Intersections of Algebraic Plane Curves

Nic Ford

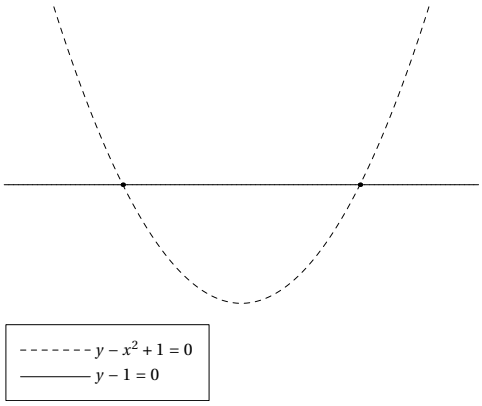
Mathcamp 2023

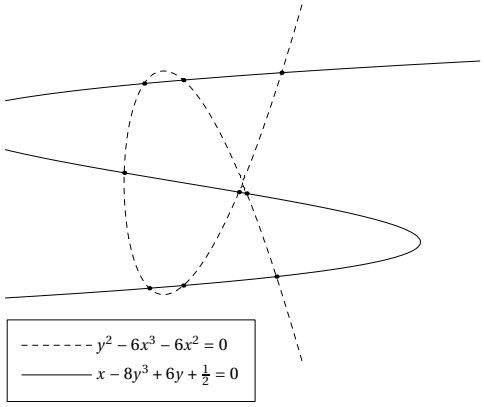
1 Introduction

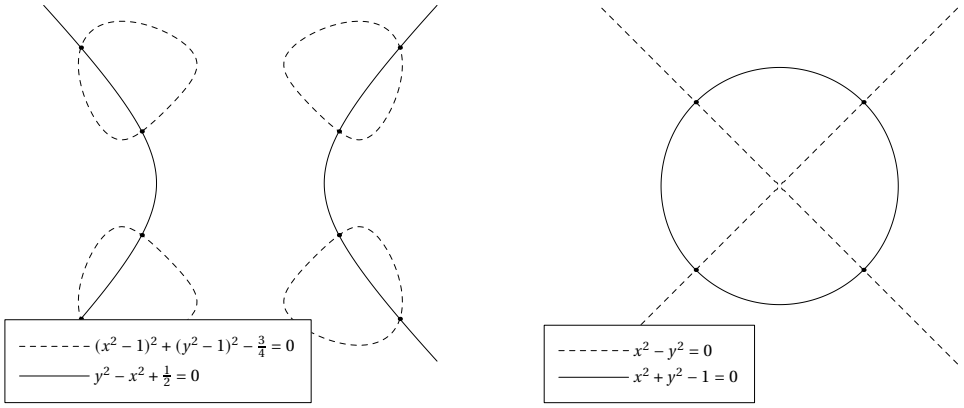
This entire class will be focused on a single question. Given two curves in the plane, each of which is described by a polynomial equation, we will be interested in counting the number of points where the two curves intersect. More formally, the objects we're going to look at are defined as follows:

Definition 1.1. For any nonconstant polynomial $f(x,y)$ in two variables, we'll define the **algebraic plane curve** cut out by f to be the set of points (x,y) in the plane which satisfy the equation $f(x,y)=0$.

We exclude constant polynomials because, if we did include them, the "curve" they cut out would be either empty or the whole plane, depending on whether the constant is zero or not. Here are some pairs of algebraic plane curves with their intersection points marked:



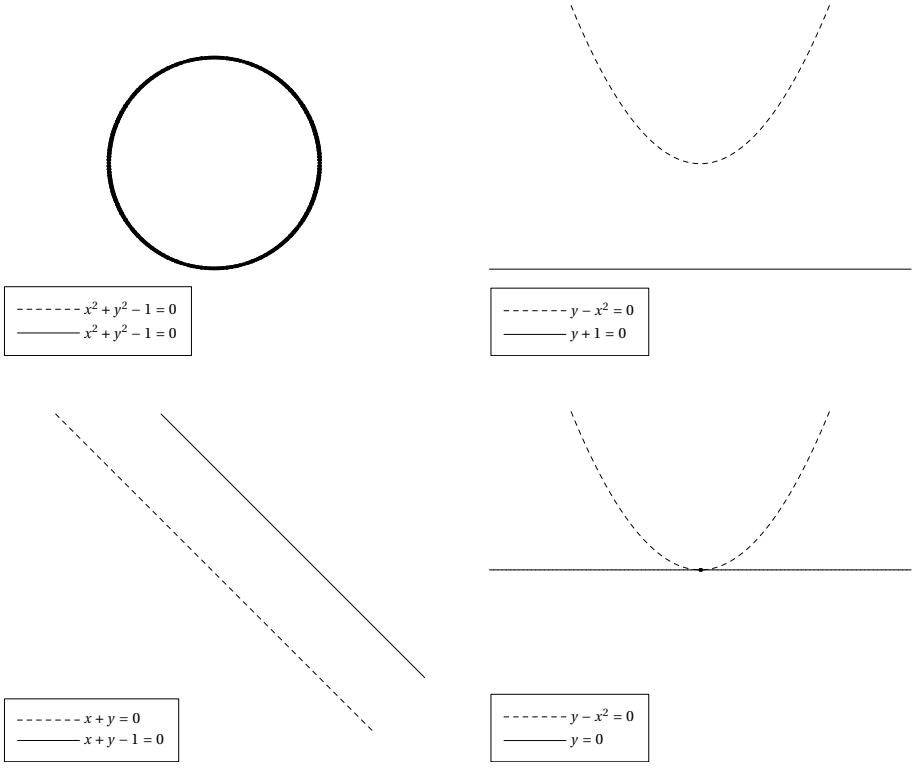




In each of these examples, there's an interesting pattern to notice if we look at the *degrees* of the polynomials which cut out each curve: the number of intersection points is always exactly equal to the product of the degrees of the two polynomials. This is the result we'll spend this class proving:

Theorem 1.2 (Bézout's Theorem, False Version)**.** *Let f and g be nonzero polynomials in two variables, with* $\deg f = d$ *and* $\deg g = e$*. Then the curves cut out by f and g intersect in exactly de points.*

One thing you might have noticed about this theorem, if you either thought about it hard enough or read its title closely enough, is that it's not true. Here are a few counterexamples:



If we're going to spend this entire class proving this theorem, we're going to have to find a way to modify the statement into one that's true! By studying these counterexamples closely, we can see a few things we might do:

- The first example involves the intersection of a curve with itself, which of course gives infinitely many intersection points. The solution to this problem will be to simply disallow it.

  It's not quite as simple as asking for f to not be equal to g, though. Notice that if we factor f as a product of two polynomials, say $f(x,y) = f_1(x,y)f_2(x,y)$, then the set of points where $f = 0$ is the set of points where *either* f_1 or f_2 is zero. This means that we'll get a similar problem if f and g have a common factor — if, say, $f = f_1 f_2$ and $g = f_1 g_2$ — even if f and g aren't equal. We'll have to disallow this case too.

- Next, look at the second example, where $f(x,y) = y - x^2$ and $g(x,y) = y + 1$. Any intersection point of these two curves would be at a point where $y = x^2$ and $y = -1$, which of course would mean that $x^2 = -1$. There is of course no real number that satisfies this equation, but there are *complex* numbers which do: we can take $x = i$ or $x = -i$, which gives us the two intersection points $(i,-1)$ and $(-i,-1)$. This suggests that we're going to need to work over the complex numbers.

- The third example consists of the two parallel lines cut out by $f(x,y) = x + y$ and $g(x,y) = x + y - 1$. The complex numbers won't help you here — if we had some (x,y) that satisfied both of these equations we would be able to conclude that $0 = 1$. Notice, though, that if I turn one of the two lines just a little bit, the lines will no longer be parallel, and so they will intersect. As the lines get closer and closer to being parallel, this point moves further and further away, shooting "off to infinity" in the direction of the parallel lines.

  This suggests that, if we somehow augment the plane with some extra "points at infinity," then this process might be able to approach some sort of limit, and we might be able to recover Bézout's Theorem. We'll start on this project in the very next section.

- Finally, look at the fourth example, where $f(x,y) = y - x^2$ and $g(x,y) = y$. An intersection point of these two curves must happen at a point where $y = x^2$ and $y = 0$, which means $x^2 = 0$. Unlike the case which inspired us to work over the complex numbers, this last equation only has the solution $x = 0$, and we correspondingly get $(0,0)$ as the only intersection point, even over $\mathbb{C}$. And, while we won't be able to prove this precisely until we have a formal definition of points at infinity, it seems visually clear that these two curves go off to infinity in totally different directions, and so probably don't intersect at infinity either.

  The solution to this problem is perhaps a bit more abstract than the others. You might recall that, when counting roots of polynomials in one variable, in order for the total count to equal the degree you need to count the roots *with multiplicity*. The same will be true for Bézout's Theorem. (In fact, as this example shows, the task of counting the roots of some polynomial $p(x)$ can be seen as a *special case* of Bézout's Theorem — look at the intersection of $y - p(x) = 0$ and $y = 0$.)

  In this case, because we ended up with the equation $x^2 = 0$, it's probably easy to believe that the "correct" multiplicity to assign is 2, and this does give us the right count for Bézout's Theorem. In general, though, when neither of the curves is a line like $y = 0$, the definition of intersection multiplicity is trickier, and it will have to wait until we're close to the end of our journey.

It will turn out that, once all four of these problems have been addressed, Bézout's Theorem will actually be true. The first two are relatively straightforward to solve: disallow pairs of polynomials with common factors, and work over the complex numbers. The last two, though, will be quite a bit harder: it's not clear yet precisely what we even *mean* by points at infinity or points with multiplicity.

That's the task ahead of us for the next few days. We will give precise definitions of these last two concepts and we'll use them to give a proof of the new, corrected version of the theorem.

