1

Geometry, Algebra, and Algorithms

This chapter will introduce some of the basic themes of the book. The geometry we are interested in concerns *affine varieties*, which are curves and surfaces (and higher dimensional objects) defined by polynomial equations. To understand affine varieties, we will need some algebra, and in particular, we will need to study *ideals* in the polynomial ring $k[x_1, \ldots, x_n]$. Finally, we will discuss polynomials in one variable to illustrate the role played by *algorithms.*

§1 Polynomials and Affine Space

To link algebra and geometry, we will study polynomials over a field. We all know what polynomials are, but the term *field* may be unfamiliar. The basic intuition is that a field is a set where one can define addition, subtraction, multiplication, and division with the usual properties. Standard examples are the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$, whereas the integers $\mathbb{Z}$ are not a field since division fails (3 and 2 are integers, but their quotient 3/2 is not). A formal definition of field may be found in Appendix A.

One reason that fields are important is that linear algebra works over *any* field. Thus, even if your linear algebra course restricted the scalars to lie in $\mathbb{R}$ or $\mathbb{C}$, most of the theorems and techniques you learned apply to an arbitrary field k. In this book, we will employ different fields for different purposes. The most commonly used fields will be:

- The rational numbers $\mathbb{Q}$: the field for most of our computer examples.
- The real numbers $\mathbb{R}$: the field for drawing pictures of curves and surfaces.
- The complex numbers $\mathbb{C}$: the field for proving many of our theorems.

On occasion, we will encounter other fields, such as fields of rational functions (which will be defined later). There is also a very interesting theory of finite fields—see the exercises for one of the simpler examples.

We can now define polynomials. The reader certainly is familiar with polynomials in one and two variables, but we will need to discuss polynomials in n variables $x_1, \ldots, x_n$ with coefficients in an arbitrary field k. We start by defining monomials.

Definition 1. *A* **monomial** *in $x_1, \ldots, x_n$ is a product of the form*

$$x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_n^{\alpha_n},$$

where all of the exponents $\alpha_1, \ldots, \alpha_n$ are nonnegative integers. The **total degree** *of this monomial is the sum $\alpha_1 + \cdots + \alpha_n$.*

We can simplify the notation for monomials as follows: let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be an n-tuple of nonnegative integers. Then we set

$$x^\alpha = x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdots x_n^{\alpha_n}.$$

When $\alpha = (0, \ldots, 0)$, note that $x^\alpha = 1$. We also let $|\alpha| = \alpha_1 + \cdots + \alpha_n$ denote the total degree of the monomial x^α.

Definition 2. *A* **polynomial** *f in $x_1, \ldots, x_n$ with coefficients in k is a finite linear combination (with coefficients in k) of monomials. We will write a polynomial f in the form*

$$f = \sum_\alpha a_\alpha x^\alpha, \quad a_\alpha \in k,$$

where the sum is over a finite number of n-tuples $\alpha = (\alpha_1, \ldots, \alpha_n)$. The set of all polynomials in $x_1, \ldots, x_n$ with coefficients in k is denoted $k[x_1, \ldots, x_n]$.

When dealing with polynomials in a small number of variables, we will usually dispense with subscripts. Thus, polynomials in one, two, and three variables lie in $k[x]$, $k[x, y]$ and $k[x, y, z]$, respectively. For example,

$$f = 2x^3y^2z + \frac{3}{2}y^3z^3 - 3xyz + y^2$$

is a polynomial in $\mathbb{Q}[x, y, z]$. We will usually use the letters *f, g, h, p, q, r* to refer to polynomials.

We will use the following terminology in dealing with polynomials.

Definition 3. *Let $f = \Sigma_\alpha a_\alpha x^\alpha$ be a polynomial in $k[x_1, \ldots, x_n]$.*

(i) *We call a_α the* **coefficient** *of the monomial x^α.*
(ii) *If $a_\alpha \neq 0$, then we call $a_\alpha x^\alpha$ a* **term** *of f.*
(iii) *The* **total degree** *of f, denoted* $\deg(f)$*, is the maximum $|\alpha|$ such that the coefficient a_α is nonzero.*

As an example, the polynomial $f = 2x^3y^2z + \frac{3}{2}y^3z^3 - 3xyz + y^2$ given above has four terms and total degree six. Note that there are two terms of maximal total degree, which is something that cannot happen for polynomials of one variable. In Chapter 2, we will study how to *order* the terms of a polynomial.

The sum and product of two polynomials is again a polynomial. We say that a polynomial *f divides* a polynomial g provided that $g = fh$ for some $h \in k[x_1, \ldots, x_n]$.

One can show that, under addition and multiplication, $k[x_1, \ldots, x_n]$ satisfies all of the field axioms except for the existence of multiplicative inverses (because, for example, $1/x_1$ is not a polynomial). Such a mathematical structure is called a commutative ring

(see Appendix A for the full definition), and for this reason we will refer to $k[x_1, \ldots, x_n]$ as a *polynomial ring*.

The next topic to consider is affine space.

Definition 4. *Given a field k and a positive integer n, we define the n-dimensional* **affine space** *over k to be the set*

$$k^n = \{(a_1, \ldots, a_n) : a_1, \ldots, a_n \in k\}.$$

For an example of affine space, consider the case $k = \mathbb{R}$. Here we get the familiar space $\mathbb{R}^n$ from calculus and linear algebra. In general, we call $k^1 = k$ the *affine line* and k^2 the *affine plane.*

Let us next see how polynomials relate to affine space. The key idea is that a polynomial $f = \Sigma_\alpha a_\alpha x^\alpha \in k[x_1, \ldots, x_n]$ gives a function

$$f : k^n \to k$$

defined as follows: given $(a_1, \ldots, a_n) \in k^n$, replace every x_i by a_i in the expression for f. Since all of the coefficients also lie in k, this operation gives an element $f(a_1, \ldots, a_n) \in k$. The ability to regard a polynomial as a function is what makes it possible to link algebra and geometry.

This dual nature of polynomials has some unexpected consequences. For example, the question "is $f = 0$?" now has two potential meanings: is f the zero polynomial?, which means that all of its coefficients a_α are zero, or is f the zero function?, which means that $f(a_1, \ldots, a_n) = 0$ for all $(a_1, \ldots, a_n) \in k^n$. The surprising fact is that these two statements are not equivalent in general. For an example of how they can differ, consider the set consisting of the two elements 0 and 1. In the exercises, we will see that this can be made into a field where $1 + 1 = 0$. This field is usually called $\mathbb{F}_2$. Now consider the polynomial $x^2 - x = x(x - 1) \in \mathbb{F}_2[x]$. Since this polynomial vanishes at 0 and 1, we have found a nonzero polynomial which gives the zero function on the affine space $\mathbb{F}_2^1$. Other examples will be discussed in the exercises.

However, as long as k is infinite, there is no problem.

Proposition 5. *Let k be an infinite field, and let $f \in k[x_1, \ldots, x_n]$. Then $f = 0$ in $k[x_1, \ldots, x_n]$ if and only if $f : k^n \to k$ is the zero function.*

Proof. One direction of the proof is obvious since the zero polynomial clearly gives the zero function. To prove the converse, we need to show that if $f(a_1, \ldots, a_n) = 0$ for all $(a_1, \ldots, a_n) \in k^n$, then f is the zero polynomial. We will use induction on the number of variables n.

When $n = 1$, it is well known that a nonzero polynomial in $k[x]$ of degree m has at most m distinct roots (we will prove this fact in Corollary 3 of §5). For our particular $f \in k[x]$, we are assuming $f(a) = 0$ for all $a \in k$. Since k is infinite, this means that f has infinitely many roots, and, hence, f must be the zero polynomial.

