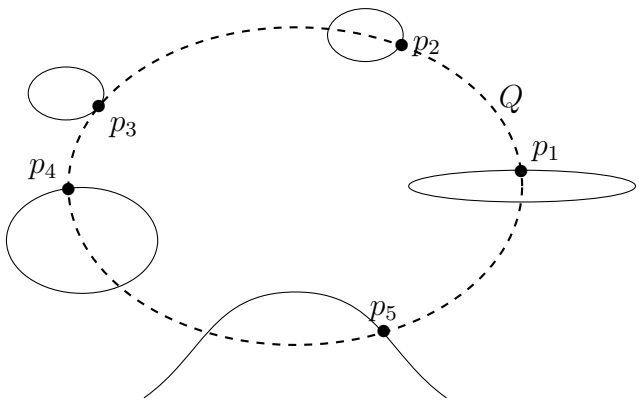


□

Proof of Harnack for quartics. In this case, we are trying to prove that if there are four ovals there can be no other component. Suppose for contradiction that we have a curve with four ovals O_1, O_2, O_3, O_4 and one more component Γ. Pick four points p_1, p_2, p_3, p_4 with $p_k \in O_k$, and a fifth point $p_5 \in \Gamma$. There is a conic curve, Q, passing through these five points (we will see why next lecture). By Lemma 12.15, for each $k = 1, 2, 3, 4$ we either have $i_{p_k}(Q, O_k) \geq 2$ or else there is another point $q_k \in Q \cap O_k$. Finally, we have $i_{p_5}(Q, C) \geq 1$. This gives

$$9 \leq i_{p_5}(Q, C) + \sum_{k=1}^{4} \sum_{p \in Q \cap O_k} i_p(Q, C),$$

while Bézout's theorem tells us that this quantity is bounded by $2 \times 4 = 8$, giving a contradiction.



□

The strategy of argument is similar for higher degree, using Theorem 11.2 to produce a suitable auxiliary curve (like the line or conic in these proofs) passing through M points on different ovals and a bunch of other points on a putative $M + 1$st component. The existence of suitable auxiliary curves is guaranteed by what we do in the next lecture.

Exercise 11.16. Write out the proof for general d.

12 Existence of curves with constraints

12.1 Finding curves

Theorem 12.1. *Through any two points there is a straight line (i.e. an algebraic curve of degree 1).*

We will prove something stronger.

Theorem 12.2. *Given $d(d+3)/2$ points in $\mathbb{A}^2(k)$, there is a curve of degree d passing through them.*

The proof will actually give an algorithm for finding the curve. Let's work out an example before we give the proof.

Example 12.3. Find a conic curve passing through the points

$$(-1,0), \quad (0,-1), \quad (1,0), \quad (0,1), \quad \left(\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}\right).$$

A general conic is given by the equation

$$a_{20}x^2 + a_{02}y^2 + a_{110}xy + a_{10}x + a_{01}y + a_{00} = 0.$$

The condition that this conic passes through $(-1,0)$ is obtained by substituting $x = -1, y = 0$ into this equation, which gives:

$$a_{20} - a_{10} + a_{00} = 0.$$

The other point constraints tell us that:

$$\begin{aligned}
0 &= a_{02} - a_{01} + a_{00} \\
0 &= a_{20} + a_{10} + a_{00} \\
0 &= a_{02} + a_{01} + a_{00} \\
0 &= \frac{1}{2}(a_{20} + a_{02} + a_{11}) + \frac{1}{\sqrt{2}}(a_{10} + a_{01}) + a_{00}
\end{aligned}$$

This is a system of five simultaneous linear equations for our six coefficients, and it has the general solution

$$a_{11} = a_{10} = a_{01} = 0, \qquad a_{20} = a_{02} = -a_{00}.$$

If we write $\alpha = a_{00}$ then this means our conic has the equation

$$\alpha - \alpha(x^2 + y^2) = 0.$$

In fact, the overall factor of α does not affect the conic, so we can divide out and rearrange to get

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

Proof of Theorem 11.2. The strategy of proof will be just like in the example: a curve of degree d has the equation

$$\sum_{i+j \le d} a_{ij} x^i y^j = 0.$$

Impose the point constraints to get a system of simultaneous linear equations for the "variables" a_{ij}. If we have N variables and we impose fewer than $N-1$ constraints, we necessarily have free variables in the solution (like α in the example) and by picking a value for these free variables (like $\alpha = 1$ in the example) we get a nonzero choice of coefficients a_{ij} yielding a curve passing through our chosen points.

All that remains is to see where the number $d(d+3)/2$ is coming from. The lemma below shows is that it is precisely one less than the number of coefficients a_{ij} with $i+j \leq d$, so the lemma will complete the proof of Theorem 11.2. □

Lemma 12.4. *The total number of monomials in three variables of degree at most d is $N = (d+1)(d+2)/2$, and $N-1 = d(d+3)/2$.*

Proof. Recall that a monomial of degree d is an expression like $x^i y^j$ with $i+j=d$. For example, x^2y is a monomial of degree 3. We can write the monomials of degree $\leq d$ in a triangle. On the bottom row we write the monomials of degree d:

$$x^d, \quad x^{d-1}y, \quad x^{d-2}y^2, \quad \cdots \quad y^d$$

There are $d+1$ of these. On the next row up, we write all the monomials of degree $d-1$:

$$x^{d-1}, \quad x^{d-2}y, \quad x^{d-3}y^2, \quad \cdots \quad y^{d-1}.$$

There are d of these. We continue, writing the monomials with degree $d-k$ on the kth row above the bottom, and we form a triangle with base $d+1$, and the monomial 1 at the apex.

$$\begin{array}{ccccccccc}
 & & & & 1 & & & & \\
 & & & \iddots & & \ddots & & & \\
 & & x^{d-2} & & \cdots & & y^{d-2} & & \\
 & x^{d-1} & & x^{d-2}y & & \cdots & & y^{d-1} & \\
x^d & & x^{d-1}y & & x^{d-2}y^2 & & \cdots & & y^d
\end{array}$$

The total number of monomials is therefore the $(d+1)$st triangular number,

$$N = (d+1)(d+2)/2.$$

Finally,

$$\begin{aligned}
N-1 &= \frac{(d+1)(d+2)}{2} - 1 \\
&= \frac{d^2+3d+2-2}{2} \\
&= \frac{d(d+3)}{2},
\end{aligned}$$

as required. □

Example 12.5. Let's try a more complicated problem: finding a cubic curve with sin-

gularities at $(0,0)$, $(1,0)$ and $(0,1)$. The general cubic is

$$\begin{aligned} f(x,y) = a_{30}x^3 + a_{21}x^2y + a_{12}xy^2 + a_{03}y^3 \\ + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \\ + a_{10}x + a_{01}y \\ + a_{00}. \end{aligned}$$

with ten coefficients. The condition that the cubic has a singularity at (a,b) is equivalent to the vanishing of

$$f(a,b), \qquad \frac{\partial f}{\partial x}(a,b), \qquad \frac{\partial f}{\partial y}(a,b).$$

Imposing this at the three points gives us nine equations in ten variables, so we will find a solution. It is an exercise for you to find these equations and this solution!