Exercises

In these exercises, and for the entire rest of this class, you should feel free to assume the Fundamental Theorem of Algebra, which says that any non-constant one-variable polynomial over $\mathbb{C}$ has a zero.

Exercises that are especially important for the rest of the class have been marked with ⚠.

1.1. How many intersection points are there between the curves $x^2+3y^2-4=0$ and $3x^2+y^2-4=0$? Draw a picture of these two curves and their intersection points. If Bézout Theorem is true, what are the intersection multiplicities at all of the points you found? Are there any intersection points at infinity?

1.2. Give an example of a nonconstant polynomial $f(x,y)$ with real coefficients such that there are no points $(x,y)\in\mathbb{R}^2$ for which $f(x,y)=0$. (When this happens, we say that the curve corresponding to f has **no real points**.)

1.3. In Exercise 1.2, you found a curve with no real points. In this exercise, we'll show that this can't happen over the complex numbers. (This is another reason, in addition to Bézout's Theorem, that the study of algebraic curves is much nicer over $\mathbb{C}$ than over $\mathbb{R}$.) Throughout this exercise, let $f(x,y)$ be a nonconstant polynomial with complex coefficients.

(a) Suppose the curve cut out by f has no points on the horizontal line $y=a$. What can we conclude about $f(x,a)$, thought of as a polynomial in x?

(b) Prove that this can only happen for finitely many values of a. *[Hint: Think of f as a polynomial in x with coefficients that are polynomials in y. If, for some a, there are no solutions to the equation $f(x,a)=0$, what does that mean about these coefficients?]*

(c) Conclude that there must be at least one point $(x,y)\in\mathbb{C}^2$ with $f(x,y)=0$.

1.4. ⚠ In Exercise 1.3, you showed that for every non-constant two-variable complex polynomial $f(x,y)$, there is at least one point in $\mathbb{C}^2$ where it vanishes. In this exercise, you'll show that (as long as $f\neq 0$) there is also at least one point where f *doesn't* vanish.

(a) Suppose f is zero at every point on the horizontal line $y=a$. Prove that $y-a$ divides f, that is, there is some polynomial $g(x,y)$ such that $f(x,y)=(y-a)g(x,y)$. *[Hint: First, plug $y=(y-a)+a$ into f and write f as a polynomial in $(y-a)$ and x.]*

(b) Conclude that if f is not the zero polynomial then there can only be finitely many a's with the property that f vanishes on the entire horizontal line $y=a$.

(c) Finally, conclude that there are infinitely many points where f doesn't vanish.

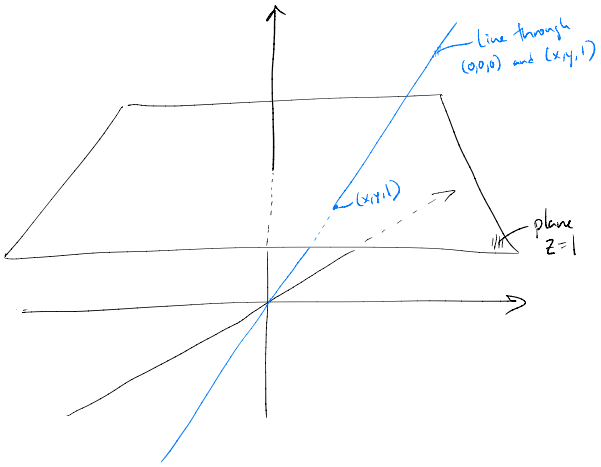
2 The Projective Plane

In the last section, we identified four problems with the original, incomplete statement of Bézout's Theorem: the polynomials can't have a common factor, we need to work over the complex numbers, we need to count intersection points with multiplicities, and we need to count "points at infinity." Of these four problems, the first two are by far the simplest: for the first, just don't allow polynomials with common factors; for the second, consider polynomials with complex coefficients and look for intersection points with complex coordinates. The problem of how to define multiplicities will need to wait until we've discussed an object called the *resultant*, which will be the main tool we'll use to prove the theorem.

So in this section, we'll learn how to formalize the concept of curves intersecting at infinity. We will do this by defining a space called the *projective plane* which is slightly larger than the ordinary plane, and which in particular contains the points at infinity we're looking for. For every algebraic plane curve, there will be a unique way to extend it to a curve in this new, slightly larger plane. A pair of curves will then be said to "intersect at infinity" if their extensions to the projective plane intersect at a point that isn't on the original, ordinary plane.

2.1 Points as Lines in Space

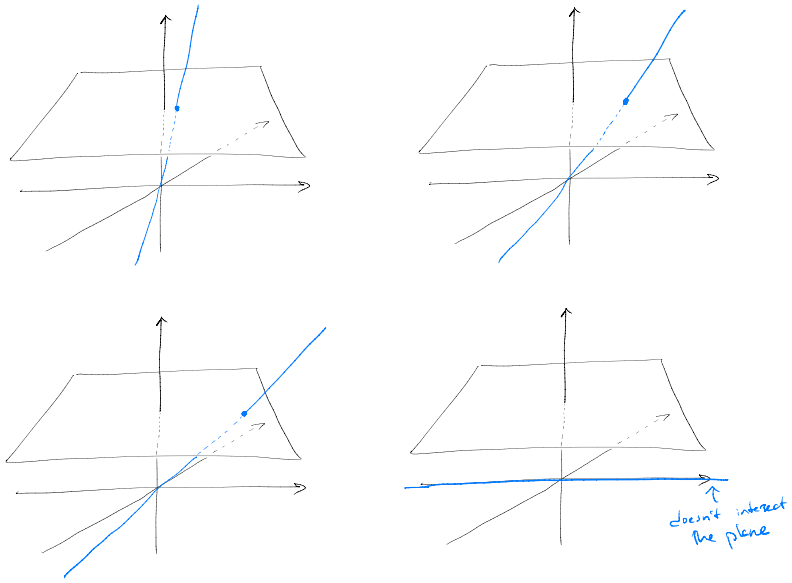
We'll define the projective plane with the help of the following picture. (Even though we just made a big fuss about working over the complex numbers, all of the pictures in this section are going to use the *real* numbers, since $\mathbb{R}^2$ is much easier to visualize than $\mathbb{C}^2$.) Take the plane $\mathbb{R}^2$, and stick it in 3-space as the plane $z = 1$. That is, the point $(x, y) \in \mathbb{R}^2$ will be represented by the point $(x, y, 1) \in \mathbb{R}^3$.



To every point in the plane, we can associate a line through the origin in $\mathbb{R}^3$: since the last coordinate of the point $(x, y, 1)$ isn't 0, we know that $(x, y, 1)$ and $(0, 0, 0)$ aren't the same point, and so there's a unique line passing through both of them. Moreover, this accounts for *almost* all of the lines through the origin in $\mathbb{R}^3$. Given a line through the origin in $\mathbb{R}^3$, we can look at where it intersects the plane $z = 1$, and this will tell us what point it should correspond to. The only

way this will not work is if the line *doesn't* intersect the plane $z = 1$, that is, if it's contained in the xy-plane.

In other words, there is almost, but not quite, a one-to-one correspondence between points in the plane and lines through the origin in $\mathbb{R}^3$. If these extra lines *were* to correspond to points in the plane, where would those points be? We can answer this by taking a line that *does* correspond to a point in the plane, rotating the line until it lies flat in the xy-plane, and keeping track of what happens to the point in the plane as we do this.



In the example shown, the point on the plane always lies along the x-axis, which reflects the fact that the line always stays in the xz-plane, so the point can never have a nonzero y coordinate. Notice that, as the line approaches the xy-plane, the point in the plane gets further and further away from the origin along the x-axis. It might, therefore, make sense to say that when the line *does* finally lie completely flat, it ought to correspond to a point that is "infinitely far away" in the x direction.

This whole story inspires the following definition.

Definition 2.1. The **real projective plane** $\mathbb{RP}^2$ is the set of lines through the origin in $\mathbb{R}^3$. The lines in $\mathbb{R}^3$ that are contained in the xy-plane are called **points at infinity** in the projective plane.

It's important to remember that a *point* of the projective plane is a *line* in $\mathbb{R}^3$. This is confusing enough and easy enough to forget that we'll often use slightly different language to describe the situation, saying that a point $p \in \mathbb{RP}^2$ "corresponds to" a line L through the origin in $\mathbb{R}^3$, rather than the more technically accurate statement that it just "is" that line. My hope is that this will help us keep the distinction between points and lines clear when it matters.

The setup we've described here has one consequence which might not be obvious. The x-axis looks like it has two "ends", one far to the right and one far to the left, so you might expect that this would give us two different points at infinity along the x-axis. But if we take our picture of the rotating line and *keep* rotating the line past the point where it lies in the xy-plane, we

see that the corresponding point in the plane is now once again very far from the origin along the x-axis, but now to the left rather than to the right. In other words, our "point at infinity" is "close" to points on the far left side of the x-axis as well as points on the far right side.

This means that our original mental picture, where there are two points at infinity on the x-axis, one in each direction, requires a small adjustment. There is actually just *one* point at infinity along the x-axis, and it should be thought of as *both* infinitely far to the left *and* infinitely far to the right.

The same thing will be true of every line in the plane, as we'll see more formally in just a moment: when we pass from the ordinary plane to the projective plane, every line gets augmented with one additional point, which we think of as lying "at infinity" along that line in either direction. A good way to visualize this is that, rather than capping off a line at both ends, which would make it look something like a closed line segment, we are instead *attaching* the two ends to each other, resulting in something more like a circle.

2.2 Homogeneous Coordinates

Our goal is to talk about algebraic curves in the projective plane. In order to do this, we're going to need some way to put coordinates on points in $\mathbb{RP}^2$, since otherwise we won't have any numbers to plug into our polynomials. We have already chosen to represent a point from the ordinary plane using the point $(x, y, 1) \in \mathbb{R}^3$, which gives us one possible idea for how to give coordinates to a point in the projective plane: given some point in the projective plane, which corresponds to a line L through the origin in $\mathbb{R}^3$, we could choose coordinates by picking the point on L which has 1 as its z coordinate.

