On Bézout's Theorem

Dathan Ault-McCoy

Abstract

Algebraic geometry is concerned with the study of the properties of certain geometric objects (particularly solution sets of systems of polynomial equations) using the methods and results of abstract algebra. One of the earliest results to this end (known in some form as far back as Newton's time) is Bézout's theorem, which relates the number of points at which two polynomial curves intersect to the degrees of the generating polynomials. The goal of this paper is to reproduce the somewhat novel proof of the theorem given in R. P. Hulst's bachelor's thesis [4] (which is itself an elaboration of [3]), and in doing so get a glimpse of a few key concepts that lie at the heart of algebraic geometry, including projective space and intersection multiplicity.

Contents

Introduction

Geometry is one of the oldest fields of mathematics, and although the geometry of Euclid is little more than a curiosity today, the study of shapes has, in various forms, remained central to much of mathematics. Modern mathematical methods have produced a plethora of ways of understanding shapes, many of which stray far from the classic domain of circles and polygons. One common place is which questions of geometry arise from other pursuits is when considering the solution sets of some equation or equations. Level curves and conic sections both offer examples of this. Such solution sets or locii often define objects with decidedly geometric properties that can be of interest in their own right. Defining shapes in this way offers the possibility of relating the geometric properties of the solution sets to the numerical or algebraic properties of the equations. In the case of the conic sections, for instance, it is possible to determine geometric parameters such

as focii and radii directly from the coefficients of the polynomials. The field of algebraic geometry explores this connection between geometry and algebra by studying the zero locii of polynomials (of arbitrary degree and in arbitrarily many variables) using methods from abstract algebra, which provides tools for systematically analyzing various forms of "arithmetic".

This paper is concerned with one particular result from algebraic geometry known as Bézout's theorem. It is one of the oldest results in the field, known in some form as early as the 17th century, and as such there are a number of different statements attached to the name. One classic version of the result states that the number of points at which two polynomial curves intersect in $\mathbb{R}^2$ is no greater than the product of the degrees of the generating polynomials. However, it turns out that with some modifications—assigning an appropriate multiplicity to intersections and considering curves in a larger "projective space" with points added at infinity—it is possible to make sense of this inequality as an exact equality. Much of the paper is spent developing the notions which will allow us to make these modifications precisely.

There is a great variety of proofs of Bézout's theorem, many of which rely on advanced machinery. In 2009, Jan Hilmar and Chris Smyth published an article [3] which presents a novel proof of the theorem using only a division algorithm and some basic properties of intersection multiplicity. Several years later, R. P. Hulst then greatly expanded on Hilmar and Smyth's proof in their bachelor's thesis [4], filling in many details which were excluded in the original article. This paper is intended to be a further refinement of this work. We provide a more complete and hopefully more accessible exposition of the fundamental concepts, and expand on some proofs while simplifying others as appropriate. In the first three sections of the paper we introduce the tools and machinery needed to state and prove the full version of Bézout's theorem, which we finally do in section 4.

This paper does not require any commutative algebra but it does assume a knowledge of general abstract algebra, particularly the basic theory of rings and fields. A brief crash course on the necessary materials is given in the appendix.

1 A First Taste

Before introducing the machinery necessary to state and prove Bézout's theorem in its more modern form, we will briefly consider some of the key concepts in a more familiar setting which will hopefully motivate our later work.

Definition 1. Given a real polynomial in two variables $f \in \mathbb{R}[x, y]$, we define the *affine zero set* or *affine zero locus* of f as

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

We use the adjective "affine" to distinguish from zero locii in *projective space* which will be the focus of the rest of this paper. The zero locus of a polynomial as defined above will typically form some kind of curve in the plane. For instance, $Z_A(y - x^2)$ is the familiar graph of the parabola, and more generally every conic section is the zero locus of some polynomial of degree 2.

Theorem 2. *Given $f, g \in \mathbb{R}[x, y]$, $Z_A(fg) = Z_A(f) \cup Z_A(g)$.*

Proof. The product fg vanishes at a point in $\mathbb{R}^2$ if and only if either f or g vanish at that point. □

As we said in the introduction, the classic version of Bézout's theorem tells us that the number of points at which $Z_A(f)$ and $Z_A(g)$ intersect is no greater than the product of the degrees of f and g. However, as stated, this is not exactly correct. A polynomial curve will often intersect itself, for instance, at infinitely many points. We clearly want to require that the two curves in question be distinct, then, but even this is not quite enough. Consider $f(x, y) = x^2 + y^2 - 1$ and $g(x, y) = (x^2 + y^2 - 1)(y - x)$. Clearly $Z_A(f)$ is the unit circle, and it can be seen (either from Theorem 2 or by direct computation with a graphing calculator) that $Z_A(g)$ is the union of the unit circle and a line passing through the origin. Then $Z_A(f) \cap Z_A(g)$ is just the unit circle, which has infinitely many points even though f and g are distinct. The problem is that the curve of g has multiple "components"—that is, subsets which are themselves the zero locii of some polynomial—one of which coincides with the curve of f even though the full curves are not identical. This example shows that we need to require not just that f and g are distinct but that their curves share no components in common. For the polynomial g above, the components of $Z_A(g)$ correspond to the irreducible factors of g (those being $x^2 + y^2 - 1$ and $y - x$), and Theorem 2 suggests that this correspondence might hold true in general. As such, we make the following definition.

Definition 3. Two polynomials f and g are *coprime* if there is no nonconstant polynomial which is a factor of both f and g.

(We have been intentionally vague about the specific ring f and g live in so that this definition can be suitably general without getting bogged down in the algebraic details. See the section on divisibility in the appendix for a more precise discussion of coprimality.) We can now say, informally at least, that $Z_A(f)$ and $Z_A(g)$ will share no components so long as f and g are coprime. This lets us state precisely the classic version of Bézout's theorem.

Proposition 4 (Bézout's Theorem)**.** *Given any two coprime polynomials $f, g \in \mathbb{R}[x, y]$ of degrees m and n respectively, $|Z_A(f) \cap Z_A(g)| \leq mn$.*

As an almost trivial example of this theorem at work, lines in the plane are exactly the zero locii of linear (degree 1) polynomials: any two distinct linear polynomials are coprime, and indeed any two distinct lines intersect either not at all (if they are parallel) or at a single point. The culmination of this paper will be a proof of a more sophisticated result, also called Bézout's theorem, which will look rather different from Proposition 4 but will imply it as an easy corollary.

2 The Projective Plane

In the preceding section, we took the real numbers $\mathbb{R}$ as our "base field": we considered polynomials with real coefficients and curves lying in $\mathbb{R}^2$. It turns out that this is a poor choice for several reasons. First, $\mathbb{R}$ is not algebraically closed (there are polynomials, such as $x^2 + 1$, with no roots in $\mathbb{R}$) which will prevent us from stating many useful results. For the remainder of this paper, we will work only in a (fixed) algebraically closed field k and consider polynomials with coefficients in k. All the results in this paper will hold for any algebraically closed field, but there is no harm done if the reader thinks of k simply as $\mathbb{C}$. The second and more subtle problem with using $\mathbb{R}$ as we did in Section 1 is that it isn't compact. As we have already seen in complex analysis, it is often useful to consider the behavior of functions "at infinity", but this is not possible when working purely in $\mathbb{R}^2$ (or even $\mathbb{C}^2$). We address this second issue by introducing projective space.

Definition 5. Given $a, b \in k^{n+1} \setminus 0$, write $a \sim b$ if and only if $a = \lambda b$ for some $\lambda \in k$. Then $\sim$ is an equivalence relation, and we call the set of equivalence classes of $\sim$ *projective n-space*, which we denote by

$$\mathbb{P}^n = (k^{n+1} \setminus 0)/\sim \ .$$

More concisely, $\mathbb{P}^n$ is the set of linear subspaces of k^{n+1}. We denote the equivalence class $[(a_0, \ldots, a_n)]_\sim$ by $(a_0 : \ldots : a_n)$, called the *homogenous coordinates* of the point in $\mathbb{P}^n$. In general, homogenous coordinates are not unique, since $(a_0 : \ldots : a_n) = (\lambda a_0 : \ldots : \lambda a_n)$ for any nonzero $\lambda \in k$. However, if $a_n \neq 0$ then we can divide through by a_n to obtain $(b_0 : b_1 : \ldots : 1)$ where $b_j = a_j/a_n$, which *is* a unique representation of the point. This gives us a canonical copy of k^n inside of $\mathbb{P}^n$. But what about when $a_n = 0$? In this case, we can freely scale the remaining n homogenous coordinates to obtain new representations of the point without affecting a_n. In other words, points of $\mathbb{P}^n$ with $a_n = 0$ behave like points of $\mathbb{P}^{n-1}$. Since every point in $\mathbb{P}^n$ falls into one of these two categories, this lets us characterize $\mathbb{P}^n$ as the union of k^n and $\mathbb{P}^{n-1}$. Since $\mathbb{P}^0$ is just k^0 (a single point), we have that $\mathbb{P}^n = k^n \cup k^{n-1} \cup \cdots \cup k^0$. We think of each of the k^m as being attached to k^{m+1} at infinity, so, for instance, when $k = \mathbb{C}$, $\mathbb{P}^1$ is $\mathbb{C}$ with a single point attached at infinity, which is exactly the Riemann sphere. This example illustrates an important point. While $\mathbb{P}^n$ can be thought of as an n-dimensional object, it is n-dimensional with respect to k, not $\mathbb{R}$. In particular, if $k = \mathbb{C}$, then $\mathbb{P}^n$ will be $2n$-dimensional over $\mathbb{R}$.