Remark 12.6. Forcing f to vanish at a point imposes one constraint. Forcing f and its first derivatives to vanish at a point imposes three constraints. More generally, by forcing higher and higher terms in the Taylor series at a point to vanish we can impose the constraint that our curve has a singularity with at least a specified multiplicity at that point. And we can do this at several points simultaneously. Provided we don't ask for the total multiplicity be too high, we will find a solution using the same kind of ideas as above.

Exercise 12.7. How many constraints are imposed by the vanishing of f and all its derivatives of order strictly smaller than k? What is the most general theorem along the lines of Theorem 11.2 (incorporating control over the multiplicities) that you can formulate?

13 Projective geometry, I

13.1 Conic sections revisited

When we classified curves of degree 2 on worksheet 1, we found that, up to a change of coordinates, any curve of degree 2 over $\mathbb{R}$ is one of the following:

A.1.a	$x^2 + y^2 = 1$	ellipse
A.1.b	$x^2 - y^2 = 1$	hyperbola
A.1.c	$x^2 + y^2 = -1$	empty
A.2	$y = x^2$	parabola
B.1.a	$x^2 + y^2 = 0$	single point($x = y = 0$)
B.1.b	$x^2 - y^2 = 0$	two transverse lines ($x = \pm y$)
B.2.a	$x^2 = 1$	two parallel lines ($x = \pm 1$)
B.2.b	$x^2 = -1$	empty
C	$x^2 = 0$	double line($x = 0$)

If we work over $\mathbb{C}$, some of these subcases become equivalent (i.e. related by a *complex* change of coordinates). For example, $x^2 + y^2 = 1$ and $x^2 - y^2 = 1$ are related by $(x, y) \mapsto (x, iy)$. The full list of cases is:

A.1	$x^2 + y^2 = 1$	
A.2	$y = x^2$	
B.1	$x^2 + y^2 = 0$	two transverse lines ($x = \pm iy$)
B.2	$x^2 = 1$	two parallel lines ($x = \pm 1$)
C	$x^2 = 0$	double line($x = 0$)

However, as the notation suggests, we can do even better if we allow even more general coordinate changes: the so-called *projective linear* (or Möbius) transformations.

Example 13.1. Suppose the coordinate systems (x, y) and (X, Y) are related by

$$x = \frac{X}{Y+1}, \qquad y = \frac{Y-1}{Y+1},$$

or equivalently

$$X = \frac{2x}{1-y}, \qquad Y = \frac{1+y}{1-y}.$$

The equation $4Y = X^2$ becomes

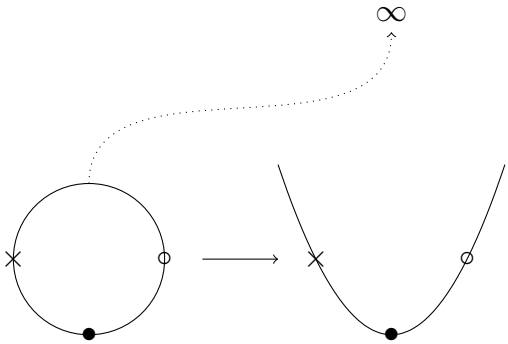
$$\frac{1+y}{(1-y)} = \frac{x^2}{(1-y)^2}$$

which rearranges into

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

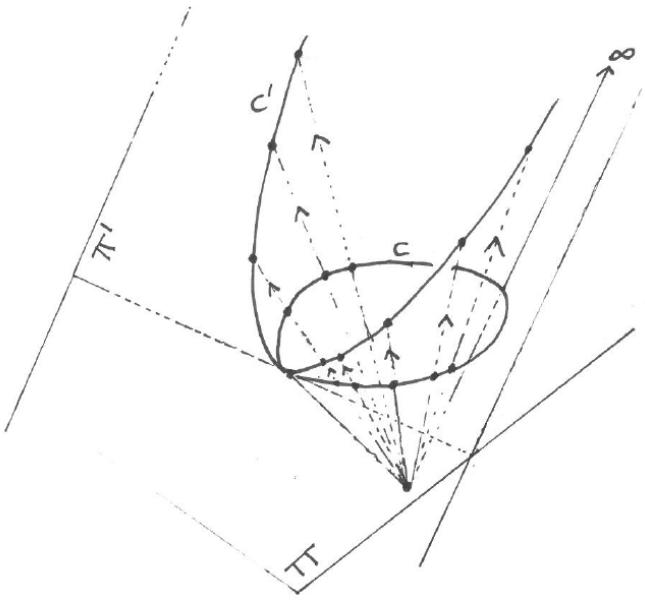
We have transformed a parabola into a circle!

You may object that this coordinate transformation makes no sense along the line $y = 1$ (or $Y = -1$) because we cannot divide by zero. Since the circle touches $y = 1$ at a single point, this point must go "to infinity[20]": how can we make rigorous sense of this?



Here is another viewpoint on the same transformation, which removes all mention of the word "infinity".

Example 13.2. Let $\Pi = \{z = 1\} \subset \mathbb{R}^3$ be the plane parallel to the xy-plane and sitting at height 1. Draw the circle $C = \{x^2+y^2 = 1\}$ in Π. Fix another plane Π', say $\{z = y+2\}$. Put a light source at the origin O, and look at the shadow[21] C' cast by C on Π'. For each point $P \neq O$ in $\mathbb{R}^3$, let L_P be the ray connecting P to O. Let $S \subset C$ be the subset of points $P \in C$ such that L_P is parallel to Π'. For each $P \in C$, the ray L_P intersects Π' if and only if $P \notin S$. For each point $P \in C$, the ray L_P is (at least partly) "in shadow" and if $P \in C \setminus S$ then the intersection point $L_P \cap \Pi'$ defines a point of C'. So following these rays gives a map $\varphi\colon C \setminus S \to C'$.



Let us work this out in detail for $\Pi' = \{z = y + 2\}$ and see what it has to do with our earlier coordinate transformation. We will pick coordinates X, Y on Π' so that $x = X$, $y = Y - 1$, $z = Y + 1$. The cone of light emanating from the origin and passing through C is a subvariety called $\mathrm{Cone}(C)$ defined by the equation $x^2 + y^2 = z^2$ (we will see how to figure out the equations of cones later). The intersection $C' = \mathrm{Cone}(C) \cap \Pi'$ is what I

[20] and no further.