Now assume that the converse is true for $n - 1$, and let $f \in k[x_1, \ldots, x_n]$ be a polynomial that vanishes at all points of k^n. By collecting the various powers of x_n, we

can write f in the form

$$f = \sum_{i=0}^{N} g_i(x_1, \ldots, x_{n-1}) x_n^i,$$

where $g_i \in k[x_1, \ldots, x_{n-1}]$. We will show that each g_i is the zero polynomial in $n-1$ variables, which will force f to be the zero polynomial in $k[x_1, \ldots, x_n]$.

If we fix $(a_1, \ldots, a_{n-1}) \in k^{n-1}$, we get the polynomial $f(a_1, \ldots, a_{n-1}, x_n) \in k[x_n]$. By our hypothesis on f, this vanishes for every $a_n \in k$. It follows from the case $n = 1$ that $f(a_1, \ldots, a_{n-1}, x_n)$ is the zero polynomial in $k[x_n]$. Using the above formula for f, we see that the coefficients of $f(a_1, \ldots, a_{n-1}, x_n)$ are $g_i(a_1, \ldots, a_{n-1})$, and thus, $g_i(a_1, \ldots, a_{n-1}) = 0$ for all i. Since $(a_1, \ldots, a_{n-1})$ was arbitrarily chosen in k^{n-1}, it follows that each $g_i \in k[x_1, \ldots, x_{n-1}]$ gives the zero function on k^{n-1}. Our inductive assumption then implies that each g_i is the zero polynomial in $k[x_1, \ldots, x_{n-1}]$. This forces f to be the zero polynomial in $k[x_1, \ldots, x_n]$ and completes the proof of the proposition. □

Note that in the statement of Proposition 5, the assertion "$f = 0$ in $k[x_1, \ldots, x_n]$" means that f is the zero polynomial, i.e., that every coefficient of f is zero. Thus, we use the same symbol "0" to stand for the zero element of k and the zero polynomial in $k[x_1, \ldots, x_n]$. The context will make clear which one we mean.

As a corollary, we see that two polynomials over an infinite field are equal precisely when they give the same function on affine space.

Corollary 6. *Let k be an infinite field, and let $f, g \in k[x_1, \ldots, x_n]$. Then $f = g$ in $k[x_1, \ldots, x_n]$ if and only if $f : k^n \to k$ and $g : k^n \to k$ are the same function.*

Proof. To prove the nontrivial direction, suppose that $f, g \in k[x_1, \ldots, x_n]$ give the same function on k^n. By hypothesis, the polynomial $f - g$ vanishes at all points of k^n. Proposition 5 then implies that $f - g$ is the zero polynomial. This proves that $f = g$ in $k[x_1, \ldots, x_n]$. □

Finally, we need to record a special property of polynomials over the field of complex numbers $\mathbb{C}$.

Theorem 7. *Every nonconstant polynomial $f \in \mathbb{C}[x]$ has a root in $\mathbb{C}$.*

Proof. This is the Fundamental Theorem of Algebra, and proofs can be found in most introductory texts on complex analysis (although many other proofs are known). □

We say that a field k is *algebraically closed* if every nonconstant polynomial in $k[x]$ has a root in k. Thus $\mathbb{R}$ is not algebraically closed (what are the roots of $x^2 + 1$?), whereas the above theorem asserts that $\mathbb{C}$ is algebraically closed. In Chapter 4 we will prove a powerful generalization of Theorem 7 called the Hilbert Nullstellensatz.

EXERCISES FOR §1

1. Let $\mathbb{F}_2 = \{0, 1\}$, and define addition and multiplication by $0 + 0 = 1 + 1 = 0, 0 + 1 = 1 + 0 = 1, 0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0$ and $1 \cdot 1 = 1$. Explain why $\mathbb{F}_2$ is a field. (You need not check the associative and distributive properties, but you should verify the existence of identities and inverses, both additive and multiplicative.)
2. Let $\mathbb{F}_2$ be the field from Exercise 1.
   a. Consider the polynomial $g(x, y) = x^2y + y^2x \in \mathbb{F}_2[x, y]$. Show that $g(x, y) = 0$ for every $(x, y) \in \mathbb{F}_2^2$, and explain why this does not contradict Proposition 5.
   b. Find a nonzero polynomial in $\mathbb{F}_2[x, y, z]$ which vanishes at every point of $\mathbb{F}_2^3$. Try to find one involving all three variables.
   c. Find a nonzero polynomial in $\mathbb{F}_2[x_1, \ldots, x_n]$ which vanishes at every point of $\mathbb{F}_2^n$. Can you find one in which all of $x_1, \ldots, x_n$ appear?
3. (Requires abstract algebra). Let p be a prime number. The ring of integers modulo p is a field with p elements, which we will denote $\mathbb{F}_p$.
   a. Explain why $\mathbb{F}_p - \{0\}$ is a group under multiplication.
   b. Use Lagrange's Theorem to show that $a^{p-1} = 1$ for all $a \in \mathbb{F}_p - \{0\}$.
   c. Prove that $a^p = a$ for all $a \in \mathbb{F}_p$. Hint: Treat the cases $a = 0$ and $a \neq 0$ separately.
   d. Find a nonzero polynomial in $\mathbb{F}_p[x]$ which vanishes at every point of $\mathbb{F}_p$. Hint: Use part (c).
4. (Requires abstract algebra.) Let F be a finite field with q elements. Adapt the argument of Exercise 3 to prove that $x^q - x$ is a nonzero polynomial in $F[x]$ which vanishes at every point of F. This shows that Proposition 5 fails for *all* finite fields.
5. In the proof of Proposition 5, we took $f \in k[x_1, \ldots, x_n]$ and wrote it as a polynomial in x_n with coefficients in $k[x_1, \ldots, x_{n-1}]$. To see what this looks like in a specific case, consider the polynomial
$$f(x, y, z) = x^5y^2z - x^4y^3 + y^5 + x^2z - y^3z + xy + 2x - 5z + 3.$$
   a. Write f as a polynomial in x with coefficients in $k[y, z]$.
   b. Write f as a polynomial in y with coefficients in $k[x, z]$.
   c. Write f as a polynomial in z with coefficients in $k[x, y]$.
6. Inside of $\mathbb{C}^n$, we have the subset $\mathbb{Z}^n$, which consists of all points with integer coordinates.
   a. Prove that if $f \in \mathbb{C}[x_1, \ldots, x_n]$ vanishes at every point of $\mathbb{Z}^n$, then f is the zero polynomial. Hint: Adapt the proof of Proposition 5.
   b. Let $f \in \mathbb{C}[x_1, \ldots, x_n]$, and let M be the largest power of any variable that appears in f. Let $\mathbb{Z}^n_{M+1}$ be the set of points of $\mathbb{Z}^n$, all coordinates of which lie between 1 and $M + 1$. Prove that if f vanishes at all points of $\mathbb{Z}^n_{M+1}$, then f is the zero polynomial.

§2 Affine Varieties

We can now define the basic geometric object of the book.