This idea, though, fails to account for our points at infinity, which after all correspond to lines that lie in the xy-plane and so don't have *any* points with 1 as their z coordinate. We would like to have a uniform system of coordinates that works for the entirety of the projective plane, and we can accomplish this if we simply remove the restriction on the z coordinate:

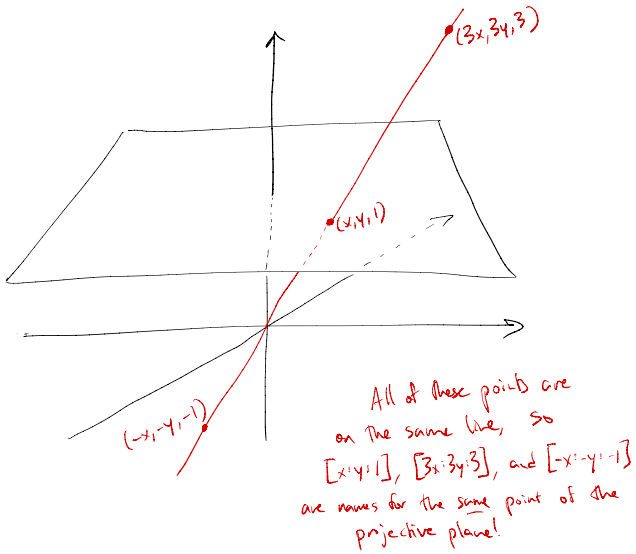
Definition 2.2. Let p be a point in $\mathbb{RP}^2$, corresponding to the line L in $\mathbb{R}^3$. We'll say a triple of real numbers x, y, z gives **homogeneous coordinates** for p if (x, y, z) is some point on L other than the origin. When referring to a point in $\mathbb{RP}^2$ using homogeneous coordinates, we will always use the notation $[x : y : z]$, not (x, y, z), to remind us that we are not talking about the point with those coordinates in $\mathbb{R}^3$.

Since any point other than the origin gives us a unique line passing through both the origin and that point, $[x : y : z]$ will always pick out a valid point of the projective plane as long as x, y, and z aren't all 0. And, since the origin is the only point where two lines through the origin can intersect, any set of homogeneous coordinates picks out a unique line, and therefore a unique point of $\mathbb{RP}^2$. But any point of $\mathbb{RP}^2$ can be described by many different sets of homogeneous coordinates: as long as (x, y, z) and (x', y', z') lie on the same line through the origin, the points $[x : y : z]$ and $[x' : y' : z']$ in $\mathbb{RP}^2$ will be identical.

How can we detect whether this happens just by looking at the coordinates themselves? If (x, y, z) is some non-origin point in $\mathbb{R}^2$, then I encourage you to verify that the line passing through $(0, 0, 0)$ and (x, y, z) is the line $\{(\alpha x, \alpha y, \alpha z) : \alpha \in \mathbb{R}\}$. So what this means is that the coordinates $[x : y : z]$ and $[x' : y' : z']$ refer to the same point of the projective plane if and only if one set of coordinates is a *scalar multiple* of the other, that is, for some $\alpha \neq 0$, we have $x' = \alpha x$, $y' = \alpha y$, and $z' = \alpha z$.

This gives us an alternative, more algebraic way of thinking about the real projective plane: it's the set of all triples of real numbers $[x : y : z]$, except that (a) we exclude the triple $[0 : 0 : 0]$,

and (b) if $\alpha \neq 0$, then $[x : y : z]$ and $[\alpha x : \alpha y : \alpha z]$ denote the same point.



So, given some homogeneous coordinates $[x : y : z]$, we can turn them into a different set of homogeneous coordinates for the same point by multiplying all three coordinates by the same factor $\alpha \neq 0$. (Using $\alpha = 0$ would be a problem, since that would give us $[0 : 0 : 0]$, and this is the only set of homogeneous coordinates that *doesn't* pick out a point of the projective plane!) In particular, if $z \neq 0$, then we are free to multiply by $1/z$, producing $[x/z : y/z : 1]$. In other words, if $z \neq 0$, then the homogeneous coordinates $[x : y : z]$ describe a point on the ordinary plane (namely the point $(x/z, y/z)$). On the other hand, if $z = 0$, then $[x : y : z]$ is a point at infinity.

2.3 Homogenization

The reason we introduced homogeneous coordinates was to enable us to extend algebraic curves from the ordinary plane to the projective plane. Let's do that now.

There is a complication that shows up when we try to define algebraic curves in the projective plane that doesn't appear when in the ordinary plane: homogeneous coordinates give us many different "names" for the same point. Suppose, for example, we tried to define an curve in the projective plane using the equation $x^2 + xy + z^2 - 1 = 0$. Is, for example, the point $[1 : 0 : 0]$ on this curve? It would seem like the answer is yes, because if you plug in $x = 1$, $y = 0$, and $z = 0$ into this equation you see that it's true. But we have a problem, because $[1 : 0 : 0]$ is the same point as $[2 : 0 : 0]$, and $2^2 + 0 \cdot 0 + 0^2 - 1 \neq 0$. So this equation doesn't seem to cut out a well-defined subset of $\mathbb{RP}^2$; the question of whether a point belongs on the curve can depend on which coordinates we use for it.

One way to see what the problem is involves looking at each term of the polynomial separately. Suppose have a term of degree d, say $kx^a y^b z^c$ where $a + b + c = d$. If we replace $[x : y : z]$ with $[\alpha x : \alpha y : \alpha z]$, then $kx^a y^b z^c$ becomes $\alpha^d \cdot kx^a y^b z^c$ — in other words, multiplying x, y, and z by α has the effect of multiplying a term of degree d by α^d. If f is a polynomial in which *every* term has degree d, the same thing will then be true of f.

So, if some polynomial f has only terms of degree d, we won't run into the situation we ran into in our example. This is because, if some triple x, y, z satisfies the equation $f(x, y, z) = 0$,

then we'll have $f(\alpha x, \alpha y, \alpha z) = \alpha^d f(x, y, z) = 0$, so the question of whether a point satisfies the equation *won't* depend on which name we use for the point. Our polynomial above didn't work because it has terms of different degrees: the first three terms have degree 2 but the last term has degree 0.

We can summarize the situation as follows:

Definition 2.3. If $f(x, y, z)$ is a polynomial and every term of f has degree d, then we say f is a **homogeneous polynomial of degree** d. If f is homogeneous, the preceding discussion shows that the set of points in the projective plane where $f = 0$ is well-defined; a subset of the projective plane that arises in this way is called a **projective plane curve**.

Note that, even though a homogeneous polynomial cuts out a well-defined *subset* of the projective plane when you look at the points where it's equal to 0, that doesn't mean it gives a well-defined *function* on the projective plane — it doesn't! The coordinates $[x : y : z]$ and $[\alpha x : \alpha y : \alpha z]$ refer to the same point, but $f(\alpha x, \alpha y, \alpha z) = \alpha^d f(x, y, z)$, so even if f is homogeneous there is no meaningful definition we can assign to "the value of f at this point." Luckily, we are only concerned with which curve f cuts out, so we only care about *whether or not* f *is zero*, and the answer to this question doesn't change if we multiply by α^d when $\alpha \neq 0$.

Our goal at the very beginning of this section was to find a way to decide whether two algebraic curves (in the *ordinary* plane) intersect at infinity. There is one task remaining: we need a way to take an algebraic curve in the ordinary plane and extend it to a curve in the projective plane. Now that we've nailed down the type of polynomials we want to use to describe projective plane curves, we can state our task more precisely. Given a polynomial $f(x, y)$ in two variables, we want to find a *homogeneous* polynomial $g(x, y, z)$ in *three* variables which cuts out a curve that, when you restrict it to the ordinary plane, is the same as the curve cut out by f.

Since the point (x, y) of the ordinary plane corresponds to the point $[x : y : 1]$ in the projective plane, we can restrict our polynomial g to the ordinary plane by just plugging in $z = 1$. So, to go the other direction, we need a way to start with f and produce a homogeneous polynomial g which, if we plug in $z = 1$, gives us back f. This is accomplished by the following procedure:

Definition 2.4. Let $f(x, y)$ be a polynomial of degree d. The **homogenization** of f is the polynomial produced from f by replacing every term $k x^a y^b$ with $k x^a y^b z^{d-a-b}$.

In other words, if the highest-degree term of f has degree d, we homogenize f by taking every term whose degree is smaller than d and adding a high enough power of z to raise its degree up to d. For example, the homogenization of $3x^2 + xy^2 - 1$ is $3x^2z + xy^2 - z^3$: the first term had degree 2, so it needed just one z to get it to degree 3, but the last term had degree 0 and so needed three z's.

As an example, let's see that we've succeeded in making two parallel lines intersect at infinity; since we already know that two non-parallel lines intersect in the ordinary plane, this will essentially give us a proof of Bézout's Theorem in the case that both degrees are 1 (as long as we believe that the intersection multiplicity is also 1).

Suppose $f(x, y) = ax + by + c$ and $g(x, y) = a'x + b'y + c$. The two lines cut out by these equations will be parallel if $a/b = a'/b'$ (or if b and b' are both zero, but let's assume they're both nonzero for now for simplicity). Let's say $a/b = a'/b' = m$. We can multiply these polynomials by a constant without changing the curves they describe, so let's multiply f by $1/b$ and g by $1/b'$. Then, (writing $k = c/b$ and $k' = c'/b'$) we have $f(x, y) = mx + y + k$ and $g(x, y) = mx + y + k'$.

When we homogenize these two polynomials, we get $f(x, y, z) = mx + y + kz$ and $g(x, y, z) = mx + y + k'z$. Now, since these are two parallel lines and not the same line twice, we must have $k \neq k'$. We therefore see that, if $z \neq 0$, then these two polynomials can't vanish simultaneously.

This is what we should expect — this corresponds to looking for an intersection point on the ordinary plane, and there aren't any!