Now we want to consider zero locii in $\mathbb{P}^n$. Unfortunately, the definition of the affine zero locus of a polynomial which we gave in Section 1 won't help us here; consider the polynomial $x + y - 2 \in k[x, y]$. This polynomial vanishes at $(1, 1)$ but not at $(2, 2)$, even though these are the same point in $\mathbb{P}^1$! The solution is to be a bit more strict about the kinds of polynomials we consider.

Definition 6. A polynomial in $k[x_1, x_2, \ldots, x_m]$ is called *homogenous of degree n* or simply homogenous if all of its terms are of degree n.

For example, $x^2 + xy$ and $xy + yz$ are homogenous, but $x^2y + xy$ and $x + 3$ are not. In general, products of homogenous polynomials are homgenous, as are monomials.

Lemma 7. *Let $f, g \in k[x, y, z]$ be nonzero polynomials. If g divides f and f is homogenous then g is homogenous.*

Proof. Suppose $f = gh$ is homogenous but g is not. Let f_n, g_n, h_n be the polynomials containing all the terms of f, g, h of degree n. Then let a and A be the smallest and largest values of n for which $g_n \neq 0$, and similarly for b, B, h. We have $f_{a+b} = g_a h_b$ and $f_{A+B} = g_A h_B$. Since g is not homogenous, $a < A$, which implies that $a + b < A + B$, but then f has terms of two distinct degrees which contradicts our assumption that f is homogenous. □

The most important characteristic of homogenous polynomials is that if $f \in k[x_1, \ldots, x_m]$ is homogenous of degree n, then

$$f(\lambda a_1, \ldots, \lambda a_n) = \lambda^n f(a_1, \ldots, a_n) \, .$$

This is clear since λ will appear n times in each term and thus can be factored out. Because of this, issues of the sort we encountered with $x + y + 2$ cannot arise with homogenous polynomials. We make this precise in the following theorem.

Theorem 8. *For any homogenous polynomial $f \in k[x_1, \ldots, x_n]$, $f(a_1, \ldots, a_n) = 0$ if and only if $f(\lambda a_1, \ldots, \lambda a_n)$ for any nonzero $\lambda \in k$.*

Proof. Suppose f is homogenous of degree n and $f(a_1, \ldots, a_n) = 0$. Then

$$f(\lambda a_1, \ldots, \lambda a_n) = \lambda^n f(a_1, \ldots, a_n) = \lambda^n 0 = 0 \,.$$

Conversely, assume $f(\lambda a_1, \ldots, \lambda a_n) = 0$. Then $f(a_1, \ldots, a_n) = f(\lambda^{-1}\lambda a_1, \ldots, \lambda^{-1}\lambda a_n)$ which vanishes by the previous implication. □

This lets us give the following definition.

Definition 9. Given a homogenous polynomial of $n+1$ variables $f \in k[x_0, \ldots, x_n]$, we define the *(projective or homogenous) zero locus* of f as

$$Z(f) = \{(a_0 : \ldots : a_n) \in \mathbb{P}^n \mid f(a_0, \ldots, a_n) = 0\} \,.$$

By Theorem 8, this is well-defined.

We have given the definition above for arbitrary dimension, but in this paper we will only consider zero locii in $\mathbb{P}^2$. We call $\mathbb{P}^2$ the *projective plane* and call zero locii lying in $\mathbb{P}^2$ *curves*, though per our earlier remark on dimension these names are potentially misleading.

Remark 10. By working in projective space we have restricted ourselves to only consider homogenous polynomials, which seems like an onerous constraint! In fact, it is not really a constraint at all. Given a polynomial $f \in k[x, y]$ (not necessarily homogenous), we can produce a homogenous polynomial $\tilde{f} \in k[x, y, z]$ by multiplying each term of f by an appropriate power of z to make it have the same degree as the highest-order term. For example, if $f(x, y) = xy^2 + 3xy + 1$, then $\tilde{f}(x, y, z) = xy^2 + 3xyz + z^3$. Clearly $\tilde{f}(x, y, 1) = f(x, y)$, so at those points of $\mathbb{P}^2$ which lie in the canonical copy of k^2, that is, those points which can be written as $(x : y : 1)$, we have that f and $\tilde{f}$ "agree". This is somewhat misleading since $\tilde{f}$ is not well-defined at points of $\mathbb{P}^2$, but at the very least we have that the points of $Z_A(f) \subset k^2$ are in natural bijection with the "non-infinite" points of $Z(\tilde{f}) \subset \mathbb{P}^2$.

3 Intersections of Curves

We now investigate the intersection of curves in the projective plane. Before getting to the meat of this section we must prove a technical lemma. This result is entirely algebraic and the reader is welcomed to take it on faith if they do not care for the details of the proof.

Lemma 11. *Let R be a UFD with K as its field of fractions. Two polynomials f and g in $R[x]$ are coprime if and only if there are polynomials $r, q \in K[x]$ such that $rf + qg = 1$.*

Proof. First, we prove that f and g are coprime as elements of $R[x]$ if and only if they are coprime as elements of $K[x]$. One direction is obvious: if there is no nonconstant polynomial in $K[x]$ which divides both f and g then there is no nonconstant polynomial in $R[x]$ which divides f and g since $R[x] \subseteq K[x]$.

Conversely, assume for contradiction that f and g are coprime in $R[x]$ but that there is some nonconstant $h \in K[x]$ which divides both f and g (in $K[x]$). Suppose in particular that $f = hp$ for $p \in K[x]$. Let c be the product of the denominators of the coefficients of h and let d be the gcd of the numerators of the coefficients, then define $h' = (c/d)h$ and $p' = (d/c)p$. By the construction of c and d, h' is in $R[x]$ and is primitve, and $f = hp = h'p'$ so $h' \mid f$ in $K[x]$.

Similarly, let a (resp. b) be the product (resp. gcd) of the denominators (resp. numerators) of the coefficients of p', and define $k = a/b$. Then kp' is in $R[x]$ and is primitive, so $kf = h'(kp')$ is also in $R[x]$ and is primitive by Gauss' lemma (Proposition 61 in the appendix). Now write k in its lowest form as a'/b', for some coprime $a', b' \in R$. If we denote the coefficients of f by $f_i \in K$, we must have that $kf_i = a'f_i/b' \in R$ and thus $b' \mid a'f_i$ (in R) for all i. Since a' and b' are coprime, we conclude that $b' \mid f_i$, but then $f_i/b' \in R$ and so $kf_i = a'(f_i/b')$ is divisible by a'. But then a' is a common divisor of the coefficients of kf, so by primitivity a' must be a unit in R. Thus $p' = b'a'^{-1}(kp') \in R[x]$. Since both h' and p' are in $R[x]$ and $f = h'p'$, this shows that $h' \mid f$ in $R[x]$ as well as $K[x]$.

By the same argument, it can be shown that $h' \mid g$ in $R[x]$, so h' is a nonconstant common divisor of f and g in $R[x]$. This contradicts our assumption that f and g are coprime in $R[x]$, so no such $h \in K[x]$ can exist whence f and g must also be coprime in $K[x]$.

Now since K is a field it is also a PID, so the result follows immediately from Bézout's identity (Theorem 57 in the appendix). □

Our first theorem on intersections is now as follows.

Theorem 12. *Given two nonzero coprime homogenous polynomials $f, g \in k[x, y, z]$, the intersection $Z(f) \cap Z(g)$ is finite.*

Proof. We can partition $\mathbb{P}^2$ into points of the forms $(1, 0, 0)$, $(a, 1, 0)$ and $(a, b, 1)$, corresponding to the canonical copies of k^0, k^1 and k^2 discussed in the previous section. There is only point of the form $(1, 0, 0)$ so it poses no problem to finiteness. In the second case, we are looking for points where $f(a, 1, 0) = g(a, 1, 0) = 0$. Suppose $f(x, 1, 0)$ and $g(x, 1, 0)$, treated as functions of x, were constantly zero. Then by homogeneity, $f(x, y, 0), g(x, y, 0) \in k[x, y]$ are both zero as well. The only way for this to occur is if the original functions $f(x, y, z)$ and $g(x, y, z)$ are either both constantly zero or both have a factor of z in every term. The former case is disallowed by assumption and in the latter case f and g would share z as a common factor, so since f and g are coprime at least one of $f(x, 1, 0)$ and $g(x, 1, 0)$ is nonzero. Since this is a nonzero polynomial of one variable it has only finitely many roots, so $f(x, 1, 0)$ and $g(x, 1, 0)$ can only vanish simultaneously at finitely many points.