[21] We will be more precise about "shadow" later.

mean by the shadow of C on Π'. In terms of X and Y, this is given by

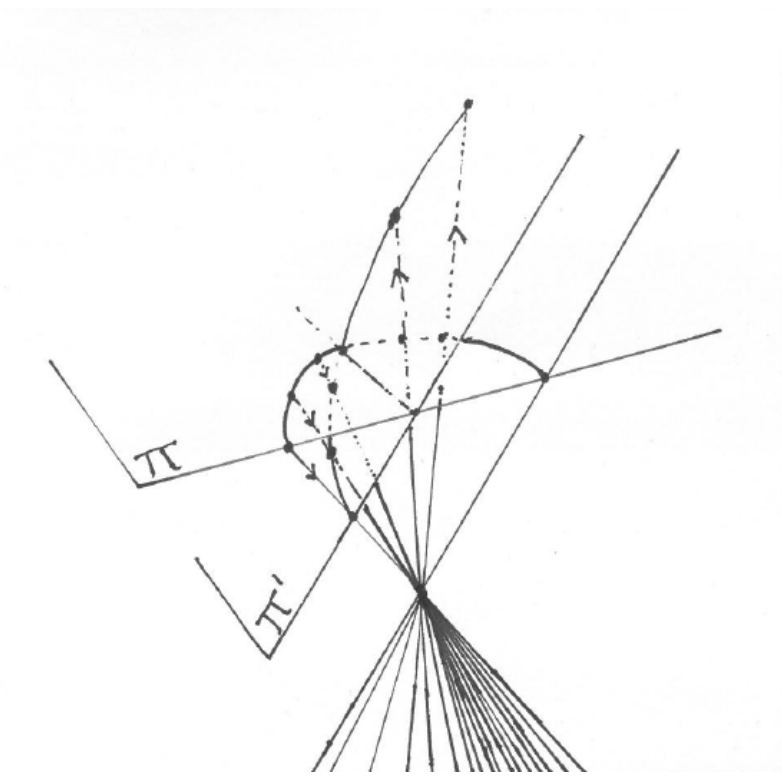
$$C' = \{(X, Y) \,:\, X^2 + (Y-1)^2 = (Y+1)^2\}.$$

This equation simplifies to $4Y = X^2$. Moreover, the map $\pi\colon C \setminus S \to C'$, $\varphi(P) = L_P \cap \Pi'$ turns out to be precisely our dodgy coordinate transformation from earlier. The inverse map $\varphi^{-1}\colon C' \to C$ is slightly easier to understand. A point $P = (X, Y-1, Y+1) \in C'$ lies on the ray $\{(\lambda X, \lambda(Y-1), \lambda(Y+1)) \,:\, \lambda \in \mathbb{R}\}$ through the origin. This intersects the plane Π precisely when $\lambda(Y+1) = 1$, i.e. $\lambda = 1/(Y+1)$, which gives the point

$$\varphi^{-1}(X, Y-1, Y+1) = \left(\frac{X}{Y+1}, \frac{Y-1}{Y+1}, 1\right) \in C.$$

This is precisely our earlier coordinate transformation. You can see that the map φ will fail to be defined at the point $P = (0, 1, 1) \in C$ because this is the unique place where the ray L_P is parallel to Π'.

Remark 13.3. The key thing in this example was taking the intersection of Π' with $\mathrm{Cone}(C)$. I used flowery language about shadows to introduce this idea, to put the picture of this intersection in your mind in a more visceral way, but don't take the metaphor too literally. For example, it is quite possible for the plane Π' to pass closer to the light source than some points of C; in that case, if we literally take the "shadow" then we may end up with a strict subset of C'. So instead of taking the shadow we should just take the intersection $\Pi' \cap \mathrm{Cone}(C)$. The picture below shows this for $\Pi' = \{z = y + 1\}$.



The idea of projective algebraic geometry is that we should treat the cone $\mathrm{Cone}(C)$ as a more fundamental entity than C: the affine curve C is just a planar slice of the cone. Of course, this is how the Greeks thought about conic sections (hence the name). The ellipse, hyperbola and parabola all arise as sections of the same cone.

13.2 Cones

Definition 13.4. A *ray* is a straight line through the origin. If $p = (x_0, \ldots, x_n) \in \mathbb{A}^{n+1}(k)$ is not the origin then there is a unique ray L_p containing p, namely

$$L_p = \{(\lambda x_0, \ldots, \lambda x_n) \,:\, \lambda \in k\}.$$

Definition 13.5. An algebraic set $V \subset \mathbb{A}^{n+1}(k)$ is called a *cone* if $(x_0, \ldots, x_n) \in V$ implies $(\lambda x_0, \ldots, \lambda x_n) \in V$ for all $\lambda \in k$. In other words, the whole ray through $x_0, \ldots, x_n)$ is contained in V. The rays contained in V are called the *rulings*[22] of V.

Definition 13.6. If V is a cone, the *projective variety* $\mathbb{P}(V)$ is the set of rulings of V.

This can be a little hard to imagine. The easiest way to get a handle on the set of rulings is to intersect with a plane.

Definition 13.7. An *affine slice* of a cone V is the intersection of V with a plane Π not containing the origin.

Each plane intersects each ruling at most once, so you get an accurate picture of a subset of the rulings. However, you miss out those rulings that are parallel to Π. These rulings are the "points at infinity" of the slice.

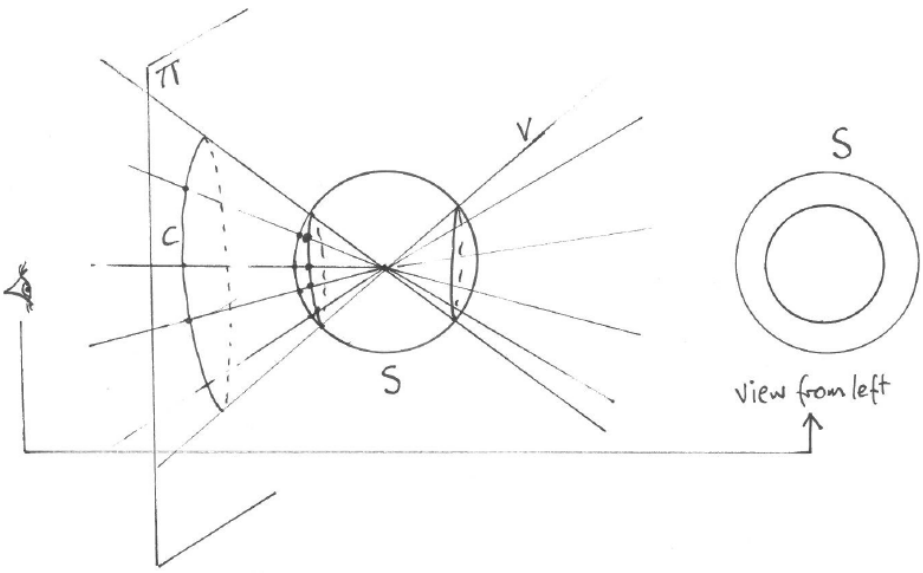
A more complete picture can be obtained by intersecting with a sphere rather than a plane. More precisely, let S be a sphere[23] centred at the origin and consider the intersection $S \cap V$. This is now a curve on the surface of the sphere. Since every ray intersects S at two (antipodal) points, the curve we see overcounts the points of the projective variety $\mathbb{P}(V)$. So just focus on a single hemisphere. Whilst following the curve, if you cross the equator and enter the other hemisphere, you should jump to the antipodal point on the equator and continue in your chosen hemisphere.