What about at infinity? The points at infinity are the points where $z = 0$. Note that this is itself a linear homogeneous polynomial; for this reason we say that the equation $z = 0$ cuts out the **line at infinity**. If $z = 0$, we are just looking for a point $[x : y : z]$ where $mx + y = 0$. I encourage you to verify that the only such point is $[1 : -m : 0]$. (Remember that if you multiply all the coordinates by a nonzero constant it's the same point!)

So we have successfully found the point at infinity where our two parallel lines intersect. Interestingly, we also see that from this that the point we get depends on the *slope* of the parallel lines. So, in particular, two lines with different slopes — that is, two non-parallel lines — will intersect the line at infinity at different points. We've therefore preserved the fact that two non-parallel lines intersect at just 1 point, even in the projective plane.

Exercises

Exercises that are especially important for the rest of the class have been marked with ⚠.

2.1. For any complex number a, consider the curves cut out by the equations $y - x^2 = 0$ and $x - a = 0$. Where do they intersect in the ordinary plane? Where do they intersect in the projective plane?

2.2. ⚠

(a) Suppose $f(x, y, z)$ is a homogeneous polynomial of degree d. Prove that

$$f(x, y, z) = z^d f(x/z, y/z, 1).$$

(b) Suppose $f(x, y, z)$ and $g(x, y, z)$ are homogeneous polynomials which both have degree d, and that $f(x, y, 1) = g(x, y, 1)$ as polynomials in x and y. Prove that $f = g$. What does this result tell you about algebraic curves?

(c) What can we conclude about f and g if we drop the requirement that they have the same degree? What does *this* result tell you about algebraic curves?

(d) Let $f(x, y)$ be a homogeneous complex polynomial in *two* variables. Prove that f can be written as a product of linear homogeneous polynomials, that is, we have

$$f(x, y) = (a_1 x + b_1 y)(a_2 x + b_2 y) \cdots (a_n x + b_n y)$$

for some a_i's and b_i's. *[Feel free to use the corresponding fact for non-homogeneous complex polynomials in one variable.]*

2.3. In Exercise 1.3, you showed that in the ordinary plane, every curve has at least one complex point. The proof was somewhat involved, but in the projective plane it's quite a bit easier. Prove that, if $f(x, y, z)$ is a nonconstant homogeneous polynomial, then there is always at least one point where f vanishes. *[Hint: One solution involves plugging $z = 0$ into f and then using Exercise 2.2d.]*

2.4. (a) Where does the circle cut out by the equation $x^2 + y^2 = 1$ intersect the line at infinity (again working over the complex numbers)? You should get exactly two points.

(b) An arbitrary circle is given by an equation of the form $(x - a)^2 + (y - b)^2 = r^2$. Prove that *every* circle passes through the two points you found in part (a). (For this reason, they are sometimes called the **circle points**.)

3 Resultants

Bézout's Theorem is all about counting the number of points where two different polynomials in two variables are both equal to zero. We'll begin our investigation of this question by first answering it for polynomials in *one* variable.

This is not just a "practice" version of the question we're actually interested in; answering the one-variable version of the question will be directly useful for answering the two-variable version. We will go into more detail in the following sections, but as a quick preview, suppose we had a way to tell, for any pair of one-variable complex polynomials f and g, whether f and g have a common root, that is, whether there's some complex number α with $f(\alpha) = g(\alpha) = 0$. We could then leverage this in our search for the points where two *two*-variable polynomials are both zero, using the following trick. Given two polynomials $p(x, y)$ and $q(x, y)$, pick some number β and plug it in for y. This has the effect of restricting our attention to the horizontal line $y = \beta$. If we then use our one-variable procedure to ask whether $p(x, \beta)$ and $q(x, \beta)$ have a common root, we will learn whether there is an intersection point on this horizontal line. As we will see in the next section, this will be an important step in our quest to count the total number of intersection points.

The tool that accomplishes this task will be called the *resultant*. In order to build it and prove that it works, we're going to have to make use of a few facts that we won't have time to prove in the main part of this article. We'll go through those now, and then move on to the resultant itself.

3.1 A Quick Review of Determinants

The first of these prerequisites involves the concept of the determinant of a matrix. It is my hope that you've encountered determinants before and that this will serve as review. If not, the quick discussion we're about to have really doesn't come close to doing the topic justice; this section would be a remarkably bad way to learn about determinants for the very first time. In particular, there is a beautiful *geometric* description of the determinant which, in my opinion, is really the best way to think about it, but which we won't touch on at all.

For our purposes, the determinant is a tool for determining whether a system of linear equations has a nontrival solution. We'll start by being precise about what that means.

Definition 3.1. A **system of m homogeneous linear equations in n variables** is a collection of equations of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0. \end{aligned}$$

(You will sometimes see linear equations with nonzero constants on the right side, but we will only be interested in the case where they're all zero; this is what is meant by "homogeneous.")

The a_{ij}'s should be thought of as fixed; we're interested in looking for values of the x_j's which make all the equations true. Any n-tuple of numbers $(x_1, \ldots, x_n)$ which satisfies all m equations in a system of linear equations is called a **solution** of the system. Note that setting all the x_j's to 0 will always give a solution no matter what the equations are; a solution with at least one nonzero number in it is called a **nontrivial solution**.

We can represent a system of equations using a two-dimensional grid of numbers called the **matrix** of the system of equations:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

When a matrix has m rows and n columns like this, we will call it an "$m \times n$ matrix."

One result that you might prove in a linear algebra class is that if there are fewer equations than variables — that is, if $m < n$ — there will always be a nontrivial solution. On the other hand, if $m = n$, the existence of a nontrivial solution depends on the values of the coefficients a_{ij}. This is the case we're interested in.

The first interesting case is when $m = n = 2$, where we have a system of the form

$$\begin{aligned} a\,x + b\,y &= 0 \\ c\,x + d\,y &= 0. \end{aligned}$$

In this case, it turns out that there's a nontrivial solution if and only if one of the equations is a multiple of the other. I encourage you to verify this if you've never done it before, and to check that this is equivalent to the statement that $a\,d - b\,c = 0$.

The quantity $a\,d - b\,c$ is the **determinant** of the 2×2 matrix

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

If M is the matrix, we'll often write its determinant as $\det M$. The utility of the determinant comes from the fact that it allows you to check whether the system has a nontrivial solution by just plugging the coefficients into this formula, which allows you to avoid having to actually find the values of x and y. When you just care about *whether* there is a nontrivial solution and not what the solution actually is — as will be the case for us when we discuss the resultant — computing the determinant can be less work than actually solving the equations.

There is a bigger, more complicated formula for the determinant of a larger system of linear equations. We will actually only need a couple of small facts about this formula, but for completeness here is one definition:

Definition 3.2. Consider an $n \times n$ matrix

$$M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

If $n = 1$, then our matrix has the form (a_{11}), and we'll say that $\det(a_{11}) = a_{11}$. Otherwise, we'll define the determinant of our $n \times n$ matrix recursively in terms of the determinant of an $(n-1) \times (n-1)$ matrix as follows.

Suppose we know how to compute $(n-1) \times (n-1)$ determinants. For any i and j, write m_{ij} for the determinant of the matrix you get by deleting the i'th row and j'th column of M. We then define

$$\det M = a_{11} m_{11} - a_{12} m_{12} + a_{13} m_{13} - \cdots \pm a_{1n} m_{1n},$$

with plus and minus signs alternating every term. (The sign of the last term therefore depends on whether n is even or odd.)

For example, in the 3×3 case, we have

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}(a_{22}a_{33}-a_{23}a_{32})-a_{12}(a_{21}a_{33}-a_{23}a_{31})+a_{13}(a_{21}a_{32}-a_{22}a_{31}).$$

I also encourage you to verify that Definition 3.2 agrees with the formula for a 2×2 determinant we discussed up above.

As mentioned above, we won't actually care about the details of this formula. There is only one fact we'll actually need, which we'll set off as a separate proposition for later convenience:

Proposition 3.3. *The determinant of an $n\times n$ matrix is a polynomial in the entries of the matrix. Each term of the polynomial contains exactly one entry from each row and each column.*

And of course, the reason we care about this otherwise quite arbitrary-looking formula is that it accomplishes the task we described at the beginning of this discussion:

Theorem 3.4. *Consider a system of n linear equations in n variables, say*

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n &= 0 \\ a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n &= 0 \\ &\vdots \\ a_{n1}x_1+a_{n2}x_2+\cdots+a_{nn}x_n &= 0, \end{aligned}$$

and write

$$M=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

for the corresponding matrix. Then the system has a nontrivial solution if and only if $\det M=0$.

3.2 Factoring Polynomials

The next set of facts we'll need concern factoring polynomials. Since many of them (with the possible exception of Proposition 3.8) are probably familiar, we won't be proving any of them here. For now, we'll just state the facts we need.

We'll start with a couple of probably familiar facts about factoring one-variable polynomials over $\mathbb{C}$.

Lemma 3.5. *Let f be a complex polynomial in one variable and let α be a complex number. Then $f(\alpha)=0$ if and only if $(x-\alpha)$ divides f, that is, if there exists some polynomial $\widetilde{f}$ such that $f(x)=(x-\alpha)\widetilde{f}(x)$.*

Lemma 3.6. *If f is a nonzero complex polynomial in one variable with degree n, then we can write*

$$f(x)=c(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n)$$

for some complex numbers $c, \alpha_1, \ldots, \alpha_n$, where $c \neq 0$ and these numbers are unique up to reordering the α_i's. (Notice that the n here counting the number of α_i's is the same as the degree of f!) For any number α, we have $f(\alpha) = 0$ if and only if α is one of the α_i's.