For the final case, we are interested in points where $f(a, b, 1) = g(a, b, 1) = 0$. Note that $f(x, y, 1)$ and $g(x, y, 1)$ can be treated as elements of $k[x][y]$, that is, polynomials in y with coefficients in $k[x]$. Since $k[x]$ is a unique factorization domain and $k(x)$ is its field of fractions, we have by Lemma 11 that there are functions $R, Q \in k(x)[y]$ such that $Rf + Qg = 1$. Elements of $k(x)[y]$ can be thought of as rational functions with numerators in $k[x, y]$ and denominators in $k[x]$, so let $h \in k[x]$ be the product of the denominators of R and Q. Then multiplying through by h it is clear that $h \in (f(x, y, 1), g(x, y, 1)) \subset k[x, y]$. In particular, at any points where $f(a, b, 1) = g(a, b, 1) = 0$ we have $h(a) = 0$, so there are only finitely many a for which $f(a, b, 1), g(a, b, 1)$ simultaneously vanish. Repeating the argument using $k[x, y] = k[y][x]$ gives similarly that there are only finitely many b such that both polynomials vanish, completing the proof. □

It turns out that a purely set-wise treatment of $|Z(f) \cap Z(g)|$, as in the above proof, is often insufficient, and that it is necessary to assign a multiplicity to intersection points. Before we can develop this notion we need some preliminary definitions.

Definition 13. Given a point $a \in \mathbb{P}^2$, the *local ring of rational functions* at a, denoted R_a, is defined as the set of all rational functions $s/t \in k(x, y, z)$ where $s, t \in k[x, y, z]$ are homogenous of the same degree and $t(a) \neq 0$.

Definition 14. Let $f_1, \ldots, f_n \in k[x, y, z]$ be homogenous polynomials and let $a \in \mathbb{P}^2$ be a point. Then the *ideal* $(f_1, \ldots, f_2)_a \subset R_a$ *generated by* $f_1, \ldots, f_n$ is the set of rational functions in R_a with numerators in $(f_1, \ldots, f_n) \subset k[x, y, z]$. That is, it's the set of all rational functions of the form

$$\frac{g_1 f_1 + \ldots g_n f_n}{s}$$

where $g_1, \ldots g_n$ and s are homogenous polynomials in $k[x, y, z]$ such that $t(a) \neq 0$ and the $g_j f_j$ are all of the same degree as s.

We first give an equivalent characterization of the ideal $(f_1, \ldots, f_n)_a$.

Lemma 15. *Let $t_1, \ldots, t_n \in k[x, y, z]$ be arbitrary homogenous polynomials such that $t_j(a) \neq 0$ and t_j is of the same degree as f_j. Then $(f_1, \ldots, f_n)_a$ is equal to the ideal of R_a generated (in the traditional sense) by $f_1/t_1, \ldots, f_n/t_n$.*

Proof. Suppose $Q \in (f_1/t_1, \ldots, f_n/t_n) \subset R_a$. Then

$$Q = \frac{g_1}{s_1}\frac{f_1}{t_1} + \cdots + \frac{g_n}{s_n}\frac{f_n}{t_n}$$

for some $g_j/s_j \in R_a$. Putting everything over a common denominator, we get

$$Q = \frac{\sum_{j=1}^{n} g_j p_j f_j}{s_1 \ldots s_n \, t_1 \ldots t_n}$$

where p_j is the product of all the $s_l t_l$ for $l \neq j$. Thus Q is in $(f_1, \ldots, f_n)_a$, so $(f_1/t_1, \ldots, f_n/t_n) \subseteq (f_1, \ldots, f_n)_a$. Conversely, suppose $Q \in (f_1, \ldots, f_n)_a$. Then

$$Q = \frac{g_1 f_1 + \ldots g_n f_n}{s}$$

for some homogenous polynomials g_j and s such that $s(a) \neq 0$. Then

$$\frac{a_j f_j}{s} = \frac{a_j t_j}{s}\frac{f_j}{t_j}$$

for all j, so Q is a linear combination of the f_j/t_j and thus lies in $(f_1/t_1, \ldots, f_n/t_n)$. □

This lemma shows that $(f_1, \ldots, f_n)_a$ really is an ideal, as suggested by the name.

Next, we note that R_a is a vector space over k with scalar multiplication $k \times R_a \to R_a$ given by the obvious operation $w \cdot f/t = (wf)/t$. Since any element $w \in k$ embeds into R_a as $w/1$, this scalar multiplication is actually just a restriction of the ring multiplication operation of R_a. In particular, since ring multiplication is well-defined under quotients so is scalar multiplication, and R_a/I is also a vector space over k.

Definition 16. Given nonzero coprime homogenous polynomials $f, g \in k[x, y, z]$, the *intersection multiplicity* of f and g at a is defined as the k-dimension of $R_a/(f, g)_a$ and is denoted by $i_{f\cap g}(a)$.

Before continuing, we should address the finiteness of the intersection multiplicity. Since we have defined $i_{f\cap g}(a)$ in terms of the dimension of a vector space, it can in principle take on any value in $\mathbb{Z}^{\geq 0} \cup \infty$, but it turns out that the multiplicity is always finite. The proof of this is beyond the scope of this paper, and in fact we could allow $i_{f\cap g}(a)$ to be infinite without affecting much of what we will do here, but we will assume the result for simplicity.

Proposition 17. *Given nonzero coprime homogenous polynomials $f, g \in k[x, y, z]$, the intersection multiplicity $i_{f\cap g}(a)$ is finite for all $a \in \mathbb{P}^2$.*

Definition 14 is very nonobvious. We will not attempt to fully motivate it here, but over the course of this section we will hopefully get a good handle on its properties. First, we note that it does count intersections as we expect.

Theorem 18. *The intersection multiplicity $i_{f\cap g}(a)$ is strictly positive if $a \in Z(f) \cap Z(g)$ and zero otherwise.*

Proof. If $a \notin Z(f) \cap Z(g)$, then either f or g is nonzero at a. Assume without loss of generality that it is f. Then $1/f \in R_a$ so f is a unit and thus $(f, g)_a = R_a$. Hence $R_a/(f, g)_a = 0$, which has dimension zero over k. Conversely, suppose $f(a) = g(a) = 0$. Then every element of $(f, g)_a$ vanishes at a, but $1/1 \in R_a$ does not, so $(f, g)_a \neq R_a$ and $R_a/(f, g)_a$ is nontrivial. $i_{f\cap g}(a) > 0$. □

It is common to package the information about intersection multiplicities in a purely formal sum, written

$$f \cap g = \sum_{a\in\mathbb{P}^2} i_{f\cap g}(a)\, a\,.$$

This is called the *intersection cycle* of f and g, and will allow us to state some results more compactly. We also define the true numerical sum

$$\#(f \cap g) = \sum_{a\in\mathbb{P}^2} i_{f\cap g}(a)\,,$$

which we call the *intersection number* of f and g. Theorem 18 implies that all terms of $f \cap g$ corresponding to points not lying in $Z(f)\cap Z(g)$ are zero, so by Theorem 12 $f \cap g$ is always a finite sum. These results, together with Proposition 17 similarly tell us that $\#(f \cap g)$ is always finite.

We will now prove several results which allow us to compute intersection cycles and multiplicities without appealing directly to Definition 14. In all of these lemmas, f, g and h are assumed to be nonzero homogenous polynomials in $k[x, y, z]$ with degree and coprimality as necessary for all intersection cycles in the statement of the lemma to be well-defined.

Lemma 19. $f \cap g = g \cap f$.

Proof. This follows from the fact that $(f, g)_a = (g, f)_a$ and thus $i_{f\cap g}(a) = i_{g\cap f}(a)$ for all $a \in \mathbb{P}^2$. □

Lemma 20. $f \cap (g + fh) = f \cap g$.

Proof. An element of $(f, g + fh) \subset k[x, y, z]$ is of the form $pf + q(g + fh)$. By rearranging this as $(p + qh)f + qg$, we see that this also lies in (f, g), so $(f, g + fh) \subset (f, g)$. Conversely, any element $pf + qg$ of (f, g) can be rewritten as $(p - qh)f + q(g + fh)$, which lies in $(f, g + fh)$. Thus $(f, g + fh) = (f, g)$, so $(f, g + fh)_a = (f, g)_a$ and $i_{f\cap(g+fh)}(a) = i_{f\cap g}(a)$ for all $a \in \mathbb{P}^2$ □

Lemma 21. $f \cap gh = f \cap g + f \cap h$.

Proof. Let a be a point of $\mathbb{P}^2$. Fix some homogeous polynomial t_0 which does not vanish at a and is of the same degree as h, then consider the maps

$$R_a/(f, g)_a \xrightarrow{T} R_a/(f, gh)_a \xrightarrow{M} R_a/(f, h)_a$$

given by $T(Q + (f, g)_a) = (h/t_0)Q + (f, gh)_a$ and $M(Q + (f, gh)_a) = Q + (f, h)_a$. First we show that these maps are well-defined. That is, given $Q, R \in R_a$, we want to show that if Q and R are equivalent in the domain of the map, then their images are equivalent in the codomain. This follows for M because $(f, gh)_a \subseteq (f, h)_a$, so if $Q - R \in (f, gh)_a$ then $Q - R \in (f, h)_a$. For T, note that if $R - r \in (f, g)_a$, then

$$Q - R = \frac{af + bg}{s}$$

so

$$\frac{h}{t_0}R - \frac{h}{t_0}R = \frac{phf + qgh}{st_0}$$

which is in $(f, gh)_a$.