Example 13.8. Consider the cone $V = \{X^2 = Y^2 + Z^2\}$. Let $\Pi = \{X = 1\}$, let $\Pi' = \{Z = 1\}$, and let $S = \{X^2 + Y^2 + Z^2 = 1\}$.

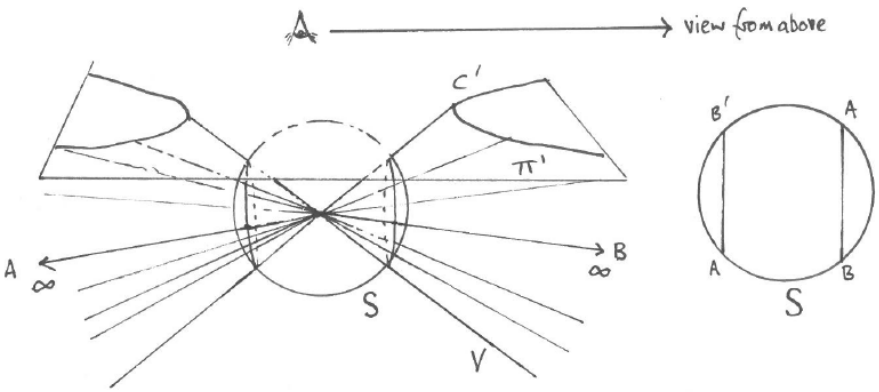
- The intersection $V \cap \Pi$ is the circle $\{1 = Y^2 + Z^2\} \subset \Pi$.
- The intersection $V \cap \Pi'$ is the hyperbola $\{X^2 = Y^2 + 1\} \subset \Pi'$.
- The intersection $C \cap S$ consists of two circles on the sphere S: the two equations $X^2 = 1 - Y^2 - Z^2$ and $X^2 = Y^2 + Z^2$ tell us that $Y^2 + Z^2 = 1/2$ and $X = \pm 1/\sqrt{2}$ on the intersection.

[22] Because you draw them on with a straight-line ruler.

[23] Recall that a sphere is the set of points at a fixed distance from a fixed centre. In $\mathbb{R}^3$, this forms a 2-dimensional surface, like the surface of the Earth. In other words, we are including only the crust, not the mantle and core.



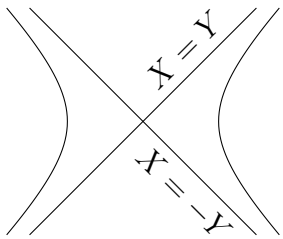
If we focus on the hemisphere $X \geq 0$ of S then we just see one circle. Note that projecting radially out gives a correspondence between this hemisphere and the plane Π which takes this circle to $V \cap \Pi$. We see that the projective variety $\mathbb{P}(V)$ is a single loop.



If we focus instead on the hemisphere $Z \geq 0$ of S then we see two half-circles: again, projecting outward, these correspond to the two pieces of the hyperbola $V \cap \Pi'$. How is that a single loop? The picture shows the hemisphere viewed from above. When we leave the hemisphere at the point A then we should re-emerge at the point A': these both lie on the same ray and hence represent the same point of $\mathbb{P}(V)$. Similarly when we leave at B we should re-emerge at B'.

The advantage of thinking in terms of spheres and hemispheres is that it brings the "points at infinity" in to where we can see them. In this example, A and B are points at infinity from the point of view of the affine slice $Z = 1$.

The points A and B of $\mathbb{P}(V)$ correspond to rulings of V which are parallel to Π'. In other words, they are contained in the parallel plane $\{Z = 0\}$. Setting $Z = 0$ in the equation $X^2 = Y^2 + Z^2$ for V we get $X^2 = Y^2$, i.e. $X = Y$ or $X = -Y$. These are precisely the asymptotes of the hyperbola $\{X^2 = Y^2 + 1\} = V \cap \Pi'$. More on this later.



13.3 Equations for cones

Since intersections of algebraic sets are algebraic, we see that the affine slice of a cone is an algebraic set. What are the equations for the slice? The equation of Π is (affine) linear, so can be used to express one of the coordinates linearly in terms of the others; we can substitute this into the equations of V to find the equations of the slice. For example, we often use $\Pi = \{z = 1\} \subset k^3$, so if $F(x, y, z) = 0$ is one of the equations defining V then $f(x, y) := F(x, y, 1)$ is one of the equations for the slice.

What about going back the other way? Let's choose the plane

$$\Pi = \{(x, y, 1) \,:\, (x, y) \in \mathbb{A}^2(k)\}$$

and let $C = \{f(x, y) = 0\} \subset \Pi$ be a curve.

Lemma 13.9. *Write $f(x, y) = \sum a_{ij} x^i y^j$ and let $d = \deg(f)$. Let $F(X, Y, Z)$ be the polynomial obtained from f by replacing each monomial $x^i y^j$ by $X^i Y^j Z^{d-i-j}$, so that each term has degree d. Let $V = \{F(X, Y, Z) = 0\} \subset \mathbb{A}^3(k)$. Then V is a cone with $V \cap \Pi = C$.*

Proof. We have

$$F(\lambda x, \lambda y, \lambda) = \sum a_{ij} (\lambda x)^i (\lambda y)^j \lambda^{d-i-j} = \lambda^d F(x, y, 1),$$

so if $(x, y, 1) \in C$ then $F(x, y, 1) = f(x, y) = 0$, and so $F(\lambda x, \lambda y, \lambda) = \lambda^d 0 = 0$ for all $\lambda \in k$, so $\{F = 0\}$ contains all the rays through points of C. □

Definition 13.10. We call the polynomial F in Lemma 13.9 the *homogenisation* of f.

Example 13.11. If $f(x, y) = x^2 + y^2 - 1$ then $F(X, Y, Z) = X^2 + Y^2 - Z^2$ and the cone on the circle is given by $\{X^2 + Y^2 = Z^2\}$ as claimed earlier.

Example 13.12. Let $f(x, y) = x^2 - y^2 - 1$ (which defines a hyperbola). We get $F(X, Y, Z) = X^2 - Y^2 - Z^2$. This was precisely the cone we used in Example 13.8.

The homogenisation of f has the property that every term has degree exactly d. Such a polynomial is called *homogeneous*. We now take a moment to discuss homogeneous polynomials in more detail.