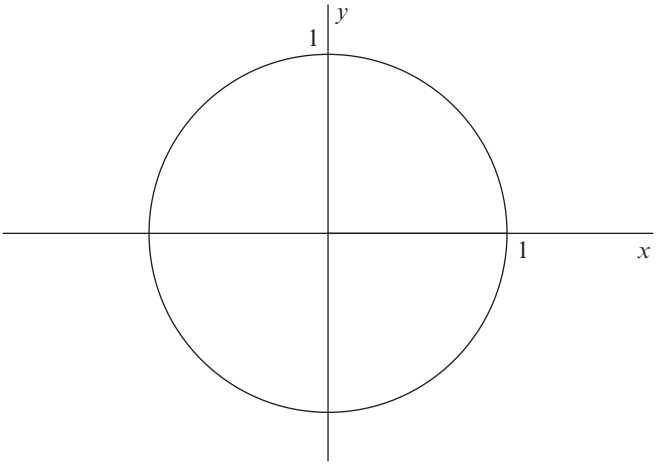
Definition 1. *Let k be a field, and let $f_1, \ldots, f_s$ be polynomials in $k[x_1, \ldots, x_n]$. Then we set*

$$\mathbf{V}(f_1, \ldots, f_s) = \{(a_1, \ldots, a_n) \in k^n : f_i(a_1, \ldots, a_n) = 0 \text{ for all } 1 \leq i \leq s\}.$$

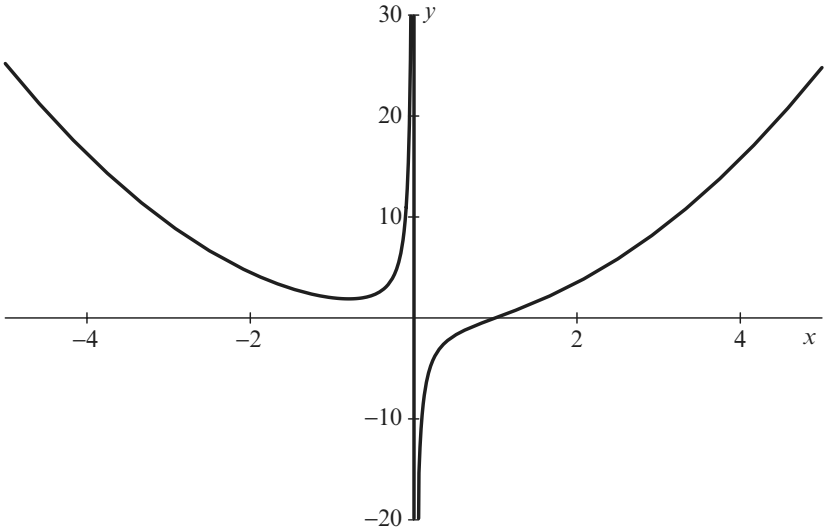
We call $\mathbf{V}(f_1, \ldots, f_s)$ the **affine variety** *defined by $f_1, \ldots, f_s$.*

Thus, an affine variety $\mathbf{V}(f_1, \ldots, f_s) \subset k^n$ is the set of all solutions of the system of equations $f_1(x_1, \ldots, x_n) = \cdots = f_s(x_1, \ldots, x_n) = 0$. We will use the letters *V*, *W*, etc. to denote affine varieties. The main purpose of this section is to introduce the reader to *lots* of examples, some new and some familiar. We will use $k = \mathbb{R}$ so that we can draw pictures.

We begin in the plane $\mathbb{R}^2$ with the variety $\mathbf{V}(x^2 + y^2 - 1)$, which is the circle of radius 1 centered at the origin:

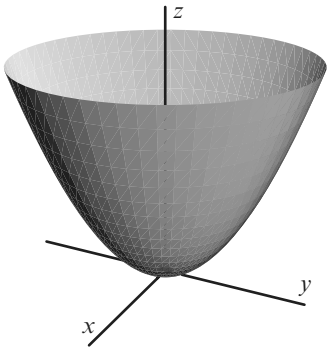


The conic sections studied in analytic geometry (circles, ellipses, parabolas, and hyperbolas) are affine varieties. Likewise, graphs of polynomial functions are affine varieties [the graph of $y = f(x)$ is $\mathbf{V}(y - f(x))$]. Although not as obvious, graphs of rational functions are also affine varieties. For example, consider the graph of $y = \frac{x^3-1}{x}$:

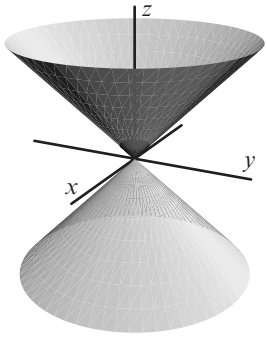


It is easy to check that this is the affine variety $\mathbf{V}(xy - x^3 + 1)$.

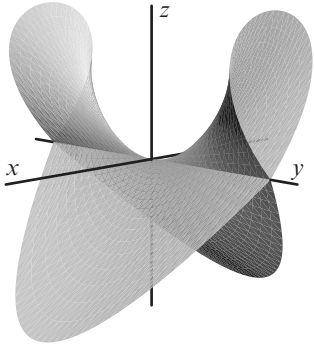
Next, let us look in the 3-dimensional space $\mathbb{R}^3$. A nice affine variety is given by paraboloid of revolution $\mathbf{V}(z - x^2 - y^2)$, which is obtained by rotating the parabola $z = x^2$ about the z-axis (you can check this using polar coordinates). This gives us the picture:



You may also be familiar with the cone $\mathbf{V}(z^2 - x^2 - y^2)$:

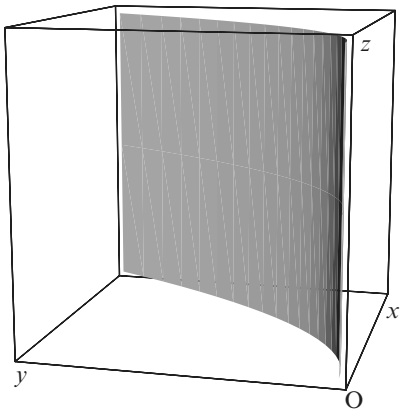


A much more complicated surface is given by $\mathbf{V}(x^2 - y^2z^2 + z^3)$:

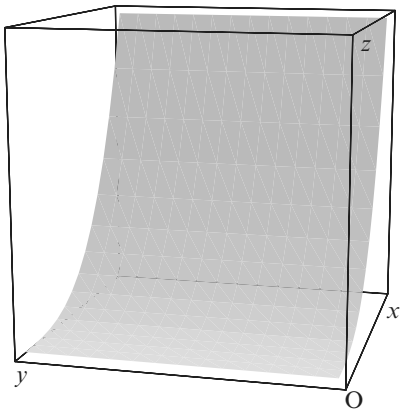


In these last two examples, the surfaces are not smooth everywhere: the cone has a sharp point at the origin, and the last example intersects itself along the whole y-axis. These are examples of *singular points*, which will be studied later in the book.

An interesting example of a curve in $\mathbb{R}^3$ is the *twisted cubic*, which is the variety $\mathbf{V}(y - x^2, z - x^3)$. For simplicity, we will confine ourselves to the portion that lies in the first octant. To begin, we draw the surfaces $y = x^2$ and $z = x^3$ separately:

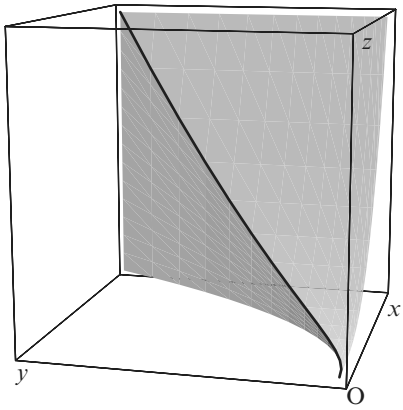


$y = x^2$



$z = x^3$

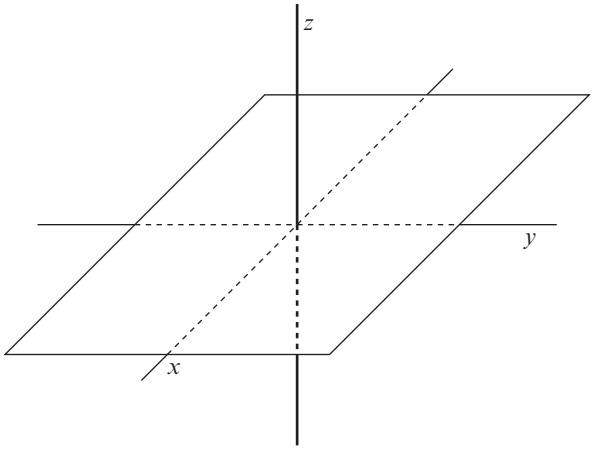
Then their intersection gives the twisted cubic:



The Twisted Cubic

Notice that when we had one equation in $\mathbb{R}^2$, we got a curve, which is a 1-dimensional object. A similar situation happens in $\mathbb{R}^3$: one equation in $\mathbb{R}^3$ usually gives a surface, which has dimension 2. Again, dimension drops by one. But now consider the twisted cubic: here, two equations in $\mathbb{R}^3$ give a curve, so that dimension drops by two. Since

each equation imposes an extra constraint, intuition suggests that each equation drops the dimension by one. Thus, if we started in $\mathbb{R}^4$, one would hope that an affine variety defined by two equations would be a surface. Unfortunately, the notion of dimension is more subtle than indicated by the above examples. To illustrate this, consider the variety $\mathbf{V}(xz, yz)$. One can easily check that the equations $xz = yz = 0$ define the union of the (x, y)-plane and the z-axis:



Hence, this variety consists of two pieces which have different dimensions, and one of the pieces (the plane) has the "wrong" dimension according to the above intuition.

We next give some examples of varieties in higher dimensions. A familiar case comes from linear algebra. Namely, fix a field k, and consider a system of m linear equations in n unknowns $x_1, \ldots, x_n$ with coefficients in k:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1, \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \tag{1}$$

The solutions of these equations form an affine variety in k^n, which we will call a *linear variety.* Thus, lines and planes are linear varieties, and there are examples of arbitrarily large dimension. In linear algebra, you learned the method of row reduction (also called Gaussian elimination), which gives an algorithm for finding all solutions of such a system of equations. In Chapter 2, we will study a generalization of this algorithm which applies to systems of polynomial equations.

Linear varieties relate nicely to our discussion of dimension. Namely, if $V \subset k^n$ is the linear variety defined by (1), then V need not have dimension $n - m$ even though V is defined by m equations. In fact, when V is nonempty, linear algebra tells us that V has dimension $n - r$, where r is the rank of the matrix (a_{ij}). So for linear varieties, the dimension is determined by the number of *independent* equations. This intuition applies to more general affine varieties, except that the notion of "independent" is more subtle.

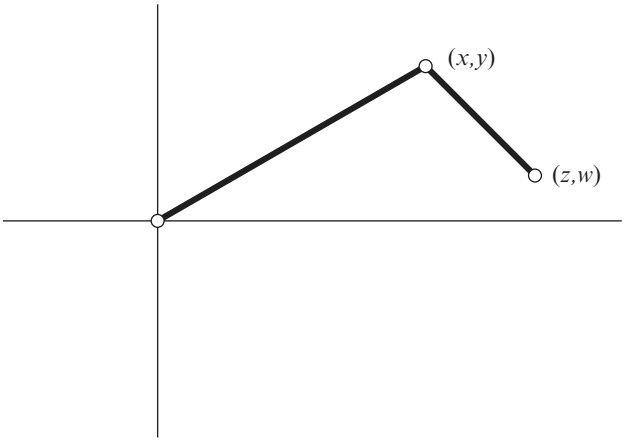
Some complicated examples in higher dimensions come from calculus. Suppose, for example, that we wanted to find the minimum and maximum values of $f(x, y, z) = x^3 + 2xyz - z^2$ subject to the constraint $g(x, y, z) = x^2 + y^2 + z^2 = 1$. The method of Lagrange multipliers states that $\nabla f = \lambda \nabla g$ at a local minimum or maximum [recall that the gradient of f is the vector of partial derivatives $\nabla f = (f_x, f_y, f_z)$]. This gives us the following system of four equations in four unknowns, x, y, z, λ, to solve:

$$\begin{aligned} 3x^2 + 2yz &= 2x\lambda, \\ 2xz &= 2y\lambda, \\ 2xy - 2z &= 2z\lambda, \\ x^2 + y^2 + z^2 &= 1. \end{aligned} \tag{2}$$

These equations define an affine variety in $\mathbb{R}^4$, and our intuition concerning dimension leads us to hope it consists of finitely many points (which have dimension 0) since it is defined by four equations. Students often find Lagrange multipliers difficult because the equations are so hard to solve. The algorithms of Chapter 2 will provide a powerful tool for attacking such problems. In particular, we will find all solutions of the above equations.

We should also mention that affine varieties can be the empty set. For example, when $k = \mathbb{R}$, it is obvious that $\mathbf{V}(x^2 + y^2 + 1) = \emptyset$ since $x^2 + y^2 = -1$ has no real solutions (although there are solutions when $k = \mathbb{C}$). Another example is $\mathbf{V}(xy, xy - 1)$, which is empty no matter what the field is, for a given x and y cannot satisfy both $xy = 0$ and $xy = 1$. In Chapter 4 we will study a method for determining when an affine variety over $\mathbb{C}$ is nonempty.

To give an idea of some of the applications of affine varieties, let us consider a simple example from robotics. Suppose we have a robot arm in the plane consisting of two linked rods of lengths 1 and 2, with the longer rod anchored at the origin:



The "state" of the arm is completely described by the coordinates (x, y) and (z, w) indicated in the figure. Thus the state can be regarded as a 4-tuple $(x, y, z, w) \in \mathbb{R}^4$.

However, not all 4-tuples can occur as states of the arm. In fact, it is easy to see that the subset of possible states is the affine variety in $\mathbb{R}^4$ defined by the equations

$$x^2 + y^2 = 4,$$
$$(x - z)^2 + (y - w)^2 = 1.$$

Notice how even larger dimensions enter quite easily: if we were to consider the same arm in 3-dimensional space, then the variety of states would be defined by two equations in $\mathbb{R}^6$. The techniques to be developed in this book have some important applications to the theory of robotics.

So far, all of our drawings have been over $\mathbb{R}$. Later in the book, we will consider varieties over $\mathbb{C}$. Here, it is more difficult (but not impossible) to get a geometric idea of what such a variety looks like.

Finally, let us record some basic properties of affine varieties.

Lemma 2. *If $V, W \subset k^n$ are affine varieties, then so are $V \cup W$ and $V \cap W$.*

Proof. Suppose that $V = \mathbf{V}(f_1, \ldots, f_s)$ and $W = \mathbf{V}(g_1, \ldots, g_t)$. Then we claim that

$$V \cap W = \mathbf{V}(f_1, \ldots, f_s, g_1, \ldots, g_t),$$
$$V \cup W = \mathbf{V}(f_i g_j : 1 \le i \le s, 1 \le j \le t).$$

The first equality is trivial to prove: being in $V \cap W$ means that both $f_1, \ldots, f_s$ and $g_1, \ldots, g_t$ vanish, which is the same as $f_1, \ldots, f_s, g_1, \ldots, g_t$ vanishing.