The story gets a bit more complicated when we consider polynomials in more than one variable. In that setting, even over the complex numbers, it's possible for polynomials to fail to factor nontrivially even when their degree is bigger than 1. We therefore introduce the following definition:

Definition 3.7. If f and g are polynomials, we say that g **divides** f if $f = gh$ for some polynomial h. A polynomial f is called **irreducible** if the only polynomials that divide it are constants and constant multiples of f. In other words, whenever we have $f = gh$ for some polynomials g and h, one of g or h must be a constant.

I encourage you to check, for example, that the polynomial $xy - 1$ can't be written in the form $(ax + by + c)(dx + ey + f)$, making $xy - 1$ an irreducible polynomial of degree 2. Despite this complication, we still have a result similar to Lemma 3.6:

Proposition 3.8. *Any polynomial f can be factored into irreducibles, that is, we can write*

$$f = f_1 f_2 \cdots f_k$$

where each f_i is irreducible. Furthermore, the f_i's are uniquely determined by f (up to multiplying each one by a constant).

For such an intuitively believable statement, the proof of the uniqueness part of Proposition 3.8 is surprisingly involved, quite a bit more so than Lemmas 3.5 and 3.6. We will actually need the full power of this result only once, in the proof of Lemma 3.9 below. That lemma will itself only be used for multi-variable polynomials once in the next section, when we show in the proof of Lemma 4.3 that the resultant of two polynomials with no common factors is nonzero. Every other application of the results from this subsection only depends on the one-variable case, which is handled by the much simpler Lemmas 3.5 and 3.6.

3.3 The Resultant

In our review of determinants just now, we saw that it there was an expression we could write down in terms of the coefficients of the linear equations in the system which is equal to zero if and only if all the equations have a common solution. Inspired by this, we are going to look for a way to tell whether two *polynomial* equations have a common solution by looking at the coefficients of the polynomials.

For polynomials in one variable over the complex numbers, the existence of a common *solution* is equivalent to the existance of a common *factor*. (This follows from the results about factoring polynomials we just stated; I encourage you to try to prove it.) But it will turn out in the next section that we'll need to apply the technology of resultants to polynomials in multiple variables, where these two concepts are *not* equivalent. For this reason, we will phrase all of our results in terms of common factors rather than common solutions so that they apply both to the one-variable complex case we are interested in here and to the cases we will care about later on.

Our plan will be to turn the question of the existence of a common factor into a question about solutions of a system of linear equations, which we already have a handle on thanks to our knowledge of the determinant. The key to making this work will be the following result.

Lemma 3.9. *Suppose f and g are nonzero complex polynomials, possibly with several variables, with* $\deg(f) = r$ *and* $\deg(g) = s$*. Then f and g have a common factor if and only if there exist polynomials A and B such that:*

- *A and B are not both the zero polynomial,*
- $\deg(A) < s$ *and* $\deg(B) < r$*, and*
- $Af + Bg = 0$.

Proof. First, suppose f and g have a common factor, say $f(x) = h(x)\widetilde{f}(x)$ and $g(x) = h(x)\widetilde{g}(x)$ for some polynomials $\widetilde{f}$ and $\widetilde{g}$, which are nonzero because f and g are. But then

$$\widetilde{g}f + (-\widetilde{f})g = \widetilde{g}h\widetilde{f} - \widetilde{f}h\widetilde{g} = 0,$$

so we can take $A = \widetilde{g}$ and $B = -\widetilde{f}$. (I encourage you to verify that the condition on the degrees is satisfied.)

Conversely, suppose there exist A and B satisfying the conditions listed above. First, note that while we have only assumed that A and B aren't both zero, in fact neither can be, because if (say) $B = 0$ and $A \neq 0$, then we would have $Af = Af + Bg = 0$, and (as you'll show in Exercise 3.3) this would mean that $f = 0$, which is a contradiction.

Using Proposition 3.8, write $f = f_1 \cdots f_k$, $g = g_1 \cdots g_l$, and $B = b_1 \cdots b_m$. Because the factorizations from Proposition 3.8 are unique, and $Af = -Bg$, each of the f_i factors from the left side of this equation must appear on the right, either as one of the g_i's or one of the b_i's. But, since we've assumed that $\deg(B) < r = \deg(f)$, it's not possible for them to all be accounted for by the b_i's — the sum of the degrees of the b_i's is $\deg(B)$, which is smaller than the sum of the degrees of the f_i's. So in fact at least one f_i must be equal to some g_j, that is, f and g have a common factor. □

As promised, we can turn the question about the existence of A and B into a system of linear equations. For now, we'll only worry about the one-variable case. Suppose we have

$$f(x) = f_r x^r + \cdots + f_1 x + f_0$$
$$g(x) = g_s x^s + \cdots + g_1 x + g_0.$$

Let's also write

$$A(x) = a_{s-1} x^{s-1} + \cdots + a_1 x + a_0$$
$$B(x) = b_{r-1} x^{r-1} + \cdots + b_1 x + b_0,$$

where we treat the a_i's and b_i's as unknowns; the fact that we go up to x^{s-1} and x^{r-1} here reflects the assumptions on the degrees of A and B from the lemma statement.

Now, $Af + Bg$ will be a polynomial of degree at most $r + s - 1$, so it can be specified by a list of $r + s$ coefficients, going from the coefficient of x^{r+s-1} down to the constant term, and of course $Af + Bg = 0$ if and only if all $r + s$ of these coefficients are equal to zero. And the key realization here is that looking at each of these coefficients gives a *linear* equation in the a_i's and b_i's! In other words, the question of whether there exist polynomials A and B of the right degrees such that $Af + Bg = 0$ is equivalent to the question of whether there exist *numbers* $a_0, a_1, \ldots, a_{s-1}, b_0, b_1, \ldots, b_{r-1}$ satisfying a system of linear equations.

The form these linear equations take will probably be clearer if we first look at a small example. Suppose $r = 2$ and $s = 3$. Then we have five unknowns: a_0, a_1, a_2, b_0, and b_1. I encourage you to check that, in order for $Af + Bg$ to equal 0, the following five equations need to be satisfied:

$$\begin{aligned}
f_0 a_0 + \qquad\qquad & g_0 b_0 \qquad\qquad = 0 \\
f_1 a_0 + f_0 a_1 + \qquad & g_1 b_0 + g_0 b_1 = 0 \\
f_2 a_0 + f_1 a_1 + f_0 a_2 + & g_2 b_0 + g_1 b_1 = 0 \\
f_2 a_1 + f_1 a_2 + & g_3 b_0 + g_2 b_1 = 0 \\
f_2 a_2 + \qquad & g_3 b_1 = 0
\end{aligned}$$

Since this gives five linear equations in five unknowns, we see that there exists a nontrivial solution of this system of equations — and therefore an A and a B with $Af + Bg = 0$ which aren't both zero, and therefore a common factor of f and g — if and only if

$$\det \begin{pmatrix}
f_0 & & & g_0 & \\
f_1 & f_0 & & g_1 & g_0 \\
f_2 & f_1 & f_0 & g_2 & g_1 \\
 & f_2 & f_1 & g_3 & g_2 \\
 & & f_2 & & g_3
\end{pmatrix} = 0.$$

(Here we're following the common convention that empty entries in a matrix are zero.)

This brings us, finally, to the definition we've been building up to.

Definition 3.10. Let f and g be complex polynomials, with $\deg(f) = r$ and $\deg(g) = s$, and say

$$\begin{aligned}
f(x) &= f_r x^r + \cdots + f_1 x + f_0 \\
g(x) &= g_s x^s + \cdots + g_1 x + g_0.
\end{aligned}$$

The **Sylvester matrix** of f and g is the $(r+s) \times (r+s)$ matrix

$$\begin{pmatrix}
f_0 & & & & g_0 & & & \\
f_1 & f_0 & & & g_1 & g_0 & & \\
\vdots & \vdots & \ddots & & \vdots & \vdots & \ddots & \\
f_r & f_{r-1} & \ddots & f_0 & g_s & g_{s-1} & \ddots & g_0 \\
 & f_r & & \vdots & & g_s & & \vdots \\
 & & \ddots & & & & \ddots & \\
 & & & f_r & & & & g_s
\end{pmatrix}.$$

That is, the first s columns each contain the coefficients of f in order, starting in the j'th row in column j, and the last r columns each contain the coefficients of g arranged in a similar way. (While this picture of the matrix makes it look like the final f_0 and the final g_0 start on the same row as each other, this doesn't necessarily have to be true, as you can see in the example from earlier.)

The determinant of the Sylvester matrix is called the **resultant** of f and g, and it's written $\operatorname{Res}(f, g)$.

The resultant is the object we'll be studying for the rest of this class, and all of the discussion leading up to it shows us that it accomplishes our goal from the beginning of this section: it

gives us a formula in terms of the coefficients of our two polynomials that can detect whether the two polynomials have a common factor. Everything we've learned in this section can be summarized as follows:

Theorem 3.11. *If f and g are nonzero polynomials with* $\deg(f) = r$ *and* $\deg(g) = s$*, then f and g have a common factor if and only if* $\operatorname{Res}(f,g) = 0$.

Let's see this in action in a quick example. Consider the polynomials $x^2-(b+1)x+b$ and $x-2$. Since the first polynomial factors as $(x-b)(x-1)$, it's easy to see without using resultants that these polynomials have a common factor if and only if $b=2$. But let's try to reproduce this result using Theorem 3.11.

The Sylvester matrix in this case will be

$$\begin{pmatrix} b & -2 & 0 \\ -(b+1) & 1 & -2 \\ 1 & 0 & 1 \end{pmatrix},$$

which I encourage you to verify. We can compute the determinant of this matrix using the formula we gave earlier, which gives a resultant of $b+2(-(b+1)+2) = 2-b$. So, indeed, the resultant is zero if and only if $b=2$, exactly as we expected from Theorem 3.11.

This fact will be the main tool we'll use to tackle the proof of Bézout's Theorem. If we are working with one-variable polynomials, then our earlier discussion implies that we can conclude that f and g have a common *root* if and only if their resultant is zero. We will discuss how to apply this result to multiple-variable polynomials in the next section.