Definition 13.13. We say that a polynomial f is homogeneous of degree d if

$$f(\lambda x_1, \ldots, \lambda x_n) = \lambda^d f(x_1, \ldots, x_n).$$

For example:

- $x^2 + y^2 - z^2$ is homogeneous (of degree 2);
- $x^3 + y^3 - z^3$ is homogeneous (of degree 3);
- $y^2 - x^3$ is not homogeneous[24] (of any degree).
- $y^2 + y^3$ is also not homogeneous[25]
- More generally, $x_1^{i_1} x_2^{i_1} \cdots x_n^{i_n}$ is homogeneous of degree $i_1 + i + 2 + \cdots + i_n$. We say that $x_1^{i_1} x_2^{i_1} \cdots x_n^{i_n}$ is a monomial of degree $i_1 + \cdots + i_n$.
- Any linear combination of monomials of degree d is homogeneous of degree d.

This means we can write any polynomial f as

$$f = f_0 + f_1 + f_2 + \cdots + f_d$$

where f_ℓ is the homogeneous polynomial of degree ℓ obtained by grouping all the monomials of degree ℓ.

Lemma 13.14. *If $f_1, \ldots, f_m$ are homogeneous (possibly of different degrees $d_1, \ldots, d_m$) then the variety $V_k(f_1, \ldots, f_m)$ is a cone.*

Proof. If $(x_1, \ldots, x_n) = (\lambda x_1', \ldots, \lambda x_n')$ then $f_j(x_1, \ldots, x_n) = \lambda^{d_j} f_j(x_1', \ldots, x_n')$, so

$$(x_1, \ldots, x_n) \in V_k(f_1, \ldots, f_m) \text{ if and only if } (x_1', \ldots, x_n') \in V_k(f_1, \ldots, f_m). \qquad \square$$

Lemma 13.15. *Suppose that the ground field has infinitely many elements. Let V be a cone and let $f \in \mathbb{I}(V)$ (i.e. f is a polynomial which vanishes on V). By grouping the monomials of f, write*

$$f = f_0 + f_1 + \cdots + f_d$$

with each f_ℓ homogeneous of degree ℓ. Then $f_\ell \in \mathbb{I}(V)$ for all ℓ.

Proof. If $f(p) = 0$ then $f(\lambda p) = 0$ for all $\lambda \in k$. But

$$f(\lambda p) = f_0(p) + \lambda f_1(p) + \cdots + \lambda^d f_d(p).$$

Pick $d + 1$ different values of λ (possible because k is infinite), say $\lambda_0, \ldots, \lambda_d$. We get $d + 1$ equations

$$\begin{aligned} f_0(p) + \lambda_0 f_1(p) + \cdots + \lambda_0^d f_d(p) &= 0, \\ f_0(p) + \lambda_1 f_1(p) + \cdots + \lambda_1^d f_d(p) &= 0, \\ \vdots \qquad\qquad\qquad & \quad \vdots \\ f_0(p) + \lambda_d f_1(p) + \cdots + \lambda_d^d f_d(p) &= 0. \end{aligned}$$

for the $d + 1$ unknowns $f_0(p), \ldots, f_d(p)$. Rewrite this as a matrix equation

$$\begin{pmatrix} 1 & \lambda_0 & \cdots & \lambda_0^d \\ 1 & \lambda_1 & \cdots & \lambda_1^d \\ \vdots & \vdots & & \vdots \\ 1 & \lambda_d & \cdots & \lambda_d^d \end{pmatrix} \begin{pmatrix} f_0(p) \\ f_1(p) \\ \vdots \\ f_d(p) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

[24] It is *weighted homogeneous*: if y scales as λ^3 and x as λ^2 then the polynomial scales as λ^6. Weighted projective geometry is as interesting and versatile as what we are doing now.

[25] This one is not even weighted homogeneous.

The matrix here is called the *Vandermonde matrix*. Its determinant is[26]

$$\prod_{i \neq j} (\lambda_i - \lambda_j) \neq 0$$

therefore we can invert the matrix and we find that $f_0(p) = \cdots = f_d(p) = 0$. $\square$

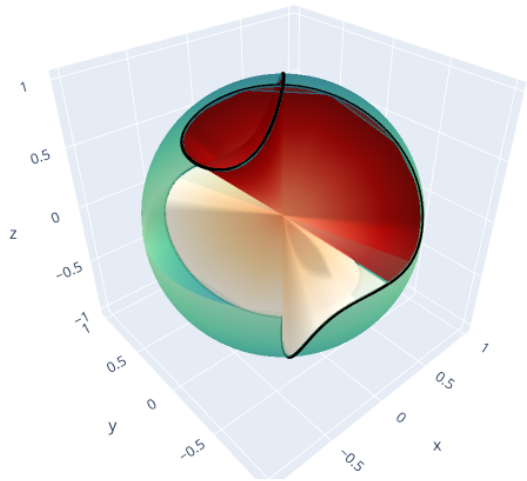
This means that if we are looking at cones, we can restrict attention to homogeneous polynomials.

13.4 Geometry at infinity

Lemma 13.9 tells us that if $C = \{f = 0\}$ and F is the homogenisation of f then $\mathrm{Cone}(C)$ is contained in the cone $V = \{F = 0\}$, but it is usually not true that $\mathrm{Cone}(C) = V$: the cone V will usually contain rulings that miss the original plane Π. For example, in Example 13.12, the rulings $X = \pm Y$, $Z = 0$ correspond to points at infinity in the hyperbola. So, remarkably, this construction is telling us which points we need to add at infinity in our projective variety.

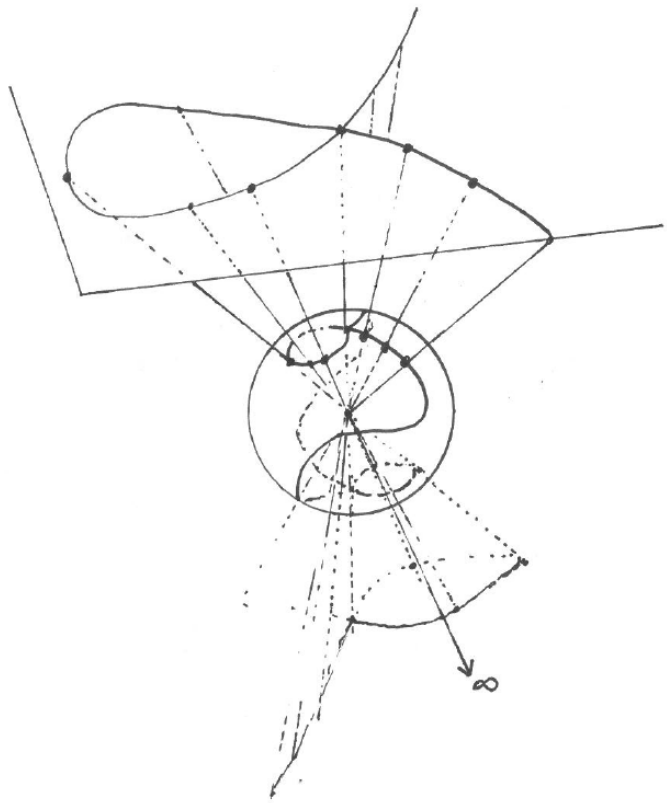
How do we find these points at infinity more generally? If our plane Π is $\{Z = 1\}$ then it will fail to see any rays in the parallel plane $\{Z = 0\}$. So the points at infinity are precisely the rulings contained in $\{Z = 0\}$. So the equation of these lines is $\{F(X, Y, 0) = 0\}$.