The second equality takes a little more work. If $(a_1, \ldots, a_n) \in V$, then all of the f_i's vanish at this point, which implies that all of the $f_i g_j$'s also vanish at $(a_1, \ldots, a_n)$. Thus, $V \subset \mathbf{V}(f_i g_j)$, and $W \subset \mathbf{V}(f_i g_j)$ follows similarly. This proves that $V \cup W \subset \mathbf{V}(f_i g_j)$. Going the other way, choose $(a_1, \ldots, a_n) \in \mathbf{V}(f_i g_j)$. If this lies in V, then we are done, and if not, then $f_{i_0}(a_1, \ldots, a_n) \neq 0$ for some i_0. Since $f_{i_0} g_j$ vanishes at $(a_1, \ldots, a_n)$ for all j, the g_j's must vanish at this point, proving that $(a_1, \ldots, a_n) \in W$. This shows that $\mathbf{V}(f_i g_j) \subset V \cup W$. □

This lemma implies that finite intersections and unions of affine varieties are again affine varieties. It turns out that we have already seen examples of unions and intersections. Concerning unions, consider the union of the (x, y)-plane and the z-axis in affine 3-space. By the above formula, we have

$$\mathbf{V}(z) \cup \mathbf{V}(x, y) = \mathbf{V}(zx, zy).$$

This, of course, is one of the examples discussed earlier in the section. As for intersections, notice that the twisted cubic was given as the intersection of two surfaces.

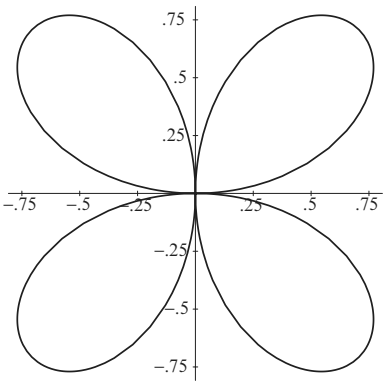
The examples given in this section lead to some interesting questions concerning affine varieties. Suppose that we have $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$. Then:

- (Consistency) Can we determine if $\mathbf{V}(f_1, \ldots, f_s) \neq \emptyset$, i.e., do the equations $f_1 = \cdots = f_s = 0$ have a common solution?
- (Finiteness) Can we determine if $\mathbf{V}(f_1, \ldots, f_s)$ is finite, and if so, can we find all of the solutions explicitly?
- (Dimension) Can we determine the "dimension" of $\mathbf{V}(f_1, \ldots, f_s)$?

The answer to these questions is yes, although care must be taken in choosing the field k that we work over. The hardest is the one concerning dimension, for it involves some sophisticated concepts. Nevertheless, we will give complete solutions to all three problems.

EXERCISES FOR §2

1. Sketch the following affine varieties in $\mathbb{R}^2$:
   a. $\mathbf{V}(x^2 + 4y^2 + 2x - 16y + 1)$.
   b. $\mathbf{V}(x^2 - y^2)$.
   c. $\mathbf{V}(2x + y - 1, 3x - y + 2)$.
   In each case, does the variety have the dimension you would intuitively expect it to have?
2. In $\mathbb{R}^2$, sketch $\mathbf{V}(y^2 - x(x-1)(x-2))$. Hint: For which x's is it possible to solve for y? How many y's correspond to each x? What symmetry does the curve have?
3. In the plane $\mathbb{R}^2$, draw a picture to illustrate
$$\mathbf{V}(x^2 + y^2 - 4) \cap \mathbf{V}(xy - 1) = \mathbf{V}(x^2 + y^2 - 4, xy - 1),$$
and determine the points of intersection. Note that this is a special case of Lemma 2.
4. Sketch the following affine varieties in $\mathbb{R}^3$:
   a. $\mathbf{V}(x^2 + y^2 + z^2 - 1)$.
   b. $\mathbf{V}(x^2 + y^2 - 1)$.
   c. $\mathbf{V}(x + 2, y - 1.5, z)$.
   d. $\mathbf{V}(xz^2 - xy)$. Hint: Factor $xz^2 - xy$.
   e. $\mathbf{V}(x^4 - zx, x^3 - yx)$.
   f. $\mathbf{V}(x^2 + y^2 + z^2 - 1, x^2 + y^2 + (z-1)^2 - 1)$.
   In each case, does the variety have the dimension you would intuitively expect it to have?
5. Use the proof of Lemma 2 to sketch $\mathbf{V}((x-2)(x^2 - y), y(x^2 - y), (z+1)(x^2 - y))$ in $\mathbb{R}^3$. Hint: This is the union of which two varieties?
6. Let us show that all finite subsets of k^n are affine varieties.
   a. Prove that a single point $(a_1, \ldots, a_n) \in k^n$ is an affine variety.
   b. Prove that every finite subset of k^n is an affine variety. Hint: Lemma 2 will be useful.
7. One of the prettiest examples from polar coordinates is the four-leaved rose



This curve is defined by the polar equation $r = \sin(2\theta)$. We will show that this curve is an affine variety.

a. Using $r^2 = x^2 + y^2$, $x = r\cos(\theta)$ and $y = r\sin(\theta)$, show that the four-leaved rose is contained in the affine variety $\mathbf{V}((x^2+y^2)^3 - 4x^2y^2)$. Hint: Use an identity for $\sin(2\theta)$.
b. Now argue carefully that $\mathbf{V}((x^2 + y^2)^3 - 4x^2y^2)$ is contained in the four-leaved rose. This is trickier than it seems since r can be negative in $r = \sin(2\theta)$.

Combining parts a and b, we have proved that the four-leaved rose is the affine variety $\mathbf{V}((x^2 + y^2)^3 - 4x^2y^2)$.

8. It can take some work to show that something is *not* an affine variety. For example, consider the set

$$X = \{(x, x) : x \in \mathbb{R},\ x \neq 1\} \subset \mathbb{R}^2,$$

which is the straight line $x = y$ with the point (1, 1) removed. To show that X is not an affine variety, suppose that $X = \mathbf{V}(f_1, \ldots, f_s)$. Then each f_i vanishes on X, and if we can show that f_i also vanishes at (1, 1), we will get the desired contradiction. Thus, here is what you are to prove: if $f \in \mathbb{R}[x, y]$ vanishes on X, then $f(1, 1) = 0$. Hint: Let $g(t) = f(t, t)$, which is a polynomial $\mathbb{R}[t]$. Now apply the proof of Proposition 5 of §1.
9. Let $R = \{(x, y) \in \mathbb{R}^2 : y > 0\}$ be the upper half plane. Prove that R is not an affine variety.
10. Let $\mathbb{Z}^n \subset \mathbb{C}^n$ consist of those points with integer coordinates. Prove that $\mathbb{Z}^n$ is not an affine variety. Hint: See Exercise 6 from §1.
11. So far, we have discussed varieties over $\mathbb{R}$ or $\mathbb{C}$. It is also possible to consider varieties over the field $\mathbb{Q}$, although the questions here tend to be *much* harder. For example, let n be a positive integer, and consider the variety $F_n \subset \mathbb{Q}^2$ defined by

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

Notice that there are some obvious solutions when x or y is zero. We call these *trivial solutions.* An interesting question is whether or not there are any nontrivial solutions.
a. Show that F_n has two trivial solutions if n is odd and four trivial solutions if n is even.
b. Show that F_n has a nontrivial solution for some $n \geq 3$ if and only if Fermat's Last Theorem were false.