Exercises

Exercises that are especially important for the rest of the class have been marked with ⚠.

3.1. Consider the polynomials x^2-1 and $(x-a)^2-1$, where a is a constant. Without using resultants, for which values of a will these two polynomials have a common root? Now, compute the resultant of these two polynomials and show that you get the same answer. *[Hint: The resultant you get should be a^4-4a^2.]*

3.2. Consider the polynomials x^2+y^2-1 and $x-a$.

(a) Treat y and a as constants and compute the resultant of these two polynomials. (Your answer should be a polynomial in y and a.)

(b) Interpret this result as the answer to the following question: where does the unit circle intersect the vertical line $x=a$? *[Hint: In terms of the picture of the circle and the line, what does it mean when you change the value of a? What does it mean when you change the value of y?]*

(c) This question is of course not hard to answer without using resultants. Does the answer you got in the previous part make sense? If we are looking for *real* points, the existence of an intersection point should depend on the value of a. How is that reflected in your answer?

3.3. ⚠

(a) Suppose f and g are both complex polynomials. Prove that fg is the zero polynomial if and only if either f or g is the zero polynomial.

(b) Suppose f, g, and h are polynomials and f is not the zero polynomial. Use the previous part to prove that, if $fg = fh$, then $g = h$.

3.4. Verify that the form of the matrix appearing in Definition 3.10 is correct, that is, that it indeed corresponds to the system of linear equations you get by looking at the coefficients of $Af + Bg$.

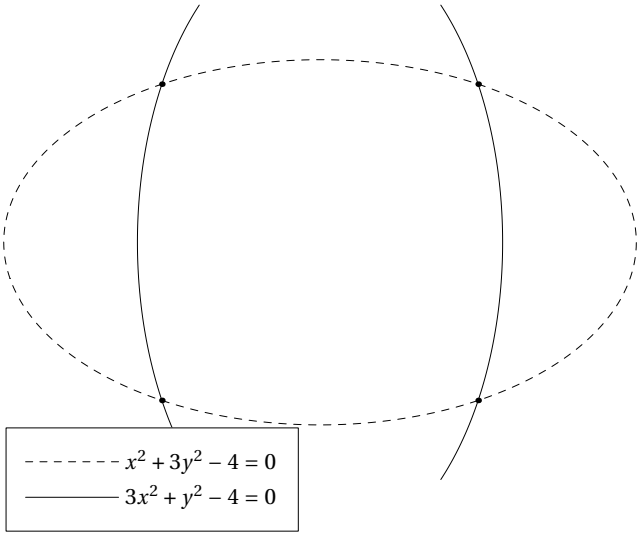
3.5. *[Only if you know a bit of calculus.]*

(a) Let f be a one-variable complex polynomial. Prove that $\mathrm{Res}(f, f') = 0$ if and only if f has a multiple root, where f' is the derivative of f. *[Hint: Factor f in the form $c(x-a_1)\cdots(x-a_n)$ and use the product rule to compute the derivative.]*

(b) The resultant of f and f' is called the **discriminant** of f. Compute the discriminant of $f(x) = a\,x^2 + b\,x + c$. The expression you get should be familiar, especially if you factor out $-a$. Explain why you ended up with it.

4 Bézout's Theorem

With the resultant in hand, we now have all the tools in place to prove Bézout's Theorem. To see how the resultant will help us, it will be helpful to look at an example. So consider the two ellipses cut out by the equations $x^2+3y^2-4=0$ and $3x^2+y^2-4=0$, drawn below.



Both from the picture and from just plugging values into the equations directly, we can see that $(1,1)$, $(1,-1)$, $(-1,1)$ and $(-1,-1)$ are all intersection points of these two curves. I encourage you to try to show directly that these are in fact the *only* intersection points, even over the complex numbers.

Because we want to also consider intersection points in projective space, we should homogenize these two polynomials, to produce $f(x,y,z)=x^2+3y^2-4z^2$ and $g(x,y,z)=3x^2+y^2-4z^2$. I encourage you to also check that, in this particular example, we don't get any new intersection points at infinity. (This amounts to plugging in $z=0$ to restrict to the line at infinity, and then verifying that the only solution of the resulting equations is $x=0$, $y=0$; since $[0:0:0]$ isn't a point of projective space, this means there are no additional intersection points.)

So the four points we've identified are in fact it. How could we have used resultants to pick them out? The trick is to think of f and g as polynomials just in the variable x, with *coefficients* which are polynomials in y and z. From this perspective, for example, f has two nonzero terms: an x^2 term with a coefficient of 1, and the constant term $3y^2-4z^2$. The resultant will then be the determinant of a matrix whose entries are polynomials in y and z and so (thanks to Proposition 3.3) will itself be a polynomial in y and z.

When we employ this trick — treating multivariable polynomials like polynomials in one of the variables and taking the resultant — we'll stick a subscript on the Res in the notation to indicate which variable we're singling out. I encourage you to go back through our discussion of resultants and convince yourself that everything still goes through if the coefficients are polynomials rather than numbers.

In our case, the resultant is

$$\begin{aligned}\mathrm{Res}_x(f,g) &= \det\begin{pmatrix} 3y^2-4z^2 & 0 & y^2-4z^2 & 0 \\ 0 & 3y^2-4z^2 & 0 & y^2-4z^2 \\ 1 & 0 & 3 & 0 \\ 0 & 1 & 0 & 3 \end{pmatrix} \\ &= 64(y^2-z^2)^2 \\ &= 64(y-z)^2(y+z)^2.\end{aligned}$$

What does this tell us about intersection points? Imagine we plug in specific values of y and z, say $y=0$ and $z=1$. This has the effect of restricting to points of the form $[x:0:1]$, that is, points on the x-axis. After plugging in these values, our resultant is 64. This is not 0, so Theorem 3.11 tells us there is no point of the form $[x:0:1]$ where f and g are both equal to 0. In other words, there is no intersection point on the x-axis.

It's easy to see that the resultant is 0 if and only if $y=z$ or $y=-z$. Each of these equations cuts out a line in the projective plane; in fact, if we restrict to the ordinary plane by plugging in $z=1$, we see that they become the horizontal lines $y=1$ and $y=-1$. But, by Theorem 3.11, the resultant is 0 if and only if there is some α such that the two original polynomials are equal to zero when we plug in $x=\alpha$.

So *the intersection points of our two curves occur exactly on the lines $y=z$ and $y=-z$*. From here, if we wanted to find the x coordinates of the intersection points, it would just be a matter of plugging each of these equations back into the original equations for the curves and solving for x, which I encourage you to do.

The task we've set ourselves, though, is *counting* the number of intersection points. Notice that in this example, our resultant was a homogeneous polynomial in y and z of degree 4. It's an encouraging sign that 4 is also the number of intersection points we expect to find between two curves of degree 2. But the method we used here doesn't quite count the number of intersection points; it counts the number of *horizontal lines* that *contain* intersection points. In this case, we got only two distinct horizontal lines.

Since we're looking to count intersection points with multiplicity, you might be encouraged by the fact that the $z-y$ and $z+y$ factors in the resultant both appeared squared; maybe the 2 is the multiplicity? But sadly, this can't be right: as we saw when we counted the points by hand, there are actually four *distinct* intersection points, two on each of these horizontal lines, and so if Bézout's Theorem is going to be true, these points all have to have multiplicity 1.

4.1 Changing Coordinates

So we want to count intersection points, but we have a tool that lets us count horizontal lines containing intersection points. This means that we'll be in good shape as long as each horizontal line can only contain one intersection point. After all, that way, the two counts will be the same. If this property fails to hold, our strategy will simply be to change our coordinate system to one where it does hold.

You'll explore the computational details in depth in the exercises, but in order for that to be possible, we need to precisely describe which types of coordinate changes we'll be using.

Definition 4.1. A **linear change of coordinates** on the projective plane is a transformation of the form

$$x' = a_1x + b_1y + c_1z; \quad y' = a_2x + b_2y + c_2z; \quad z' = a_3x + b_3y + c_3z.$$

which is *invertible*, that is, for which it is always possible to write x, y, and z in terms of x', y', and z'.

Though we won't prove this here, such a coordinate change is invertible if and only if

$$\det\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \neq 0.$$

Given such a coordinate change, the invertibility requirement means that it will always be possible to take any homogeneous polynomial $f(x,y,z)$ and write it as a function of x', y', and z'; I encourage you to check that this results in a polynomial that is still homogeneous and has the same degree as f.

In addition to the possibility that two intersection points lie on the same horizontal line, there is also one more thing the intersection points could do to ruin our plan. Our strategy involves identifying which horizontal lines contain an intersection point, so our count will also be off if there is an intersection point that lies on *every* horizontal line. If we were working in the ordinary plane, this would not be an issue, since horizontal lines are parallel. But we're working in the *projective* plane, where there is a point that's on every horizontal line: the point $[1:0:0]$.

We therefore need to impose one more condition on our coordinate system. We can summarize everything we need as follows:

Definition 4.2. Suppose $f(x,y,z)$ and $g(x,y,z)$ are homogeneous polynomials with no common factors. Suppose that, after performing a linear change of coordinates, we have that no two intersection points of f and g lie on the same horizontal line, and that $f(1,0,0) \neq 0$ and $g(1,0,0) \neq 0$. We will then say that we are working in a coordinate system that is **compatible** with f and g.

(In order for $[1:0:0]$ to not be an intersection point, it's enough if one of f or g doesn't vanish there. We are imposing the stronger condition that *neither* vanishes there just for later convenience; it would be possible, but more annoying, to make the proof work with the weaker condition.) You'll show in Exercise 4.3 that, as long as f and g have no common factors, it is always possible to find a coordinate system that is compatible with f and g. For now, we'll move forward assuming that this has been accomplished.

4.2 Multiplicities and the Proof

In our example at the beginning of this section, we saw that, when we regarded our two polynomials as polynomials just in the variable x and computed their resultant, we got a homogeneous polynomial in y and z which split into linear factors, and each factor gave us one of the horizontal lines that contained an intersection point. As the following two lemmas show, this will happen in general.