Example 13.16. Consider the nodal cubic curve $C = \{y^2 = x^3 + x^2\}$. The homogenisation of $f(x, y) = y^2 - x^3 - x^2$ is $F(X, Y, Z) = Y^2Z - X^3 - X^2Z$. The figure below shows the cone $Y^2Z = X^3 + X^2Z$ and the curve where it intersects the sphere (we've chosen radius < 1 to separate it from the plane $Z = 1$).



If we project this radially onto the $Z = 1$ plane we get the cubic curve $y^2 = x^3 + x^2$. There is precisely one point at infinity: if we set $Z = 0$ then the equation becomes $X^3 = 0$ and the only ruling satisfying $X^3 = Z = 0$ is the Y-axis. We see this affine slice and some of the rulings (including the point at infinity) below.

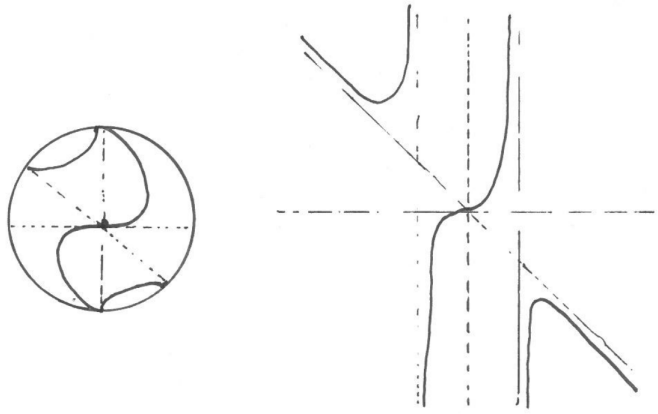
[26] Exercise!



Below you can see the curve drawn in a hemisphere (viewed from above). When you reach the equator at A, you re-emerge at A' and we see that the curve is (topologically) a figure 8.

To get a better understanding of the geometry at infinity, we can pick a different affine slice. For example, let Π' be the plane $\{Y = 1\}$; write the points of Π' as $(x, 1, z)$. The rays "at infinity" in our first slice intersect Π' at points $(x, 1, 0)$, so there is actually a whole line of points at infinity. (In fact, we're still missing the ray pointing in the $(1, 0, 0)$-direction, so there is a whole loop of points at infinity, but this loop intersects each affine slice in a straight line).

Example 13.17. Continuing Example 13.16, we use the affine slice $\Pi' = \{Y = 1\}$ and get the curve $\{z = x^3 + x^2 z\} \subset \Pi'$. The line at infinity from the Π-slice looks like the x-axis in this slice; the point at infinity of the cubic corresponds to the ruling which intersects Π' at $(0, 1, 0)$, i.e. $x = z = 0$. We can see that the curve intersects this line with multiplicity 3.



In the Π'-slice, the rulings at infinity can be found from $Y^2Z = X^3+X^2Z$ by setting $Y = 0$ to get $X^2(X + Z) = 0$. The curve C' has three asymptotes: two vertical (corresponding to $X^2 = 0$) and one along $X = -Z$ (corresponding to $X + Z = 0$).

14 Projective geometry, II

14.1 Projective space

Spaces of lines are hard to imagine. It's not too bad when the lines are real lines and the space is 3-dimensional, but when we work over $\mathbb{C}$ (or worse) and when we work in higher dimensions, our intuition is very bad. For this reason, mathematicians have invented a space called *projective space* which provides a very geometric language for working with spaces of lines and formally manipulating cones.

Definition 14.1. The n-dimensional projective space over a field k, written $\mathbb{P}^n(k)$, is the set of all lines through the origin in $\mathbb{A}^{n+1}(k)$.

To specify a line L through the origin, you only need to specify another point $p \in L$ (not the origin); all the other points are then related to p by scaling:

$$L = \{x \in \mathbb{A}^n(k) \,:\, x = \lambda p \text{ for some } \lambda \in k\}.$$

If $p \neq 0$, we write $[p]$ for the unique line through 0 and p. If $p = (x_0, \ldots, x_n)$ then we usually write

$$[p] = [x_0 : \cdots : x_n].$$

The numbers $x_0, \ldots, x_n$ are called *homogeneous coordinates* on $\mathbb{P}^n(k)$. They are not really well-defined coordinates, because many sets of coordinates correspond to the same point. For example $[1 : 1 : 0] = [2 : 2 : 0] \in \mathbb{P}^2(k)$. Another way to think about this is to define an equivalence relation $\sim$ on $\mathbb{A}^{n+1}(k) \setminus \{0\}$:

$$p \sim q \text{ if and only if } p = \lambda q \text{ for some } \lambda \in k.$$

Then $\mathbb{P}^n(k) = \mathbb{A}^{n+1}(k)/\sim$.

When we pick an affine slice, we are essentially fixing one of the homogeneous coordinates: since each line intersects the slice in a single point, the other coordinates have uniquely determined values. We call this an *affine coordinate chart* $\mathbb{A}^n(k) \subset \mathbb{P}^n(k)$.

Example 14.2. The slice $Z = 1$ gives us an affine chart consisting of points $[X : Y : 1] \in \mathbb{P}^2(k)$. If $p = [X : Y : Z] \in \mathbb{P}^2(k)$ then p is in this affine chart if and only if $Z \neq 0$:

- If $Z \neq 0$ then $[X : Y : Z] \sim [X/Z : Y/Z : 1]$, which is in the chart.
- If $Z = 0$ then every rescaling of $[X : Y : 0]$ has the third homogeneous coordinate equal to zero, so this corresponds to a line that does not intersect the slice.

The coordinates $x = X/Z$ and $y = Y/Z$ are now well-defined coordinates on the affine chart: if you scale $[X : Y : Z]$ by λ then x and y are left unchanged:

$$X/Z \mapsto \lambda X/\lambda Z = X/Z, \qquad Y/Z \mapsto \lambda X/\lambda Z = Y/Z.$$

We cannot choose affine coordinates globally on $\mathbb{P}^n(k)$, just as we cannot draw a map of the whole Earth on a flat piece of paper. However, we can cover $\mathbb{P}^n(k)$ with $n + 1$ coordinate charts:

$$\{x_0 \neq 0\},\ \{x_1 \neq 0\},\ \ldots, \{x_n \neq 0\}.$$

Where these charts overlap, we can convert between the coordinates by passing to homogeneous coordinates.