Fermat's Last Theorem states that, for $n \geq 3$, the equation

$$x^n + y^n = z^n$$

has no solutions where x, y, and z are nonzero integers. The general case of this conjecture was proved by Andrew Wiles in 1994 using some very sophisticated number theory. The proof is *extremely* difficult.
12. Find a Lagrange multipliers problem in a calculus book and write down the corresponding system of equations. Be sure to use an example where one wants to find the minimum or maximum of a polynomial function subject to a polynomial constraint. This way the equations define an affine variety, and try to find a problem that leads to complicated equations. Later we will use Groebner basis methods to solve these equations.
13. Consider a robot arm in $\mathbb{R}^2$ that consists of three arms of lengths 3, 2, and 1, respectively. The arm of length 3 is anchored at the origin, the arm of length 2 is attached to the free end of the arm of length 3, and the arm of length 1 is attached to the free end of the arm of length 2. The "hand" of the robot arm is attached to the end of the arm of length 1.
a. Draw a picture of the robot arm.
b. How many variables does it take to determine the "state" of the robot arm?
c. Give the equations for the variety of possible states.
d. Using the intuitive notion of dimension discussed in this section, guess what the dimension of the variety of states should be.

14. This exercise will study the possible "hand" positions of the robot arm described in Exercise 13.
    a. If (u, v) is the position of the hand, explain why $u^2 + v^2 \leq 36$.
    b. Suppose we "lock" the joint between the length 3 and length 2 arms to form a straight angle, but allow the other joint to move freely. Draw a picture to show that in these configurations, (u, v) can be *any* point of the annulus $16 \leq u^2 + v^2 \leq 36$.
    c. Draw a picture to show that (u, v) can be any point in the disk $u^2 + v^2 \leq 36$. Hint: These positions can be reached by putting the second joint in a fixed, special position.
15. In Lemma 2, we showed that if V and W are affine varieties, then so are their union $V \cup W$ and intersection $V \cap W$. In this exercise we will study how other set-theoretic operations affect affine varieties.
    a. Prove that finite unions and intersections of affine varieties are again affine varieties. Hint: Induction.
    b. Give an example to show that an infinite union of affine varieties need not be an affine variety. Hint: By Exercises 8–10, we know some subsets of k^n that are not affine varieties. Surprisingly, an infinite intersection of affine varieties is still an affine variety. This is a consequence of the Hilbert Basis Theorem, which will be discussed in Chapters 2 and 4.
    c. Give an example to show that the set-theoretic difference $V - W$ of two affine varieties need not be an affine variety.
    d. Let $V \subset k^n$ and $W \subset k^m$ be two affine varieties, and let

$$V \times W = \{(x_1, \ldots, x_n, y_1, \ldots, y_m) \in k^{n+m} : (x_1, \ldots, x_n) \in V, (y_1, \ldots, y_m) \in W\}$$

    be their cartesian product. Prove that $V \times W$ is an affine variety in k^{n+m}. Hint: If V is defined by $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$, then we can regard $f_1, \ldots, f_s$ as polynomials in $k[x_1, \ldots, x_n, y_1, \ldots, y_m]$, and similarly for W. Show that this gives defining equations for the cartesian product.

§3 Parametrizations of Affine Varieties

In this section, we will discuss the problem of describing the points of an affine variety $\mathbf{V}(f_1, \ldots, f_s)$. This reduces to asking whether there is a way to "write down" the solutions of the system of polynomial equations $f_1 = \cdots = f_s = 0$. When there are finitely many solutions, the goal is simply to list them all. But what do we do when there are infinitely many? As we will see, this question leads to the notion of parametrizing an affine variety.

To get started, let us look at an example from linear algebra. Let the field be $\mathbb{R}$, and consider the system of equations

$$\begin{aligned} x + y + z &= 1, \\ x + 2y - z &= 3. \end{aligned} \tag{1}$$

Geometrically, this represents the line in $\mathbb{R}^3$ which is the intersection of the planes $x + y + z = 1$ and $x + 2y - z = 3$. It follows that there are infinitely many solutions. To describe the solutions, we use row operations on equations (1) to obtain the

equivalent equations

$$x + 3z = -1,$$
$$y - 2z = 2.$$

Letting $z = t$, where t is arbitrary, this implies that all solutions of (1) are given by

$$\begin{aligned} x &= -1 - 3t, \\ y &= 2 + 2t, \\ z &= t \end{aligned} \tag{2}$$

as t varies over $\mathbb{R}$. We call t a *parameter*, and (2) is, thus, a *parametrization* of the solutions of (1).

To see if the idea of parametrizing solutions can be applied to other affine varieties, let us look at the example of the unit circle

$$x^2 + y^2 = 1. \tag{3}$$

A common way to parametrize the circle is using trigonometric functions:

$$x = \cos(t),$$
$$y = \sin(t).$$

There is also a more algebraic way to parametrize this circle:

$$\begin{aligned} x &= \frac{1 - t^2}{1 + t^2}, \\ y &= \frac{2t}{1 + t^2}. \end{aligned} \tag{4}$$

You should check that the points defined by these equations lie on the circle (3). It is also interesting to note that this parametrization does not describe the whole circle: since $x = \frac{1-t^2}{1+t^2}$ can never equal -1, the point $(-1, 0)$ is not covered. At the end of the section, we will explain how this parametrization was obtained.

Notice that equations (4) involve quotients of polynomials. These are examples of *rational functions*, and before we can say what it means to parametrize a variety, we need to define the general notion of rational function.

Definition 1. *Let k be a field. A* **rational function** *in $t_1, \ldots, t_m$ with coefficients in k is a quotient f/g of two polynomials $f, g \in k[t_1, \ldots, t_m]$, where g is not the zero polynomial. Furthermore, two rational functions f/g and h/k are equal, provided that $kf = gh$ in $k[t_1, \ldots, t_m]$. Finally, the set of all rational functions in $t_1, \ldots, t_m$ with coefficients in k is denoted $k(t_1, \ldots, t_m)$.*

It is not difficult to show that addition and multiplication of rational functions are well defined and that $k(t_1, \ldots, t_m)$ is a field. We will assume these facts without proof.

Now suppose that we are given a variety $V = \mathbf{V}(f_1, \ldots, f_s) \subset k^n$. Then a *rational parametric representation* of V consists of rational functions $r_1, \ldots, r_n \in k(t_1, \ldots, t_m)$

such that the points given by

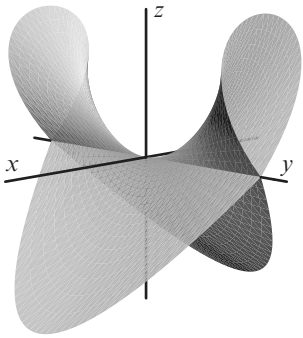
$$\begin{aligned} x_1 &= r_1(t_1, \ldots, t_m), \\ x_2 &= r_2(t_1, \ldots, t_m), \\ &\vdots \\ x_n &= r_n(t_1, \ldots, t_m) \end{aligned}$$

lie in V. We also require that V be the "smallest" variety containing these points. As the example of the circle shows, a parametrization may not cover all points of V. In Chapter 3, we will give a more precise definition of what we mean by "smallest."

In many situations, we have a parametrization of a variety V, where $r_1, \ldots, r_n$ are polynomials rather than rational functions. This is what we call a *polynomial parametric representation* of V.

By contrast, the original defining equations $f_1 = \cdots = f_s = 0$ of V are called an *implicit representation* of V. In our previous examples, note that equations (1) and (3) are implicit representations of varieties, whereas (2) and (4) are parametric.

One of the main virtues of a parametric representation of a curve or surface is that it is easy to draw on a computer. Given the formulas for the parametrization, the computer evaluates them for various values of the parameters and then plots the resulting points. For example, in §2 we viewed the surface $\mathbf{V}(x^2 - y^2z^2 + z^3)$:



This picture was not plotted using the implicit representation $x^2 - y^2z^2 + z^3 = 0$. Rather, we used the parametric representation given by

$$\begin{aligned} x &= t(u^2 - t^2), \\ y &= u, \\ z &= u^2 - t^2. \end{aligned} \tag{5}$$

There are two parameters t and u since we are describing a surface, and the above picture was drawn using t, u in the range $-1 \le t, u \le 1$. In the exercises, we will derive this parametrization and check that it covers the entire surface $\mathbf{V}(x^2-y^2z^2+z^3)$.