We now make the following claims regarding the properties of T and M.

1. *T and M are k-linear.* This is immediate from the definitions.

2. *M is surjective.* This follows since M is simply the projection of a quotient ring onto a smaller quotient. More explicitly, given any element (equivalence class) of $R_a/(f, h)_a$, choose an arbitrary representative $q \in (f, gh)_a$ of that element. Then $M(q + (f, gh)_a) = q + (f, h)_a$ which is exactly the original element.

3. *T is injective.* Take $Q, R \in R_a$ and suppose $T(Q + (f, g)_a) = T(R + (f, g)_a)$. Then $\frac{h}{t_0}Q - \frac{h}{t_0}R \in (f, gh)_a$, so the numerator of $h(Q - R)$ must be a linear combination of terms containing either f or gh as a factor. Since f and h are colinear, this implies that the numerator of $Q - R$ is a linear combination of terms containing either f or g as a factor, so $Q + (f, g)_a = R + (f, g)_a$.

4. $\operatorname{im} T = \ker M$. Take any $Q \in R_a$. We have

$$M(T(Q + (f, g)_a)) = M\left(\frac{h}{t_0}Q + (f, gh)_a\right) = \frac{h}{t_0}Q + (f, h)_a = 0 + (f, h)_a$$

so $\operatorname{im} T \subseteq \ker M$. Conversely, suppose $M(Q + (f, gh)_a) = Q + (f, h)_a = 0 + (f, h)_a$. Then $Q = \frac{af+bh}{s}$, so $q + (f, gh)_a = bh/s + (f, gh)_a$. But $T(bt_0/s + (f, g)_a) = bh/s + (f, gh)_a$, so $\ker M \subseteq \operatorname{im} T$.

Now, since T is linear and injective it is an isomorphism into its image, but $\operatorname{im} T = \ker M$, so $R_a/(f,g)_a \cong \operatorname{im} T = \ker M$. On the other hand, M is surjective, so $\operatorname{im} M = R_a/(f,h)_a$. Thus, by the rank-nullity theorem,

$$\begin{aligned} i_{f\cap gh}(a) &= \dim R_a/(f,gh)_a \\ &= \dim \ker M + \dim \operatorname{im} M \\ &= \dim R_a/(f,g)_a + \dim R_a/(f,h)_a \\ &= i_{f\cap g}(a) + i_{f\cap h}(a)\,. \end{aligned}$$

□

Lemma 22. *If f and g are nonconstant and linear then $\#(f \cap g) = 1$.*

Proof. Write $f(x,y,z) = f_1x + f_2y + f_3z$ and $g(x,y,z) = g_1x + g_2y + g_3z$ for $f_j, g_j \in k$. We first want to show that there is exactly one point which $Z(f)$ and $Z(g)$ share in common. We have three possibilities for the ratios f_1/f_2 and g_1/g_2: either they are distinct, the same but nonzero, or both zero. We claim that these possibilities correspond to the curves having an intersection point of the form $(x : y : 1)$, $(x : 1 : 0)$ or $(1 : 0 : 0)$ respectively. To have an intersection point of the form $(x : y : 1)$ we must have $f_1x + f_2y + f_3 = g_1x + g_2y + g_3 = 0$. These equations describe the intersection of two lines in k^2 with slopes f_1/f_2 and g_1/g_2. If the slopes are equal then the lines are parallel so there are no intersection points (f and g are coprime so the lines must be distinct), otherwise they intersect at exactly one point, as desired. Next, for there to be an intersection point of the form $(x : 1 : 0)$, we must have $f_1x + f_2 = g_1x + g_2 = 0$. If $f_1/f_2 = g_1/g_2 \neq 0$, then $x = -f_2/f_1 = -g_2/g_1$ is the unique solution to those equations. Conversely, if those equations are satisfied then f_1 and g_1 must be nonzero—if $f_1 = g_1 = 0$ then $f_2 = g_2 = 0$ as well, so f and g would be multiples of z and not coprime—and by rearranging the equation it is clear that we must have $f_1/f_2 = g_1/g_2$. Finally, $(1 : 0 : 0)$ is in the intersection exactly when $f(1,0,0) = f_1 = 0$ and $g(1,0,0) = g_1 = 0$.

Since these three possibilities for the ratios are mutually exclusive and in each case there is exactly one intersection point, we must have that $Z(f)$ and $Z(g)$ share only one point of $\mathbb{P}^2$ in common, which we will call a_0. We now show that $i_{f\cap g}(a_0) = 1$. Since k^3 has three dimensions, we can choose some polynomial $h(x,y,z) = h_1x + h_2y + h_3z$ such that the matrix

$$J = \begin{pmatrix} f_1 & f_2 & f_3 \\ g_1 & g_2 & g_3 \\ h_1 & h_2 & h_3 \end{pmatrix}$$

is invertible. Since $J(x,y,z) = (f,g,h)$, any polynomial p in indeterminates x, y, z can be rewritten as a polynomial $p \circ J^{-1}$ in f, g, h. In particular, if $Q \in R_{a_0}$ is a nonzero rational function then we can express the numerator as polynomial in f, g, h in this way and then further rewrite it as follows. First, collect all the terms with f as a factor and write these term as fs_2 where $s_2 \in k[f,g,h]$ is homogenous. The remaining terms form a polynomial in just the variables g, h, so again collect the terms with g as a factor and write them as gs_1 for $s_1 \in k[g,h]$ homogenous. Finally, the remaining terms can depend only on h, so since the original polynomial was homogenous this must be of the form $h^m s_0$ for some integer m and $s_0 \in k$. Performing the same process on the denominator, we get that

$$Q = \frac{fs_2 + gs_1 + h^m s_0}{ft_2 + gt_1 + h^m t_0}\,.$$

Note that $ft_2 + gt_1 + h^m t_0$ must be nonzero at a_0 but both f and g vanish there, so $t_0 \neq 0$. Then we have

$$Q - \frac{s_0}{t_0} = \frac{f(s_2 t_0 - s_0 t_2) + g(s_1 t_0 - s_0 t_1)}{f t_0 t_2 + g t_0 t_1 + t_0^2 h^m}$$

which is an element of $(f, g)_{a_0}$. Thus, every element of R_{a_0} is equal to a constant modulo $(f, g)_{a_0}$, so $R_{a_0}/(f, g)_{a_0}$ has dimension 1 over k. □

These last two results are particularly powerful. On the one hand, Lemma 21 lets us compute the intersection cycles of polynomials in terms of the cycles of their products, which will be the key ingredient in our proof of Bézout's theorem. For now, we just note that it gives us the identities

$$f \cap g^n = n(f \cap g)$$

and

$$f \cap \lambda g = f \cap \lambda + f \cap g = f \cap g$$

for $\lambda \in k$ nonzero. On the other hand, Lemma 22 tells us that if f and g are linear and we find a single point $a \in \mathbb{P}^2$ such that $f(a) = g(a) = 0$, then a is in fact the only intersection point of the curves of f and g and it has multiplicity one, so $f \cap g = a$. For instance, this gives us immediately that $y \cap z = (1 : 0 : 0)$.

4 Bezout's Theorem

We can now state the full version of Bézout's theorem.

Theorem 23 (Bézout's Theorem Revisited)**.** *Let $f, g \in k[x, y, z]$ be nonzero coprime homogenous polynomials of orders m and n respectively. Then $\#(f \cap g) = mn$.*

Before continuing on to the proof, we first show that Theorem 23 does indeed imply Proposition 4 as promised. This will essentially go by extending the curves in $\mathbb{R}^2$ to curves in $\mathbb{P}^2$ and noting that we have at worst added new intersection points.

Proof. Suppose we have two coprime polynomials $f, g \in \mathbb{R}[x, y]$ with degrees m and n respectively. We can extend $Z_A(f)$ and $Z_A(g)$ to $\mathbb{C}^2$ simply by treating f and g as elements of $\mathbb{C}[x, y]$ and looking at the set of complex points where the polynomials vanish. Note that f and g are also coprime as elements of $\mathbb{C}[x, y]$. Since f and g are coprime in $\mathbb{R}[x, y]$, it follows from Lemma 11 that there are $R, Q \in \mathbb{R}(x)[y]$ such that $Rf + Qg = 1$. Since this is a formal equality of polynomials it also holds for values in $\mathbb{C}$, so by converse implication of the same lemma f and g are coprime in $\mathbb{C}[x, y]$ as well. Now by Remark 10, we can obtain homogenous polynomials $\tilde{f}, \tilde{g} \in \mathbb{C}[x, y, z]$ such that $\tilde{f}(x, y, 1) = f(x, y)$ and $\tilde{g}(x, y, 1) = g(x, y)$. Note that the degrees of $\tilde{f}$ and $\tilde{g}$ are still m and n. I claim again that $\tilde{f}$ and $\tilde{g}$ are coprime. Suppose $\tilde{f} = qr$ and $\tilde{g} = pr$. Then $f(x, y) = q(x, y, 1)r(x, y, 1)$ and $g(x, y) = p(x, y, 1)r(x, y, 1)$, violating our assumption that f and g are coprime.