Lemma 4.3. *Suppose $f(x,y,z)$ and $g(x,y,z)$ are homogeneous polynomials with no common factors, with $\deg(f) = d$ and $\deg(g) = e$, and suppose $f(1,0,0) \neq 0$ and $g(1,0,0) \neq 0$. Then $\mathrm{Res}_x(f,g)$ is a nonzero homogeneous polynomial of degree de in y and z.*

Proof. We'll start by writing f and g as polynomials in x with coefficients which are polynomials in y and z. Let's say

$$f(x) = f_d x^d + \cdots + f_1 x + f_0$$
$$g(x) = g_e x^e + \cdots + g_1 x + g_0.$$

Since $f(1,0,0) \neq 0$, we also know that $f_d \neq 0$, that is, f has degree exactly d when thought of as a polynomial in x; similarly, g has degree e as a polynomial in x. This, along with the assumption that f and g have no common factors, is enough to satisfy the hypotheses of Theorem 3.11 and allow us to conclude that $\text{Res}_x(f,g) \neq 0$. [1]

By Proposition 3.3, the determinant of the Sylvester matrix is a polynomial in matrix's entries. Since each of these entries is a polynomial in y and z, the resultant is also a polynomial in y and z. It now just remains to show that it is homogeneous of degree de. You'll do this in Exercise 4.5. □

Recall from Exercise 2.2d that any homogeneous polynomial $f(y,z)$ in two variables can be factored completely into linear homogeneous polynomials. That is, it can be written in the form

$$f(y,z) = (a_1 y - b_1 z)(a_2 y - b_2 z)\cdots(a_d y - b_d z),$$

where $d = \deg(f)$.

If we are working in a compatible coordinate system and $[a:b:c]$ is an intersection point of f and g, then $cy - bz = 0$ is one of the horizontal lines containing an intersection point, so it must be one of the linear factors that appears when we apply this result to $\text{Res}_x(f,g)$. (At least, some constant multiple of it is — you can always factor out a constant and multiply it into one of the other factors, but because of unique factorization this is the only change you can make.) Unlike in the example we started this section with, after picking a compatible coordinate system we now know that $[a:b:c]$ will be the *only* intersection point on the line $cy - bz = 0$. So this will, finally, give us our definition of intersection multiplicity: the multiplicity will just be the number of times this factor appears.

More precisely:

Definition 4.4. Suppose $f(x,y,z)$ and $g(x,y,z)$ are homogeneous polynomials with no common components, and that we are working in a coordinate system compatible with f and g. If $p = [a:b:c]$ is an intersection point of f and g, then the **multiplicity** of p in the intersection of f and g is the largest number m such that $\text{Res}_x(f,g)$ is divisible by $(cy - bz)^m$. If p is not an intersection point, we'll define the multiplicity to be 0.

Notice that this definition of multiplicity only depends on the y and z coordinates of the point, that is, it only depends on the horizontal line the point lies on. So it is really only a sensible definition *after* we have imposed the restriction that no two intersection points lie on the same horizontal line.

With all of these results in place, the proof of Bézout's Theorem falls out pretty quickly:

Theorem 4.5 (Bézout's Theorem, True Version)**.** *Suppose $f(x,y,z)$ and $g(x,y,z)$ are homogeneous polynomials with no common factors, with* $\deg(f) = d$ *and* $\deg(g) = e$*. Then, in any coordinate system compatible with f and g, the algebraic curves cut out by f and g intersect in exactly de points, counted with multiplicity.*

[1] This conclusion is not quite as simple as it might seem, for a couple of reasons.

First, we're applying Theorem 3.11 to polynomials whose coefficients are *polynomials* rather than numbers — the theorem tells us that, if the resultant is zero, then f and g have a common factor, which in this context would be an actual, three-variable polynomial that divides both f and g, which we have assumed doesn't exist. If this step is unclear, make sure to go back through the discussion leading up to Theorem 3.11 and convince yourself it's still true if the coefficients are polynomials rather than numbers!

In particular, when we proved Theorem 3.11, we relied on Theorem 3.4, which said that a system of linear equations has a nontrivial solution if and only if the determinant of the corresponding matrix is 0. In order for Theorem 3.11 to still be true in our present case, we need Theorem 3.4 to still be true when both the entries of the matrix and the entries of the solution we're looking for are polynomials. This is in fact true, although we won't prove it here.

Proof. Consider the resultant $\mathrm{Res}_x(f,g)$. By Lemma 4.3, this is a nonzero homogeneous polynomial of degree $d\,e$; let's write it $R(y,z)$. By Exercise 2.2d, it factors in the form

$$R(y,z) = (a_1 y - b_1 z)(a_2 y - b_2 z)\cdots(a_{de} y - b_{de} z),$$

where each of these factors is nonzero.

Suppose $[a:b:c]$ is an intersection point of f and g. This means that, if we plug $y=b, z=c$ into f and g, the resulting polynomials in x have a common root at $x=a$. So by Theorem 3.11, we must have that $R(b,c)=0$. Because all the factors of R are nonzero, this can only happen if one of the factors of R is $(c\,y - b\,z)$, up to a constant multiple. The number of times this factor appears is, by Definition 4.4, the intersection multiplicity of our point $[a:b:c]$.

Conversely, if $(t\,y - s\,z)$ is one of the linear factors of R, we know that $R(s,t)=0$, and so by Theorem 3.11 once again we know that $f(x,s,t)$ and $g(x,s,t)$ have a common root, say at $x=r$, which means that $[r:s:t]$ is an intersection point.

So, to each intersection point, we have associated m of the linear factors of R, where m is that point's intersection multiplicity, and we have shown that every linear factor of R is accounted for in this way. Our count could still be off, though, if two different intersection points were associated to the same linear factor of R. Suppose two distinct points $[a:b:c]$ and $[a':b':c']$ were assigned the same linear factor. This would mean (possibly after multiplying by a constant) that $(c\,y - b\,z)$ and $(c'y - b'z)$ were the same polynomial, that is, we would have $b'=b$ and $c'=c$. But this would place our two intersection points on the same horizontal line, which we have assumed can't happen.

This completes the proof: there are $d\,e$ linear factors; each intersection point accounts for m of them, where m is its multiplicity; and every intersection point is included in this count. So the sum of the multiplicities must be exactly $d\,e$. □

At long last, we've finished the goal we set out in the introduction: after modifying the statement to account for all the problems we identified with the first version, we have succeeded in turning Bézout's Theorem into something we can prove,

But there is one potentially unsatisfying aspect of our setup that has to do with our definition of intersection multiplicities. We defined the intersection multiplicity of f and g at a point p by first picking a compatible coordinate system and then computing $\mathrm{Res}_x(f,g)$ in that coordinate system. How do we know that, if we had picked a *different* (but still compatible) coordinate system, we would have gotten the same number out as our multiplicity?

This is, in fact, a big weakness of the resultant-based approach to defining intersection multiplicities, and if you ever go on to study this topic in an algebraic geometry class you will probably use a definition of intersection multiplicity that doesn't have this problem of seeming to depend on the choice of coordinates. Still, it is possible to solve it, and this will be tackled in the optional Section A.

Exercises

Exercises that are especially important for the rest of the class have been marked with ⚠.

4.1. In Exercise 2.1 you found the intersection points of the parabola $y - x^2 = 0$ and the vertical line $x - a = 0$. Using the techniques from this section, find the intersection points again and compute their multiplicities.

4.2. Consider the circle $x^2+y^2-y=0$ and the ellipse $x^2-xy+y^2-y=0$. Use the techniques from this section to find all the intersection points of these two curves in the projective plane and their multiplicities. *[Hint: The resultant you get after homogenizing should be y^4-y^3z. If you don't feel like computing a 4×4 determinant, feel free to just assume this and solve the problem from there.]*

4.3. ⚠ Let $f(x,y,z)$ and $g(x,y,z)$ be non-constant homogeneous polynomials with no common factors. In this problem, we'll show that there is a coordinate system compatible with f and g.

(a) Show that, for any point p in the projective plane and any line L containing p, there is a linear change of coordinates that takes p to $[1:0:0]$ and L to the line $y=0$. *[Hint: First show that, for any point q on L other than p, there is a change of coordinates taking $[1:0:0]$ to p and $[0:0:1]$ to q. Then show that the inverse of this change of coordinates works.]*

(b) Using a similar argument, show that there is a coordinate change we can perform after which neither curve contains $[1:0:0]$ or $[0:1:0]$. *[Hint: Exercise 1.4 might be helpful.]*

(c) Suppose we have performed the coordinate change from part (b). Prove that there are only finitely many horizontal lines and finitely many vertical lines which contain intersection points of f and g. Conclude that there are finitely many intersection points. (A horizontal line is a line that contains $[1:0:0]$, and a vertical line is a line that contains $[0:1:0]$.) *[Note: You can't use Bézout's Theorem for this, since the fact we're proving in this problem is part of our proof of Bézout's Theorem! Instead use some of the facts we proved about resultants in this section.]*

(d) Prove that there exists a line L with the following two properties: (i) no line parallel to L contains two distinct intersection points of f and g, and (ii) L is not completely contained in the union of our two curves. *[Hint: You might find some of the exercises from Section 1 helpful.]*

(e) Prove that there is a point p on L which is not on either of our two curves, and that if you apply the coordinate change from part (a) to this choice of p and L the resulting coordinate system is compatible with f and g.

4.4. In our definition of compatible coordinate systems, we asked that $f(1,0,0)$ and $g(1,0,0)$ be nonzero. We ended up using this in the proof of Lemma 4.3 in order to satisfy one of the hypotheses of Theorem 3.11.

It would in fact be possible to do this whole proof just assuming *one* of f or g is nonzero at $[1:0:0]$. But what goes wrong if we both f and g vanish at $[1:0:0]$? Specifically, what do the Sylvester matrix and the resultant look like if $f(1,0,0)=g(1,0,0)=0$, and how does this line up with the answer to the question about what horizontal lines contain intersection points?