At the same time, it is often useful to have an implicit representation of a variety. For example, suppose we want to know whether or not the point $(1, 2, -1)$ is on the above surface. If all we had was the parametrization (5), then, to decide this question, we would need to solve the equations

$$\begin{aligned} 1 &= t(u^2 - t^2), \\ 2 &= u, \\ -1 &= u^2 - t^2 \end{aligned} \tag{6}$$

for t and u. On the other hand, if we have the implicit representation $x^2 - y^2z^2 + z^3 = 0$, then it is simply a matter of plugging into this equation. Since

$$1^2 - 2^2(-1)^2 + (-1)^3 = 1 - 4 - 1 = -4 \neq 0,$$

it follows that $(1, 2, -1)$ is not on the surface [and, consequently, equations (6) have no solution].

The desirability of having both types of representations leads to the following two questions:

- (Parametrization) Does every affine variety have a rational parametric representation?
- (Implicitization) Given a parametric representation of an affine variety, can we find the defining equations (i.e., can we find an implicit representation)?

The answer to the first question is no. In fact, most affine varieties cannot be parametrized in the sense described here. Those that can are called *unirational*. In general, it is difficult to tell whether a given variety is unirational or not. The situation for the second question is much nicer. In Chapter 3, we will see that the answer is always yes: given a parametric representation, we can always find the defining equations.

Let us look at an example of how implicitization works. Consider the parametric representation

$$\begin{aligned} x &= 1 + t, \\ y &= 1 + t^2. \end{aligned} \tag{7}$$

This describes a curve in the plane, but at this point, we cannot be sure that it lies on an affine variety. To find the equation we are looking for, notice that we can solve the first equation for t to obtain

$$t = x - 1.$$

Substituting this into the second equation yields

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

Hence the parametric equations (7) describe the affine variety $\mathbf{V}(y - x^2 + 2x - 2)$.

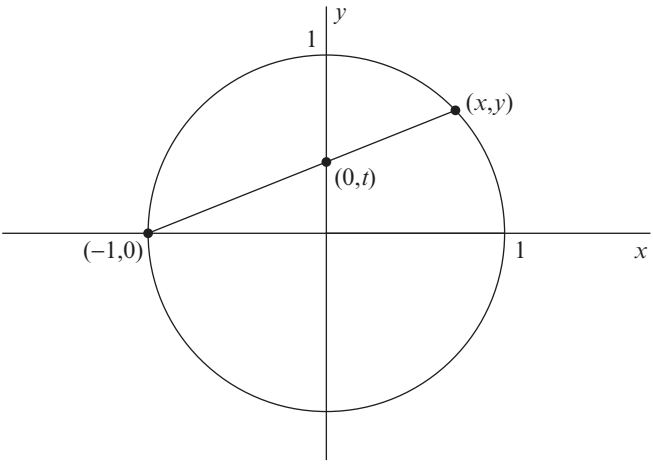
In the above example, notice that the basic strategy was to eliminate the variable t so that we were left with an equation involving only x and y. This illustrates the

role played by *elimination theory*, which will be studied in much greater detail in Chapter 3.

We will next discuss two examples of how geometry can be used to parametrize varieties. Let us start with the unit circle $x^2 + y^2 = 1$, which was parametrized in (4) via

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

To see where this parametrization comes from, notice that each nonvertical line through $(-1, 0)$ will intersect the circle in a unique point (x, y):



Each nonvertical line also meets the y-axis, and this is the point $(0, t)$ in the above picture.

This gives us a geometric parametrization of the circle: given t, draw the line connecting $(-1, 0)$ to $(0, t)$, and let (x, y) be the point where the line meets $x^2 + y^2 = 1$. Notice that the previous sentence really gives a parametrization: as t runs from $-\infty$ to ∞ on the vertical axis, the corresponding point (x, y) traverses all of the circle except for the point $(-1,0)$.

It remains to find explicit formulas for x and y in terms of t. To do this, consider the *slope* of the line in the above picture. We can compute the slope in two ways, using either the points $(-1, 0)$ and $(0, t)$, or the points $(-1, 0)$ and (x, y). This gives us the equation

$$\frac{t - 0}{0 - (-1)} = \frac{y - 0}{x - (-1)},$$

which simplifies to become

$$t = \frac{y}{x+1}.$$

Thus, $y = t(x+1)$. If we substitute this into $x^2 + y^2 = 1$, we get

$$x^2 + t^2(x+1)^2 = 1,$$

which gives the quadratic equation

$$(1+t^2)x^2 + 2t^2x + t^2 - 1 = 0. \tag{8}$$

This equation gives the x-coordinates of where the line meets the circle, and it is quadratic since there are two points of intersection. One of the points is -1, so that $x+1$ is a factor of (8). It is now easy to find the other factor, and we can rewrite (8) as

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

Since the x-coordinate we want is given by the second factor, we obtain

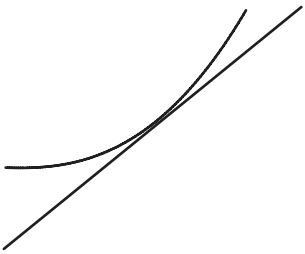
$$x = \frac{1-t^2}{1+t^2}.$$

Furthermore, $y = t(x+1)$ easily leads to

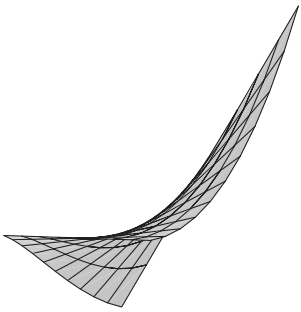
$$y = \frac{2t}{1+t^2}$$

(you should check this), and we have now derived the parametrization given earlier. Note how the geometry tells us exactly what portion of the circle is covered.

For our second example, let us consider the twisted cubic $\mathbf{V}(y - x^2, z - x^3)$ from §2. This is a curve in 3-dimensional space, and by looking at the tangent lines to the curve, we will get an interesting surface. The idea is as follows. Given one point on the curve, we can draw the tangent line at that point:



Now imagine taking the tangent lines for *all* points on the twisted cubic. This gives us the following surface:



This picture shows several of the tangent lines. The above surface is called the *tangent surface* of the twisted cubic.

To convert this geometric description into something more algebraic, notice that setting $x = t$ in $y - x^2 = z - x^3 = 0$ gives us a parametrization

$$\begin{aligned} x &= t, \\ y &= t^2, \\ z &= t^3 \end{aligned}$$

of the twisted cubic. We will write this as $\mathbf{r}(t) = (t, t^2, t^3)$. Now fix a particular value of t, which gives us a point on the curve. From calculus, we know that the tangent vector to the curve at the point given by $\mathbf{r}(t)$ is $\mathbf{r}'(t) = (1, 2t, 3t^2)$. It follows that the tangent line is parametrized by

$$\mathbf{r}(t) + u\mathbf{r}'(t) = (t, t^2, t^3) + u(1, 2t, 3t^2) = (t + u, t^2 + 2tu, t^3 + 3t^2u),$$

where u is a parameter that moves along the tangent line. If we now allow t to vary, then we can parametrize the entire tangent surface by

$$\begin{aligned} x &= t + u, \\ y &= t^2 + 2tu, \\ z &= t^3 + 3t^2u. \end{aligned}$$

The parameters t and u have the following interpretations: t tells where we are on the curve, and u tells where we are on the tangent line. This parametrization was used to draw the picture of the tangent surface presented earlier.