There are now four different intersection counts we can consider:

1. The number of intersection points $|Z_A(f) \cap Z_A(g)|$ in $\mathbb{R}^2$ (counted without multiplicity)

2. The number of intersection points in $\mathbb{C}^2$ (counted without multiplicity)

3. The number of intersection points $|Z(\tilde{f}) \cap Z(\tilde{g})|$ in $\mathbb{P}^2$ (counted without multiplicity)

4. The intersection number $\#(f \cap g)$ in $\mathbb{P}^2$ (counted with multiplicity)

Theorem 23 tells us that (4) is exactly mn while Proposition 4 states that (1) is no more than mn. I claim that each count is no greater than the next, which will prove the result. (1) $\leq$ (2) simply because the zero locii in $\mathbb{R}^2$ are subsets of the locii in $\mathbb{C}^2$. (2) $\leq$ (3) follows in a similar manner, though it is slightly subtler—the locii in $\mathbb{C}^2$ are not a priori subsets of the locii in $\mathbb{P}^2$, but they are in canonical bijection with such subsets, which is sufficient. Finally, (3) $\leq$ (4) follows from Theorem 18.

□

Now, the proof of Theorem 23 will proceed by strong induction on the x degree of g. We first prove the base case when $g \in k[y, z]$, then give an algorithm which will allow us to compute $f \cap g$ in terms of the intersection cycles of lower degree polynomials. Finally, we tie together these results in a short inductive proof. First we need the following lemma regarding factoring homogenous polynomials.

Lemma 24. *Let $f \in k[x, y]$ be a nonzero homogenous polynomial in two variables with degree n. Then f factors as*

$$f(x, y) = \lambda y^t \prod_{j=1}^{s} (x - \alpha_j y)$$

for some $\lambda, \alpha_j \in k$, with $\lambda \neq 0$ and $s + t = n$.

Proof. Let s be the x degree of f and let $t = n - s$. Then the term of f with the highest x degree will be of the form $\lambda y^t x^s$. Since f is homogenous, every other term of f will have a y degree of at least t, so we can factor f as

$$f(x, y) = \lambda y^t f'(x, y)$$

where f' is homogenous and has both a total and an x degree of s. Hence $f'(x, 1) \in k[x]$ will be a monic polynomial of degree s, so since k is algebraically closed it will split as

$$f'(x, 1) = \prod_{j=1}^{s} (x - \alpha_j)$$

where α_j are the roots of $f'(x, 1)$. Then by homogeneity,

$$f'(x, y) = y^t f\left(\frac{x}{y}, 1\right) = y^t \prod_{j=1}^{s} \left(\frac{x}{y} - \alpha_j\right) = \prod_{j=1}^{s} (x - \alpha_j y)$$

which proves the lemma. □

Theorem 25. *Let $f \in k[x, y, z]$, $g \in k[y, z]$ be nonzero coprime homogenous polynomials of degrees m and n respectively. Then $\#(f \cap g) = mn$.*

Proof. Using Lemma 24, we can factor g as

$$g(y,z) = \lambda z^s \prod_{i=1}^{n-s}(y - \alpha_i z)\,,$$

so

$$f \cap g = s(f(x,y,z) \cap z) + \sum_{i=1}^{n-s} f(x,y,z) \cap (y - \alpha_i z)\,.$$

First, consider the intersection cycle $f(x,y,z)\cap z$. By separating out the terms of f with nonzero z degree, we can write

$$f(x,y,z) = f(x,y,0) + zf'(x,y,z)$$

whence by Lemma 20 we have $f(x,y,z) \cap z = f(x,y,0) \cap z$. By writing

$$f(x,y,0) = \mu y^t \prod_{j=1}^{m-t}(x - \beta_j y)$$

we get

$$f(x,y,0) \cap z = t(y \cap z) + \sum_{j=1}^{m-t}(x - \beta_j y) \cap z\,.$$

Clearly $(1:0:0) \in Z(y) \cap Z(z)$ and $(\beta_j : 1 : 0) \in Z(x - \beta_j y) \cap Z(z)$, but by Lemma 22 these can be the only intersection points and they must occur with multiplicty one. Thus $y \cap z = (1:0:0)$ and $(x - \beta_j y) \cap z = (\beta_j : 1 : 0)$.

Now consider the cycle $f(x,y,z) \cap (y - \alpha_i z)$. Define $h_i(x,w,z) = f(x, w + \alpha_i z, z)$ so that $f(x,y,z) = h_i(x, y - \alpha_i z, z)$. Then as before, we can separate the terms of h_i and write

$$h_i(x,w,z) = h_i(x,0,z) + wh_i'(x,w,z),$$

so

$$f(x,y,z) = f(x,\alpha_i z, z) + (y - \alpha_i z)h_i'(x, y - \alpha_i z, z)\,.$$

Thus by Lemma 20 again, we have

$$f(x,y,z) \cap (y - \alpha_i z) = f(x,\alpha_i z, z) \cap (y - \alpha_i z)\,.$$

Since $f(x,\alpha_i z, z)$ is a now a polynomial of two variables, we can factor it yet again as

$$f(x,\alpha_i z, z) = \eta_i z^{l_i} \prod_{j=1}^{m-l_i}(x - \gamma_{ij} z)\,,$$

and thus

$$f(x,\alpha_i z, z) \cap (y - \alpha_i z) = l_i(z \cap (y - \alpha_i z)) + \sum_{j=1}^{m-l_i}(x - \gamma_{ij} z) \cap (y - \alpha_i z)\,.$$

We have immediately that $z \cap (y - \alpha_i z) = z \cap y = (1:0:0)$, and it is clear by the same arguments as before that $(x - \gamma_{ij} z) \cap (y - \alpha_i z) = (\gamma_{ij} : \alpha_i : 1)$.

Putting all of this together, we have

$$f \cap g = s\left(t(1:0:0) + \sum_{j=1}^{m-t}(\beta_i : 1 : 0)\right) + \sum_{i=1}^{n-s}\left(l_i(1:0:0) + \sum_{j=1}^{m-l_i}(\gamma_{ij} : \alpha_i : 1)\right).$$

Counting the terms in the cycle then gives that $\#(f \cap g) = mn$. □

We now describe an algorithm which will allow us to compute the intersection cycle $f \cap g$ in terms of intersections of lower-degree polynomials. In what follows, ∂f denotes the total degree of a polynomial f and $\partial_x f$ denotes its x degree. Assume that $\partial_x f \geq \partial_x g$. Then by treating f and g as polynomials of a single variable x with coefficients in $k[y, z]$, we can perform polynomial long division to obtain unique rational functions $Q, R \in k(y, z)[x]$ with $\partial_x R < \partial_x g$ such that

$$f = Qg + R\,.$$

Note that since f and g are coprime, R must be nonzero. The denominators of Q and R will be polynomials in $k[y, z]$—call their least common denominator h. Then

$$hf = qg + r$$

where $q = hQ$ and $r = hR$ are polynomials in $k[y, z]$. Now let c be the greatest common divisor of g and r. Then c divides $hg + r$, so it must divide hf as well. But if c is nonconstant, then by the coprimality of f and g it cannot divide f, so it must divide h. In fact, any common divisor of h and g will also divide r, so we must have that c is the gcd of g and h as well. Thus, divide through the equation above by c to obtain

$$h'f = qg' + r'$$

where $h = h'c$, $g = g'c$, $r = r'c$.

We now have four new polynomials of interest: g', h', r' and c. What are their properties? We take it for granted that the numerators and denominators of the rational functions Q and R obtained from polynomial division are homogenous (see [4], Lemma 3.5 for a careful treatment), so that q, r and h are all homogenous. Since g', h', r' and c all divide one of those three polynomials, they are all homogenous as well by Lemma 7. Furthermore, since g' divides g we have $\gcd(g', f) = 1$ and we already argued above that $\gcd(f, c) = 1$. Since h' and r' were obtained from h and r by dividing out $c = \gcd(g, h) = \gcd(g, r)$, so $\gcd(g', h) = \gcd(g', r) = 1$. All of this has simply been to check the technicalities which allow us to note the following:

$$f \cap g = f \cap g' + f \cap c = h'f \cap g' - h' \cap g' + f \cap c = r' \cap g' - h' \cap g' + f \cap c\,.$$

Finally, we note the x degrees of our newly obtained polynomials. Since h is the lcm of polynomials in $k[y, z]$ its x degree is 0, as is the x degree of c since c divides h. Furthermore, $\partial_x r = \partial_x R < \partial_x g$ by assumption, but since $\partial_x c = 0$ we have $\partial_x g = \partial_x g'$ and $\partial_x r = \partial_x r'$ so $\partial_x r' < \partial_x g'$.

To summarize, given any two coprime homogenous polynomials f and g, we have an algorithm which produces four new homogenous polynomials with x degree less than that of g such that $f \cap g$ can be computed in terms of the intersection cycles of the new polynomials. This gives us the final ingredient for our inductive proof of Theorem 23.

Proof. Let $f, g \in k[x, y, z]$ be coprime homogenous polynomials and assume without loss of generality that $\partial_x f \geq \partial_x g$.