Example 14.3. Take $PP^1(\mathbb{C})$. There are two charts: $x_0 \neq 0$, containing the points $[1 : z]$ and $x_1 \neq 0$, containing the points $[z' : 1]$. Each chart is a copy of $\mathbb{C}$, and misses out a single "point at infinity", so $\mathbb{P}^1(\mathbb{C})$ is a commonly-used model for $\mathbb{C} \cup \{\infty\}$. It is often drawn as a sphere: the north pole is $[1 : 0]$, and the chart $[z : 1]$ is idenfitifed with $\mathbb{C}$ by stereographic projection.

The overlap $\{x_0 \neq 0\} \cap \{x_1 \neq 0\}$ consists of points where both homogeneous coordinates are nonvanishing, i.e. $z \neq 0$ and $z' \neq 0$. To convert between these charts, we see that

$$[1 : z] \sim [z' : 1] \text{ if and only if } z' = 1/z,\ z = 1/z'.$$

From this point of view $\mathbb{P}^n$ is just a space which cannot be covered by a single coordinate chart, and the "points at infinity", rather than being somehow mysterious and ineffable, are just points that didn't fit into our chosen coordinate chart.

Definition 14.4. Given a cone $V \subset \mathbb{A}^{n+1}(k)$, the projective variety $\mathbb{P}(V) \subset \mathbb{P}^n(k)$ is the set of rulings of V considered as a subset of $\mathbb{P}^n(k)$. Given an affine chart $A = \mathbb{A}^n(k) \subset \mathbb{P}^n(k)$ with $D = \mathbb{P}^n(k) \setminus A$, the curve $A \cap \mathbb{P}(V)$ is an affine slice of $\mathbb{P}(V)$ and the "points at infinity" are the points of $\mathbb{P}(V) \cap D$.

One of the biggest payoffs of introducing projective spaces and projective varieties is that Bézout's theorem becomes an equality:

Theorem 14.5 (Projective Bézout theorem)**.** *Let k be an algebraically closed field. Let $F(X, Y, Z)$ and $G(X, Y, Z)$ be homogeneous polynomials of degrees c and c' respectively and let $C = \{F = 0\} \subset \mathbb{P}^2(k)$ and $C' = \{G = 0\} = \mathbb{P}^2(k)$ be the corresponding projective curves. If $\gcd(F, G) = 1$ then*

$$\sum_{p \in C \cap C'} i_p(C, C') = cc'.$$

Here, $i_p(C, C')$ is measured by taking an affine chart containing p and calculating it there. In other words, this theorem is saying that the points we were missing in the affine Bézout theorem (responsible for the $\leq$ symbol) were hiding at infinity. We will not prove this theorem: the proof involves a careful analysis of the maps ψ and θ in the proof of the affine Bézout theorem. Instead, let's just show that it works for lines (i.e. curves of degree 1), because the proof reduces to linear algebra (and doesn't even require k to be algebraically closed).

Projective Bézout for lines. A projective line is cut out by a nonzero linear equation, so

$$F(X, Y, Z) = aX + bY + cZ, \quad G(X, Y, Z) = dX + eY + fZ$$

for some $a, b, c, d, e, f \in k$. Since $\gcd(F, G) = 1$, (d, e, f) is not a multiple of (a, b, c). Put the matrix

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

into echelon form. Since the bottom row is not a multiple of the top, and neither row is zero, the reduced echelon form must be

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

so the kernel has dimension one. In other words, there is a unique line of vectors $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ such that

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} aX + bY + cZ \\ dX + eY + fZ \end{pmatrix}.$$

That means there is a unique point $[X : Y : Z] \in \mathbb{P}^2(k)$ in the intersection $\{F = 0\} \cap \{G = 0\}$. $\square$

Example 14.6. The affine lines $x = 0$ and $x = 1$ are parallel. They correspond to the projective lines $X = 0$ and $X = Z$ (by homogenising). These intersect at the point $[0 : 1 : 0]$. In terms of the proof, the matrix is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

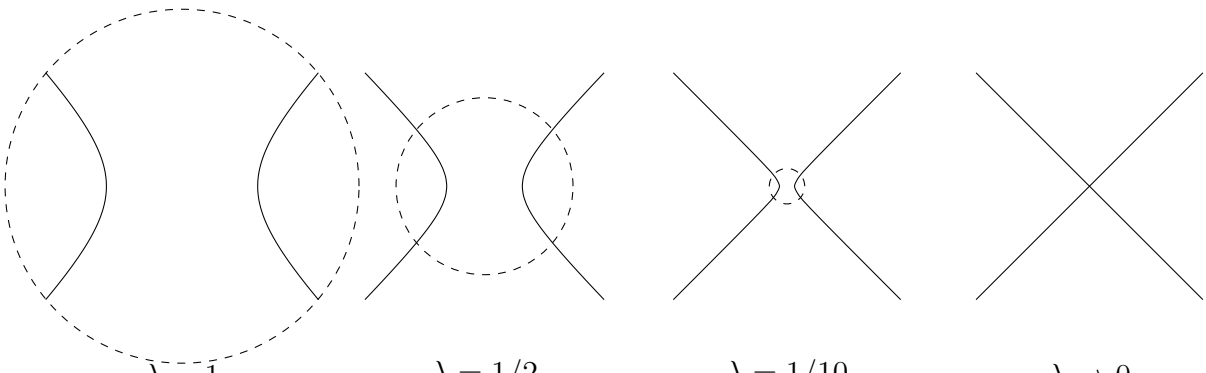
whose kernel is spanned by $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$.

14.2 Asymptotes

We will now make more precise the relationship between the points at infinity and the asymptotes of our curve.

We would like to say that a curve C is asymptotic to a collection of lines if, when viewed on a very large scale, C is very close to these lines. The further we zoom out, the harder it should be to distinguish C from its asymptotes.

It is easiest to make this rigorous by writing equations. Let $f(x, y) = 0$ be the equation of C and suppose it has degree d. The transformation $T_\lambda(x, y) = (\lambda x, \lambda x)$ rescales the plane by a factor of λ. Let $X = \lambda x$ and $Y = \lambda y$ be the rescaled coordinates. Write f in terms of these new coordinates and call the resulting polynomial $f_\lambda(X, Y)$: for example, if $f(x, y) = xy - 1$ then $f_\lambda(X, Y) = \frac{XY}{\lambda^2} - 1$. The rescaled curve $T_\lambda(C)$ is defined by the equation $f_\lambda = 0$, or equivalently $\lambda^d f_\lambda = 0$ (to clear the denominators). Now observe that $\lambda^d f_\lambda(X, Y) = F(X, Y, \lambda)$ where F is the homogenisation of f.