4.5. ⚠ In this exercise, we'll verify the degree count in Lemma 4.3. As in the lemma, we will start with two homogeneous polynomials $f(x,y,z)$ and $g(x,y,z)$, with $\deg(f)=d$ and $\deg(g)=e$. We are thinking of f and g as polynomials in x with coefficients that are polynomials in y and z and taking the resultant. Again as in the lemma, let's write

$$f(x)=f_d x^d+\cdots+f_1 x+f_0$$
$$g(x)=g_e x^e+\cdots+g_1 x+g_0.$$

(a) Argue that each f_i is homoegeneous of degree $d-i$ and each g_j is homogeneous of degree $e-j$.

(b) Write c_{ij} for the entry in the Sylvester matrix on row i and column j. Show that, if $c_{ij} \neq 0$, then it's a homogeneous polynomial of degree $d+j-i$ if $j \leq e$ and degree $j-i$ if $j > e$.

(c) Recall from Proposition 3.3 that each term in the determinant of this matrix is a product of matrix entries, with exactly one from each row and each column. Let's focus on just one such term now. Let $\sigma(j)$ be the row of the entry we're choosing from column j. Prove that our term is a homogeneous polynomial of degree

$$\sum_{j=1}^{e}(d+j-\sigma(j)) + \sum_{j=e+1}^{e+d}(j-\sigma(j)).$$

(d) Finally, argue that this sum is always equal to de, and conclude that the entire resultant is therefore a homogeneous polynomial of degree de. *[Hint: It will be helpful to remember that, as j ranges from 1 to $d+e$, $\sigma(j)$ takes on each value from 1 to $d+e$ exactly once.]*

A Coordinate Changes and Multiplicities

When we finished our proof of Bézout's Theorem we were left with one unsatisfying detail: our definition of intersection multiplicity required us to pick a coordinate system compatible with our two curves, but we never established that the number we got out as the multiplicity was actually independent of this choice. In this section, we'll close this loophole.

Our strategy will be to describe intersection multiplicities in terms of a list of axioms that don't depend on the coordinate system. If we can show that these axioms uniquely specify the intersection multiplicity and that our definition satisfies them, we'll have our result. As a bonus, in addition to giving us our coordinate-independence, this will also provide a nice way to *compute* intersection multiplicities without having to take determinants of gigantic matrices.

Definition A.1. Let $m_p(f,g)$ be a function which takes two homogeneous polynomials $f(x,y,z)$ and $g(x,y,z)$ with no common factors, along with a point p in the projective plane, and returns a nonnegative integer. We will say that m **satisfies the intersection multiplicity axioms** if it has all of the following properties:

(i) $m_p(f,g) = m_p(g,f)$

(ii) $m_p(f,g) = 0$ if and only p isn't an intersection point of f and g

(iii) if f and g are lines and p is their intersection point, then $m_p(f,g) = 1$

(iv) $m_p(f_1 f_2, g) = m_p(f_1, g) + m_p(f_2, g)$

(v) if $\deg(f) \leq \deg(g)$ and h is another homogeneous polynomial with $\deg(h) = \deg(g) - \deg(f)$, then $m_p(f, g - fh) = m_p(f,g)$.

Theorem A.2. *The intersection multiplicity axioms uniquely determine the function m, and the intersection multiplicity defined in Definition 4.4 satisfies the intersection multiplicity axioms.*

Proof. We will start by showing that the axioms uniquely determine m. Given f, g, and p, our strategy will be to repeatedly use (iv) and (v) to reduce the degrees of f and g until we can apply (ii) or (iii) to compute the multiplicity.

Since all of our axioms are indepedent of our choice of coordinates, we can change coordinates so that $p = [0:0:1]$.

Suppose we are given f, g, and p. Our proof will proceed by induction on a somewhat odd quantity: the sum of the highest power of y appearing in f and the highest power of y appearing in g. (For example, the highest power of y appearing in $x^4 + 3xyz^2 + 2xy^2z$ is 2.) In the base case, there are no y's in either f or g, so they are both homogeneous polynomials in just x and z. By Exercise 2.2d, this means they can both be written as a product of linear homogeneous polynomials, and I encourage you to verify that our axioms are enough to handle that case.

Now, suppose we *do* have some y's in f or g, but that we know that our axioms uniquely determine m whenever our quantity is strictly smaller than it is for f and g. Our first task will be to repeatedly apply (v) to cancel terms until one of the polynomials is a multiple of y, after which we will be able to use (iv) to lower the power of y in that polynomial.

A homogeneous polynomial will fail to be a multiple of y if it has a term of the form $c\,x^a z^b$. Let r be the biggest power of x in a term of this form in f, and let s be the biggest power of x in a term of this form in g. After multiplying f and g by constants, we can assume that the

coefficients of both of these terms are 1. And, after possibly using (i) to switch f and g, we can assume $r \leq s$.

Suppose $r > 0$. We are then aiming to build a polynomial out of f and g in which the smallest power of x appearing in a term of the form $x^a z^b$ is strictly smaller than s. I encourage you to convince yourself that

$$h = z^{\deg(f)+s-r} g - x^{s-r} z^{\deg(g)} f$$

does the job. By (v), we have that $m_p(f,h) = m_p(f, z^{\deg(f)+s-r} g)$, and by (iv) and (ii), this is equal to $m_p(f,g)$ (since z doesn't vanish at $p = [0:0:1]$). So, if we replace g with h, we have decreased s without changing the intersection multiplicity.

We can apply this procedure over and over again, lowering r or s each time, until one of them is 0. Note also that we never increase the power of y appearing in either polynomial, so our induction is safe.

Suppose now that we have done this and that $r = 0$. Then the only term in f that might not contain a y is $z^{\deg(f)}$. But we also know that $f(0,0,1) = 0$, which means that in fact the coefficient on $z^{\deg(f)}$ has to be zero. So, in this case, f is a multiple of y, say $f = y\widetilde{f}$. By (iv), we then know that $m_p(f,g) = m_p(\widetilde{f},g) + m_p(y,g)$. Our induction hypothesis handles the first term, so we just need to take care of the second.

To do this, write g in the form

$$g(x,y,z) = g_1(x,z) + y\, g_2(x,y,z).$$

That is, split off all the terms of g containing a y and call that $y\,g_2$. By (v) once again, we know that $m_p(y,g) = m_p(y,g_1)$. But now g_1 is a homogeneous polynomial in x and z, and so by Exercise 2.2d it's a product of *linear* homogeneous polynomials. Applying (iv) to this tells us that $m_p(y,g_1)$ is a sum of terms of the form $m_p(y, tx+uz)$, and by (iii), these are all equal to 1.

This completes the proof of uniqueness. We now have to show that the intersection multiplicity as we defined it in the last section satisfies our axioms. This will be handled in the exercises. □

Exercises

These exercises will be all about establishing that our resultant-based definition of intersection multiplicities from Definition 4.4 satisfies the axioms in this section. The axioms will all turn out to correspond to some basic facts about resultants which we could have proved shortly after defining them. These facts will, in turn, depend on knowing a few facts about how to compute determinants that didn't come up anywhere else in our discussion; I will indicate what these are when we need them.

Throughout this discussion, unless stated otherwise, f and g will be homogeneous polynomials in the variables x, y, and z, and we are working in a coordinate system which is compatible with f and g.

A.1. Prove that $\mathrm{Res}_x(f,g) = \pm\mathrm{Res}_x(g,f)$, and conclude that (i) is satisfied. *[Hint: Use the fact that interchanging two columns of a matrix multiplies its determinant by* -1*.]*

A.2. Definition 4.4 already states directly that if p isn't an intersection point of f and g then the multiplicity is 0. Prove the other direction of (ii) by verifying that the multiplicity is positive if p is an intersection point.

A.3. Prove (iii) by directly computing the determinant of the Sylvester matrix in this case.

A.4. (a) Suppose that f and g are one-variable (non-homogeneous) complex polynomials, and that we have factored them as

$$f(x) = c(x-\alpha_1)\cdots(x-\alpha_m); \qquad g(x) = c'(x-\beta_1)\cdots(x-\beta_n).$$

Prove that, for some constant K,

$$\mathrm{Res}(f,g) = K\prod_{i=1}^{m}\prod_{j=1}^{n}(\alpha_i-\beta_j).$$

[Hint: Think of f and g as homogeneous polynomials in $x,\alpha_1,\ldots,\alpha_m,\beta_1,\ldots,\beta_n$, apply Lemma 4.3, and observe what happens when you plug in $\alpha_i=\beta_j$ for some i,j.]

(b) Conclude that, if f, g, and h are one-variable polynomials,

$$\mathrm{Res}(f,gh) = L\,\mathrm{Res}(f,g)\mathrm{Res}(f,h)$$

for some constant L. (This constant is in fact 1, but this is a bit harder to prove.)

(c) Argue that this must then also be true if f, g, and h are multi-variable polynomials and we replace Res with Res_x.

(d) Conclude that (iv) is satisfied.

A.5. Prove that, if $\deg(f)\le\deg(g)$ and h is a homogeneous polynomial with $\deg(h)=\deg(g)-\deg(f)$, then $\mathrm{Res}_x(f,g-fh)=\mathrm{Res}_x(f,g)$ and conclude that (v) is satisfied.[2] *[Hint: Use the fact that adding a multiple of one column of a matrix to another column doesn't change the determinant.]*

[2]There is a subtlety here which the wording of this question glosses over: even if our coordinate system is compatible with f and g, it doesn't necessarily follow that it's compatible with f and $g-fh$. Resolving this requires showing that everything we've done only requires *one* of f or g to not contain $[1:0:0]$. You can, if you'd like, treat resolving this issue as an optional extra exercise. It will require modifying Lemma 3.9 to allow one (but not both) of f or g to have smaller than r or s respectively.