If g has an x degree of zero, then it follows from Theorem 25 that $\#(f \cap g) = \partial f \partial g$. Now suppose $\partial_x g = N \neq 0$ and assume for the inductive hypothesis that Bézout's theorem holds for any two polynomials such that the x degree of at least one of them is strictly less than N. Performing the algorithm described above, we have homogenous polynomials g', h', r', c such that

$$f \cap g = r' \cap g' - h' \cap g' + f \cap c$$

and hence

$$\#(f \cap g) = \#(r' \cap g') - \#(h' \cap g') + \#(f \cap c) \,.$$

Since r', h' and c all have x degree less than N, we have by the inductive hypothesis that

$$\#(f \cap g) = \partial r' \partial g' - \partial h' \partial g' + \partial f \partial c \,.$$

Since $g = g'c$ we have $\partial g' + \partial c = \partial g$. Similarly, since $h' f = q g' + r'$ and all the above polynomials are homogenous, it follows that $\partial(h' f) = \partial h' + \partial f = \partial r'$. Thus

$$\partial r' \partial g' - \partial h' \partial g' + \partial f \partial c = (\partial r' - \partial h') \partial g' + \partial f \partial c = \partial f \partial g' + \partial f \partial c = \partial f (\partial g' + \partial c) = \partial f \partial g \,.$$

This proves the result.

□

5 Appendix: Algebra

Introduction

In this appendix, we give a very short overview of the key ideas and results from the field of algebra, more specifically the theory of rings and fields, which are used throughout this paper. This is not in any way meant to be a comprehensive introduction. In the interest of brevity, we skip over countless crucial topics, leave out details and state important results without proof. To fill in these gaps we suggest the reader consult any introductory textbook on algebra such as that by Dummit and Foote [1], which was used as a reference when compiling this material.

The mathematical field of algebra offers a language with which to formalize and generalize familiar notions of arithmetic on various (possibly unfamiliar) objects. Consider, for instance, integers, real numbers and matrices. Each of these objects have some sort of ari thmetic operations which can be performed on them and which follow broadly similar rules. In all three cases there is a notion of addition, subtraction and multiplication, as well as additive and multiplicative identities. However, arithmetic on these objects is not identical. You cannot in general divide two integers to obtain a new integer, for instance, and matrix multiplication is not commutative. So-called "abstract algebra" gives us tools for keeping track of these differences. To start, we have the following definition.

Definition 26. A ring is a set R with two associated binary operations $+, \cdot : R \times R \to R$, canonically thought of as addition and multiplication, such that the following axioms hold.

(a) Addition and multiplication are both associative.

(b) Addition is commutative.

(c) There exist (not necessarily distinct) elements of R, denoted 0 and 1 respectively, such that $0 + a = a + 0 = a$ and $1 \cdot a = a \cdot 1 = a$ for all $a \in R$.

(d) For all $a \in R$, there exists an element, denoted by $-a$, such that $a + (-a) = 0$.

(e) For all $a, b, c \in R$, $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$.

We can now summarize some of our observations by saying that the set of integers, equipped with the standard notion of addition and multiplication, form a ring, as do the set of real numbers and the set of matrices. It is important to note what is *not* included in this definition: commutativity and the existence of inverses is required for addition but not multiplication. If a ring R does happen to have a commutative multiplicative operation then we say that R itself is a *commutative ring*. In particular, $\mathbb{R}$ and $\mathbb{Z}$ are commutative, as are all of the rings considered in this paper, but the ring of matrices is not.

Even from these limited axioms it is possible to prove a great number of facts. For instance, 0, 1 and additive inverses are all unique (though we did not require this a priori) and we have that $0 \cdot a = 0$ for all $a \in R$. Since these follow purely from the ring axioms, these statements are true for *all* rings R.

It is worth emphasizing again that elements of a ring need not have multiplicative inverses. In fact, we have a special name for those that do.

Definition 27. Let R be a ring. An element $a \in R$ is called a *unit* if there is some $b \in R$ such that $ab = ba = 1$.

Such an element b is often denoted a^{-1}. As before, it is easy to show that for any a, the element a^{-1} is unique, if it exists at all. Just because one element of a ring is a unit does not mean that we can go about dividing with reckless abandon—a single ring can have some elements which are invertible and others which are not. Furthermore, even if a is a unit, the expression c/a for some arbitrary $c \in R$ is still not well-defined: do we mean ca^{-1} or $a^{-1}c$? In a general ring these need not be the same value. So, to be able to "divide" in any familiar sense in a ring, we need stronger conditions.

Definition 28. A commutative ring R is called a *field* if every nonzero element of R is a unit.

In this paper, fields are denoted with a lowercase k. If $a, b \in k$ are elements of a field, we can now write a/b without concern, so long as $b \neq 0$. (0, of course, is never a unit since $0a$ is always 0.) This definition gives us another way to formalize some of the informal observations on different kinds of arithmetic which we began the section with. In particular, $\mathbb{R}$ and $\mathbb{C}$ are fields but $\mathbb{Z}$ is not.

With some basic definitions under our belt, we can investigate some other, slightly more exotic examples of rings. We have, for instance, the trivial ring 0, which contains only a single element. This is an important example, though not a particularly interesting one. A more interesting example can be found in polynomials.

Definition 29. Let R be a commutative ring. The *polynomial ring* $R[x]$ (read "R adjoin x") is definited as the set of formal sums

$$p(x) = \sum_{j=0}^{n} a_j x^j$$

where $a_j \in R$, $a_n \neq 0$. Addition is defined term-wise and multiplication is given by the familiar distributive formulas. The number n is called the *degree* of the polynomial p.

Polynomials and polynomial rings are one of the primary objects of study in this paper and we will discuss them extensively. The definition above is for polynomials in a single variable x, but we can also form a ring of multivariate polynomials, denoted $R[x_1, x_2, \ldots, x_m]$. Such rings can be defined directly in an analogous way to the above using more complicated formal sums, but an equivalent and often more useful approach is to define them recursively as $R[x_1, x_2, \ldots, x_m] = R[x_1, x_2, \ldots, x_{m-1}][x_m]$. As an example, this amounts to treating elements of the ring $R[x, y]$ as polynomials of a single variable y whose coefficients are *themselves* polynomials of x.

In a multivariate polynomial ring such as $R[x, y, z]$, we define the *x degree* of a polynomial f to be its degree in $R[y, z][x]$ (that is, simply the highest power on the variable x in the polynomial). The *total degree* or just *degree* of a term of f is then the sum of the term's x, y and z degrees, and the degree of the whole polynomial f is the greatest degree of any of its terms.

Ideals and Quotients

Other important examples of rings can be found in modular arithmetic. Let $\mathbb{Z}/n\mathbb{Z}$ be the set of integers $\{0, \ldots, n-1\}$ equipped with the usual addition and multiplication operators, with the caveat that the arithmetic "wraps around". For instance, treating 3 and 5 as elements of $\mathbb{Z}/7\mathbb{Z}$,

we have $3 \cdot 5 = 15$, which is not a priori in our ring, but we considre 15 to be equivalent to 8 "modulo 7" since $15 = 8 + 7$. One very useful way to think about this is to define an equivalence relation on $\mathbb{Z}$ such that $a\ b$ if and only if n divides $a - b$. $\mathbb{Z}/n\mathbb{Z}$ is then just the set of equivalence classes $\mathbb{Z}/$, and the operations are performed by choosing representatives from the equivalence classes. This idea of taking a ring and making it smaller by identifying certain elements with one another is called *quotienting* and is one of the most fundamental tools of algebra. However, not all equivalences lead to well-formed rings. For instance, taking the integers again, if we identified 5 and 7, declaring them to be equal but leaving all other numbers untouched, we get nonsense. The products $5 \cdot 3$ and $7 \cdot 3$ are formally equivalent in this "ring", but they yield numbers, 15 and 21, which are not! The problem, in essence, is that this equivalence does not respect the arithmetic structure of the integers. One approach to define more well-behaved equivalence relations is to choose some subset I of our ring and declare the elements of I to be 0 in the new quotient ring. Other elements a, b of the new ring are then equivalent if and only if $a - b \in I$. If we let R be $\mathbb{Z}$ and I be the set of integer multiples of n, then we get exactly our characterization of $\mathbb{Z}/n\mathbb{Z}$ from above. However, even this approach is not quite sufficient. We also need the elements of I to be behave like zero in a certain sense. In particlar, we certainly want the original additive identity to be in I, and sums and products of elements which we have declared to be 0 should also be treated as 0.

Definition 30. Let R be a ring. A subset $I \subset R$ is called an *ideal* of R if the following properties are satisfied.

(a) $0 \in I$.

(b) If $a, b \in I$ then $a + b \in I$.

(c) If $a \in I$ and $r \in R$ then $ar, ra \in I$.

This lets us give the formal definition of a quotient ring.

Definition 31. Let R be a ring and I be an ideal of R. Define an equivalence relation $\sim$ on R such that $a \sim b$ if and only if $a - b \in I$. The set of equivalence classes of $\sim$ is denoted R/I (read "R mod I") and is called a *quotient ring* of R. The equivalence class of $a \in R$ is denoted $a + I$. The operations on R/I are defined by choosing a representative of each equivalence class, performing the operation in R, then taking equivalence classes again at the end.

First, we prove that this gives us a well-defined ring.