$\lambda = 1$ $\quad$ $\lambda = 1/2$ $\quad$ $\lambda = 1/10$ $\quad$ $\lambda \to 0$

We are going to let $\lambda \to 0$, as this will correspond to "zooming out" (a point (x, y) with

large radius will end up close to the origin in (X, Y) coordinates). In the limit, we are left with the curve $F(X, Y, 0) = 0$, which is precisely the collection of rulings corresponding to points "at infinity" of the affine slice C.

Another way to think about this is that the rescaled curve $T_\lambda(C)$ is the intersection of the cone $\{F(X, Y, Z) = 0\}$ with the plane $Z = \lambda$, and as $\lambda \to 0$, we approach the plane $Z = 0$.

Corollary 14.7. *An affine curve of degree d can have at most d asymptotes.*

Proof. The asymptotes are precisely the rulings satisfying $F(X, Y, 0) = 0$. Each such ruling gives a linear factor $aX + bY$ of $F(X, Y, 0)$. Since $F(X, Y, 0)$ has degree d, there are d such factors (counted with multiplicity). □

Remark 14.8. You can think of this as Bézout's theorem applied to the intersection between C and the line at infinity.

14.3 Conic classification revisited

With all these ideas in mind, we return to our classification of curves of degree 2. Rather than trying to classify *affine curves*, we can classify *projective curves*, that is cones in $\mathbb{C}^3$ cut out by a *homogeneous* equation of degree 2. Two such cones are equivalent if they are related by a linear change of coordinates of $\mathbb{C}^3$.

Remark 14.9. We can interpret such a linear map transformation as a coordinate transformation in two variables involving *rational maps* in the sense of Example 13.1. For example,

$$(x, y, z) = (AX + BY + XZ, DX + EY + FZ, GX + HY + IZ)$$

could be interpreted as a map from the $Z = 1$ plane to the $z = 1$ plane given by

$$(x, y) = \left(\frac{AX + BY + C}{GX + HY + I}, \frac{DX + EY + F}{GX + HY + I}\right).$$

Theorem 14.10. *A homogeneous equation of degree 2 in three variables over $\mathbb{C}$ is linearly equivalent to one of the following three equations:*

$$X^2 + Y^2 + Z^2 = 0, \qquad X^2 + Y^2 = 0, \qquad X^2 = 0.$$

Proof. Suppose that

$$F(X, Y, Z) = aX^2 + bY^2 + cZ^2 + dXY + eXZ + fYZ$$

is our equation. Write

$$v = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}, \text{ and } M = \begin{pmatrix} a & d/2 & e/2 \\ d/2 & b & f/2 \\ e/2 & f/2 & c \end{pmatrix}.$$

We have

$$F(X, Y, Z) = v^T M v.$$

If $A \in GL(3, \mathbb{C})$ is an invertible 3-by-3 matrix then, after changing coordinates by A, the equation of our cone becomes $w^T N w = 0$ where $w = A^{-1}v$ and $N = (A^T)MA$. Two matrices M and N which are related this way are said to be *congruent*. In MATH220, you saw that a symmetric matrix can be diagonalised: in other words M is congruent to a diagonal matrix, say

$$\begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}.$$

If $\lambda_j \neq 0$ then you can rescale the jth coordinate by $1/\sqrt{\lambda_j}$ and find a congruent matrix whose diagonal entries are either zero or one. Up to permuting the variables, this gives the three possibilities in the theorem. □

14.4 Moduli of cubics

Can we come up with a similarly slick classification of cubic curves? In the setting of projective geometry over $\mathbb{C}$, most of Newton's 78 cases become equivalent: they are just cubics which intersect the line at infinity in different ways, or whose real parts happen to look different, and the enumeration becomes more transparent:

Theorem 14.11. *A smooth homogeneous cubic is equivalent to one of the cubics in the following family:*

$$Y^2Z = X(X - Z)(X - \lambda Z), \quad \lambda \neq 0, 1.$$

Two cubics from this family are equivalent if and only if their j-invariants coincide, where

$$j(\lambda) = 256\frac{(\lambda^2 - \lambda + 1)^3}{\lambda^2(\lambda - 1)^2}.$$

There are two further irreducible singular curves:

$$Y^2Z = X^2(X + 1), \qquad Y^2Z = X^3$$

and five reducible curves:

$$X^3 = 0, \qquad X^2Y = 0, \qquad XYZ = 0, \qquad XY(X + Y) = 0, \qquad X(Y^2 + Z^2) = 0.$$

Note that the cubic $\{Y^2Z = X(X - Z)(X - \lambda Z)\}$ intersects $\{Z = 0\}$ at a single point $[0 : 1 : 0]$ with multiplicity 3. In fact, the line at infinity is a tangent to the curve at this point. Such a point (where the tangent line intersects the curve with multiplicity at least 3) is called a *flex* of the curve, and the most important part of the proof of this theorem is showing that any smooth cubic curve has at least one flex (actually it has nine). Then you can change coordinates to ensure that a flex is at $[0 : 1 : 0]$. This ensures that there are no terms X^2Y, XY^2, Y^3 in the equation. You can learn more about flexes on Sheet 5.

Remark 14.12. Over fields that are not $\mathbb{C}$, at least if the equation $6 \neq 0$ holds in your field, then you can put a cubic into the form $Y^2Z = X^3 + AXZ^2 + BZ^3$. The cubic is smooth if and only if $4A^3 + 27B^2 \neq 0$ and the j-invariant (determining equivalence over the algebraic closure) is

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

The set of inequivalent cubics with the same j-invariant is in bijection with a certain Galois cohomology group. For example, if $\sqrt{c} \notin k$ then the curves

$$Y^2 = X^3 + AXZ^2 + BZ^3 \text{ and } Y^2 = X^3 + c^2AXZ^2 + c^3BZ^3$$

have the same j-invariant but fail to be isomorphic over k. These are the only two possibilities if $j \neq 0, 1728$. See Chapter X, Proposition 5.4 of Silverman (2009) *The arithmetic of elliptic curves* for a complete description[27].

This has a different flavour from the situation in degree 2: there is a *continuum* of inequivalent smooth curves. In other words, the set of all cubic curves is itself a space! It is called a *moduli space*. The moduli space of cubics is 1-dimensional (with coordinate $j(\lambda)$). You can think of the singular curves as representing limit points (e.g. $\lambda \to 0$ or $\lambda \to 1$) of this space. The study of moduli spaces of varieties is one of the major themes of research in algebraic geometry. Moduli spaces of curves of higher degree have been very intensively studied, and have turned up a wealth of surprises. Almost nothing is known about moduli spaces of higher-dimensional varieties (for example surfaces of general type).

[27] But be warned that you might need to read Chapters I–IX first.