Theorem 32. *Let R be a ring and I be an ideal of R. Then the operations given in Definition* 31 *are well-defined and make R/I into a ring.*

Proof. First we show that the operations are well-defined. Take $a, b, a', b' \in R$ such that $a \sim a'$ and $b \sim b'$. We want to show that $a + b \sim a' + b'$ and $ab \sim a'b'$. We have that $a' = a + i$ and $b' = b + j$ for some $i, j \in I$, so $(a' + b') - (a + b) = i + j$ which is indeed in I. Similarly, $a'b' = ab + aj + bi + ij$, so $a'b' - ab = aj + bi + ij \in I$, as desired. To see that R/I forms a ring, note that the operations of R/I are just the operations of R performed on representatives, so the ring axioms are satisfied immediately. □

Given a ring R, how can we find ideals of R? First, we note that 0 (the singleton subset) and R (the whole ring) are both trivially ideals of R. What are the corresponding quotients for these ideals? Take $a, b \in R$. We have $a - b = 0$ if and only if $a = b$, so when $I = 0$ the the equivalence relation $\sim$ in Definition 31 is just equality. On the other hand, $a - b$ is *always* in R, so the when $I = R$, $\sim$ identifies every element with every other element. Thus $R/0$ is just R and R/R is the trivial ring 0.

Now we turn to a method of constructing more interesting ideals. Suppose S is an arbitrary subset of R. We already know there is at least one ideal which contains S, namely R, but we might ask what the *smallest* ideal containing S is.

Definition 33. Let R be a ring and S be a subset of R. Then define (S) as the set of finite sums of elements of the form rsr' where $s \in S$, $r, r' \in R$. We call (S) the *ideal generated by* S in R.

If $S = \{a_1, a_2, \ldots, a_n\}$ then we omit the curly braces and write (S) as $(a_1, a_2, \ldots, a_n)$. (S) is clearly an ideal since it satisfies the necessary closure properties by definition, and it is the smallest ideal containing S since it has only enough elements to do so. We can think of $R/(S)$ as the ring R where we have declared all the elements of S to be zero and made the minimum number of additional changes for our arithmetic to remain consistent. For instance, $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/(n)$.

Note that condition (c) in Definition 30 is quite strong: the products of elements of I with *any* other element in the ring must also be in I. The strength of this requirement is apparent in the following result.

Theorem 34. *Let R be a ring and I an ideal of R. If I contains a unit then $I = R$.*

Proof. Suppose $a \in I$ and $ba = 1$. Then given any element $c \in R$, $c = c(ba) = (cb)a$ which is in I. □

An important corollary of this result is the following.

Theorem 35. *The only ideals of a field k are 0 and k itself.*

Proof. Let I be an ideal. If I contains a nonzero element then it contains a unit, so by Theorem 34, $I = k$. □

Integral Domains

We now consider a common pathology in ring theory, using the modular integer rings $\mathbb{Z}/n\mathbb{Z}$ as an example. In $\mathbb{Z}/4\mathbb{Z}$, we have that $2 \cdot 3 \equiv 2 \equiv 2 \cdot 1$, even though $3 \not\equiv 1$. This shows that the cancellation law $ab = ac \implies b = c$ does not hold in all cases for rings. The chain of logic which proves the cancellation law for elementary arithmetic (that is, arithmetic in $\mathbb{Z}$ or $\mathbb{R}$) is that if $ab = ac$, then $a(b - c) = 0$, so if $a \neq 0$ we must have $b - c = 0$ and $b = c$. Every one of these manipulations is valid in an arbitrary ring except for the implicit assumption that if $a(b - c)$ is zero, one of a or $b - c$ must be zero—none of the ring axioms require that the product of two nonzero elements be nonzero, and indeed, our pathological example in $\mathbb{Z}/4\mathbb{Z}$ arose exactly because $2(3 - 1) = 2 \cdot 2 \equiv 0$.

Definition 36. Let R be a ring. An element $a \in R$ is a *zero divisor* if there is some nonzero $b \in R$ such that either ab or ba is 0.

If a is not a zero divisor then $ab = ac$ does indeed imply that $b = c$. In general, noncommutativity and the presence of zero divisors are responsible for much of the nonintuitive behavior of rings. To avoid these issues we often want to consider rings without zero divisors.

Definition 37. A commutative ring R is an *integral domain* if the only zero divisor of R is 0 itself.

$\mathbb{Z}$ is an integral domain but $\mathbb{Z}/4\mathbb{Z}$ is not.

Theorem 38. *Let R be a commutative ring. If $a \in R$ is a unit then it is not a zero divisor.*

Proof. If $ab = 0$ for some $b \in R$, then $a^{-1}ab = a^{-1}0 \implies b = 0$. □

Corollary 39. *Every field is an integral domain.*

Although we have defined $R[x]$ for arbitrary commutative rings, many useful and seemingly obvious properties of polynomials only hold in the case that R is an integral domain. For instance, we would like to say that for $f, g \in R[x]$ we always have that $\partial(fg) = \partial f + \partial g$, but taking $f(x) = g(x) = 2x+1$ in $\mathbb{Z}/4\mathbb{Z}$, we have $f(x)g(x) = (2x+1)^2 = 4x^2+4x+1 = 1$. Not only does this product have degree 0 when it "should" have degree 2, it also shows that nonconstant polynomials can be units, contrary to our intuition. Thankfully, problems of these sort cannot arise when R is an integral domain.

Theorem 40. *Let R be an integral domain.*

(a) $R[x_1, \ldots, x_n]$ is an integral domain.

(b) For all nonzero $f, g \in R[x_1, \ldots, x_n]$, $\partial(fg) = \partial f + \partial g$.

(c) The units of $R[x_1, \ldots, x_n]$ are exactly the units of R (treated as constant polynomials).

Proof. We first prove (b), from which (a) and (c) will follow. Suppose f and g have degrees m and n respectively. Then choose a term of maximal degree from both f and g and call the (nonzero) coefficients a and b, respectively. Then fg will have a term of degree $m + n$ with coefficient ab, which is nonzero since R is an integral domain. Thus $\partial(fg)$ is at least $m + n$, but clearly no product of terms of f and g will lead to a term in fg of degree greater than $m + n$, so $\partial(fg) = m + n$.

From this, it is immediate that if fg is constant (in particular, if $fg = 0$ or $fg = 1$) then both f and g must be constant. But then we are simply working in R, so it follows that the units and zero divisors of $R[x_1, \ldots, x_n]$ are exactly the units and zero divisors of R, proving (a) and (b). □

Fields of Fractions

Intuitively, a field is just a commutative ring in which one can divide. This begs the question, if R is a commutative ring, is there some way to turn R into a field by making it bigger? More precisely, can we construct a field of which R is a subset? The integers $\mathbb{Z}$ and rational numbers $\mathbb{Q}$ are an instructive example here. $\mathbb{Q}$ is in a precise sense the minimal field obtained from $\mathbb{Z}$ by simply allowing division. The analogous construction for arbitrary integral domains is as follows.

Definition 41. Let R be an integral domain. We define the *field of fractions* of R as the set K of formal sums a/b for $a, b \in R$, $b \neq 0$ under the equivalence relation

$$\frac{a}{b} \sim \frac{c}{d} \iff ad = bc\,,$$

with ring operations given by

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd} \qquad \text{and} \qquad \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}\,.$$

We first note that $\sim$ defined as above really is an equivalence relation. Symmetry and reflexivity are immediate, leaving only transitivity. Suppose $a/b \sim c/d$ and $c/d \sim e/f$. Then $ad = bc$ and $cf = ed$, so $acf = aed = ade = bce$, which implies that $af = be$ and $a/b \sim e/f$ since R is an integral domain and c, being a denominator, is nonzero. (This explains our requirement that R be an integral domain and is yet another example of how zero divisors can wreak havoc.) Having checked this, it is tedious but straightforward to confirm that the ring operations given for K are well-defined with respect to $\sim$ and obey the ring axioms. Furthermore, we can embed any element a of R into K as $a/1$, and we see that K is indeed a field since given any $a/b \in K$, $a \neq 0$, we have

$$\frac{a}{b} \cdot \frac{b}{a} \sim \frac{1}{1}\,.$$

The field of fractions of $\mathbb{Z}$ is $\mathbb{Q}$, as expected, while the field of fractions of a field k is just k itself. More interestingly, given an integral domain R, the field of fractions of $R[x_1, \ldots, x_n]$ is the field of (formal) rational functions in n variables, denoted $R(x_1, \ldots, x_n)$. These will play an important role at several points in this paper.

UFDs and PIDs

We now introduce a few more classes of rings with certain special properties.

There are two further properties that integral domains can have, but before we introduce them we need some preliminary definitions.

Definition 42. Two elements $a, b \in R$ are called *associates* if $a = ub$ for some unit $u \in R$.

For example, since the units in a polynomial ring are exactly the constants, two polynomials are associates if they are equal up to a multiplicative constant. Similarly, the units of $\mathbb{Z}$ are just 1 and -1, so two integers are associates if they are equal up to sign.

Definition 43. An element $a \in R$ is *reducible* if $a = bc$ for some non-units $b, c \in R$ and *irreducible* otherwise.

The irreducble elements of $\mathbb{Z}$ are exactly the prime numbers and their negatives.

Definition 44. An integral domain R is a *unique factorization domain* or UFD if every nonzero, nonunit element $a \in R$ factors as $a = b_1 \ldots b_n$ for each b_j irreducible in a manner which is unique up to rearrangements and associates. More precisely, the factorization should be unique in the sense that if $a = b'_1 \ldots b'_m$ is another factorization with b'_j irreducible, then there is a bijection $\varphi : \{1, \ldots, n\} \to \{1, \ldots, m\}$ such that b_j and $b'_{\varphi(j)}$ are associates for all j.

If $R = \mathbb{Z}$ then this definition is equivalent to the statement that "every integer factors as a product of primes that is unique up to sign and rearrangement", which is exactly the fundamental theorem of arithmetic. Hence $\mathbb{Z}$ is a UFD.

Definition 45. Let R be a ring. An ideal I of R is *principal* if it is generated by a single element, that is, if there is some $a \in R$ such that $I = (a)$.

Definition 46. An integral domain R is a *principal ideal domain* or PID if every ideal of R is principal.

These two special types of integral domains, UFDs and PIDs, seem unrelated, but they are rather intimately linked by the following theorem (which we will not prove).

Proposition 47. *Every PID is a UFD.*

In general, it can be quite difficult to determine whether an integral domain is a UFD or PID. However, as this paper is largely concerned with polynomial rings over fields, the following results (some of which we leave unproven) will often suffice.

Theorem 48. *Every field is a PID.*

Proof. Let k be a field. By Theorem 35, the only ideals of k are $0 = (0)$ and $k = (1)$, both of which are principal. □

Proposition 49. *If R is a UFD then $R[x]$ is a UFD.*

Proposition 50. *If k is a field then $k[x]$ is a PID.*

Note that Propositions 49 and 50 are both stated for univariate polynomial rings. Proposition 49 does generalize to multivariate rings by induction: if R is a UFD then $R[x_1]$ is a UFD, so $R[x_1][x_2] = R[x_1, x_2]$ is a UFD, and so on. However, since Proposition 50 requires a stronger condition on R than it gives us on $R[x]$, it does not generalize in the same way. Thus, if k is a field, $k[x_1, x_2, \ldots, x_n]$ is not, in general, a PID for $n > 1$.

Divisibility

In this paper, we are primarily interested in UFDs and PIDs because of their applications to the theory of divisibility, which we will now outline.

Definition 51. Let R be a commutative ring and $a, b \in R$. If $a = bc$ for some $c \in R$, then we say that a is a *multiple* of b, or that b divides or is a *divisor* of a, and we write $a|b$.

Definition 52. Let R be a commutative ring and $a, b \in R$. We say that $d \in R$ is a *common divisor* of a and b if $d|a$ and $d|b$. We say that d is a *greatest common divisor* or gcd of a and b if it is a common divisor of a and b and if all other common divisors of a and b divide d.

Greatest common divisors are not in general unique, but we do have the following result which establishes a kind of uniqueness if R is an integral domain.

Theorem 53. *If R is an integral domain with elements a and b, then any two greatest common divisors of a and b are associates.*

Proof. Let d, d' be greatest common divisors of a and b. First, suppose one of d or d' is zero, assuming without loss of generality that it is d. Then $d' | 0$—that is, d' is a zero divisor—but then since R is an integral domain this implies that $d' = 0$. Thus $d = d'$ so d and d' are associates. Now assume d and d' are both nonzero. We have $d = ud'$ and $d' = vd$ for some u and v, so $d = uvd$. Thus, since R is an integral domain, we have $uv = 1$ so u is a unit. □

Not only are gcds not necessarily distinct, they need not even exist in arbitrary commutative rings. However, if R is a UFD, then any elements $a, b \in R$ will factor uniquely (modulo the usual caveats) as products of irreducibles, so we can look at the product d of all of the factors shared between both a and b. Then d is clearly a common divisor of a and b, and we can see that it is the greatest common divisor since any other common divisor d' of a and b must itself be equal to some product of the shared irreducible factors of a and b and hence divide d. Hence, we *can* always find a gcd of any two elements of a UFD. We state this result formally for emphasis.

Theorem 54. *Any two elements of a UFD have a greatest common divisor.*

Thus, in a UFD it makes sense to (and indeed we will) write $\gcd(a, b)$ so long as we remember that the value is only unique up to associates.

In a commutative ring, a principal ideal (a) is just the set of elements of the form ra, that is, the set of multiples of a. It follows that $(a) \subseteq (b)$ if and only if $b \mid a$ (note the change in order). This gives us a way to talk about divisibility in the language of ideals, and in particular gives us the following characterization of greatest common divisors. Suppose it happened that the ideal (a, b) is principal, so that $(a, b) = (d)$ for some d. Then (d) is the smallest ideal containing both (a) and (b), so by the above correspondence d is a gcd of a and b. More precisely, $(a) \subseteq (d)$ and $(b) \subseteq (d)$ so d is a common divisor. Moreso, for any other common divisor d', we must have similarly that (d') contains both (a) and (b), so since $(d) = (a, b)$ is the intersection of all ideals containing (a) and (b) we must have $(d) \subseteq (d')$ and thus $d' \mid d$. This new characterization immediately tells us that gcds always exist in PIDs. This alone is not of interest since all PIDs are UFDs, but this correspondence will be useful shortly.

The following definition is of central importance to the rest of this paper.

Definition 55. Let R be a commutative ring. Elements a and b of R are *coprime* if there are no nonunit common divisors of a and b.

The following theorems give two further equivalent characterizations of coprimality.

Theorem 56. *If R is a commutative ring, then $a, b \in R$ are coprime if and only if 1 is a greatest common divisor of a and b.*

Proof. If 1 is a gcd of a and b, then any common divisor of a and b divides 1 and hence is a unit. Conversely, suppose a and b are coprime. Then 1 is a common divisor of a and b because 1 divides everything, and it is the greatest common divisor because every other common divisor of a and b are units and thus divide 1. □

Theorem 57 (Bézout's Identity)**.** *If R is a PID, then $a, b \in R$ are coprime if and only if there is some $r, q \in R$ such that $ra + qb = 1$.*

Proof. Since R is a PID, $(a, b) = (d)$ for some d and this d is also a gcd of a and b. We note that $(d) = (a, b)$ is the set of all linear combinations of the form $ra + qb$ for $r, q \in R$. If a and b are coprime then d is a unit, so $(d) = (a, b) = R$. In particular, $1 \in (a, b)$, so there are r and q such that $ra + qb = 1$. Conversely, suppose $ra + qb = 1$. Then $1 \in (a, b)$ so $(a, b) = R = (1)$ so 1 is a gcd of a and b. □

Remark 58. For convenience, we have thus-far limited our discussion of greatest common divisors and coprimality to the case of two elements, but these ideas generalize in an obvious manner. A gcd of a finite collection of elements S is an element which divides everything and S and which is divisible by any other common divisor of S. If we choose a single element $a \in S$, then $\gcd(S) = \gcd(a, \gcd(S \setminus a))$, so arbitrary gcds can be constructed from nested pairwise gcds. Similarly, the elements of S are coprime if there is no nonunit common divisor of S, but this is equivalent to the coprimality of every pair in S.

Having defined coprimality, we return briefly to discuss fields of fractions and give the following definition.

Definition 59. If R is an integral domain and K is its field of fractions, then an element $a/b \in K$ is said to be *reduced* or *in lowest form* if $a, b \in R$ are coprime.

If R is a UFD then it is possible to write any element $a/b \in K$ in lowest form by factoring $\gcd(a, b)$ out of both a and b, yielding an equivalent fraction.

There is a somewhat analogous definition for polynomials.

Definition 60. Let R be an integral domain. A polynomial $f = f_0 + f_1 x + \ldots f_n x^n \in R[x]$ is *primitive* if its coefficients $f_0, \ldots, f_n$ are coprime.

Once more, if R is a UFD then it is always possible to factor $f \in R[x]$ as $c_f f^*$, where $c_f \in R$ (called the *content* of f) is the gcd of the coefficients and $f^* \in R[x]$ (called the primitive part of f) is primitive. The final theorem in this appendix (which we will not prove) is known as Gauss' lemma. There are a number of results which go by this name, but most of them are easy corollaries of the statement below.

Proposition 61 (Gauss's Lemma)**.** *Let R be a UFD. If $f, g \in R[x]$ are both primitive then fg is primitive as well.*

References

[1] David Dummit, Richard Foote. *Abstract Algebra.* University of Vermont, John Wiley Sons, 2004.

[2] Andreas Gathmann. *Algebraic Geometry, notes for a class.* University of Kaiserslautern, 2002/3.
`http://mathematik.uni-kl.de/~gathmann/class/alggeom-2002/alggeom-2002.pdf`.

[3] Jan Hilmar, Chris Smyth. *Euclid meets Bézout: Intersecting algebraic plane curves with the Euclidean algorithm.* American Mathematical Monthly, vol. 117, no. 3, pp. 250-260, 2010.

[4] R. P. Hulst. *A Proof of Bézout's Theorem Using the Euclidean Algortihm.* Mathematisch Instituut, Universiteit Leiden, 2011.
http://math.leidenuniv.nl/scripties/HulstBach.pdf.