A final question concerns the implicit representation of the tangent surface: how do we find its defining equation? This is a special case of the implicitization problem mentioned earlier and is equivalent to eliminating t and u from the above parametric equations. In Chapters 2 and 3, we will see that there is an algorithm for doing this, and, in particular, we will prove that the tangent surface to the twisted cubic is defined by the equation

$$-4x^3z + 3x^2y^2 - 4y^3 + 6xyz - z^2 = 0.$$

We will end this section with an example from Computer Aided Geometric Design (CAGD). When creating complex shapes like automobile hoods or airplane wings, design engineers need curves and surfaces that are varied in shape, easy to describe, and quick to draw. Parametric equations involving polynomial and rational functions satisfy these requirements; there is a large body of literature on this topic.

For simplicity, let us suppose that a design engineer wants to describe a curve in the plane. Complicated curves are usually created by joining together simpler pieces, and for the pieces to join smoothly, the tangent directions must match up at the endpoints. Thus, for each piece, the designer needs to control the following geometric data:

- the starting and ending points of the curve;
- the tangent directions at the starting and ending points.

The *Bézier cubic*, introduced by Renault auto designer P. Bézier, is especially well suited for this purpose. A Bézier cubic is given parametrically by the equations

$$\begin{aligned} x &= (1-t)^3x_0 + 3t(1-t)^2x_1 + 3t^2(1-t)x_2 + t^3x_3, \\ y &= (1-t)^3y_0 + 3t(1-t)^2y_1 + 3t^2(1-t)y_2 + t^3y_3 \end{aligned} \tag{9}$$

for $0 \le t \le 1$, where $x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3$ are constants specified by the design engineer. We need to see how these constants correspond to the above geometric data.

If we evaluate the above formulas at $t = 0$ and $t = 1$, then we obtain

$$\begin{aligned} (x(0), y(0)) &= (x_0, y_0), \\ (x(1), y(1)) &= (x_3, y_3). \end{aligned}$$

As t varies from 0 to 1, equations (9) describe a curve starting at (x_0, y_0) and ending at (x_3, y_3). This gives us half of the needed data. We will next use calculus to find the tangent directions when $t = 0$ and 1. We know that the tangent vector to (9) when $t = 0$ is $(x'(0), y'(0))$. To calculate $x'(0)$, we differentiate the first line of (9) to obtain

$$x' = -3(1-t)^2x_0 + 3((1-t)^2 - 2t(1-t))x_1 + 3(2t(1-t) - t^2)x_2 + 3t^2x_3.$$

Then substituting $t = 0$ yields

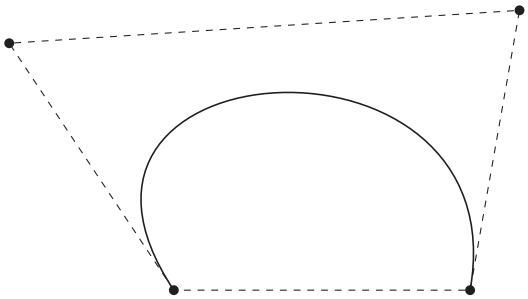
$$x'(0) = -3x_0 + 3x_1 = 3(x_1 - x_0),$$

and from here, it is straightforward to show that

$$\begin{aligned} (x'(0), y'(0)) &= 3(x_1 - x_0, y_1 - y_0), \\ (x'(1), y'(1)) &= 3(x_3 - x_2, y_3 - y_2). \end{aligned} \tag{10}$$

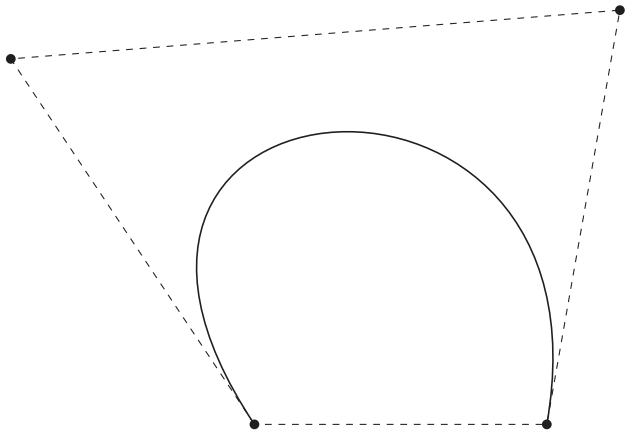
Since $(x_1 - x_0, y_1 - y_0) = (x_1, y_1) - (x_0, y_0)$, it follows that $(x'(0), y'(0))$ is three times the vector from (x_0, y_0) to (x_1, y_1). Hence, by placing (x_1, y_1), the designer can control the tangent direction at the beginning of the curve. In a similar way, the placement of (x_2, y_2) controls the tangent direction at the end of the curve.

The points (x_0, y_0), (x_1, y_1), (x_2, y_2) and (x_3, y_3) are called the *control points* of the Bézier cubic. They are usually labelled P_0, P_1, P_2 and P_3, and the convex quadrilateral they determine is called the *control polygon*. Here is a picture of a Bézier curve together with its control polygon:



In the exercises, we will show that a Bézier cubic always lies inside its control polygon.

The data determining a Bézier cubic is thus easy to specify and has a strong geometric meaning. One issue not resolved so far is the *length* of the tangent vectors $(x'(0), y'(0))$ and $(x'(1), y'(1))$. According to (10), it is possible to change the points (x_1, y_1) and (x_2, y_2) without changing the tangent directions. For example, if we keep the same directions as in the previous picture, but lengthen the tangent vectors, then we get the following curve:



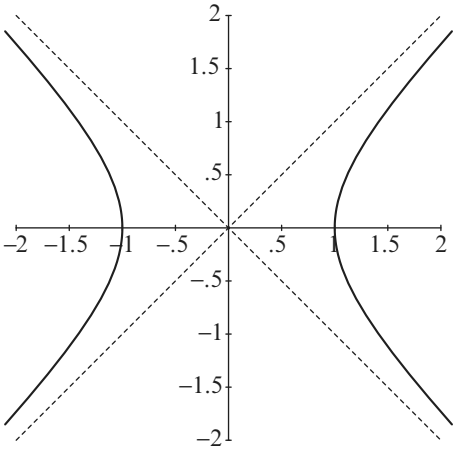
Thus, increasing the velocity at an endpoint makes the curve stay close to the tangent line for a longer distance. With practice and experience, a designer can become proficient in using Bézier cubics to create a wide variety of curves. It is interesting to note that the designer may never be aware of equations (9) that are used to describe the curve.

Besides CAGD, we should mention that Bézier cubics are also used in the page description language PostScript. The `curveto` command in PostScript has the coordinates of the control points as input and the Bézier cubic as output. This is how the above

Bézier cubics were drawn—each curve was specified by a single `curveto` instruction in a PostScript file.

EXERCISES FOR §3

1. Parametrize all solutions of the linear equations
$$x + 2y - 2z + w = -1,$$
$$x + y + z - w = 2.$$
2. Use a trigonometric identity to show that
$$x = \cos(t),$$
$$y = \cos(2t)$$
parametrizes a portion of a parabola. Indicate exactly what portion of the parabola is covered.
3. Given $f \in k[x]$, find a parametrization of $\mathbf{V}(y - f(x))$.
4. Consider the parametric representation
$$x = \frac{t}{1+t},$$
$$y = 1 - \frac{1}{t^2}.$$
   a. Find the equation of the affine variety determined by the above parametric equations.
   b. Show that the above equations parametrize all points of the variety found in part a except for the point (1,1).
5. This problem will be concerned with the hyperbola $x^2 - y^2 = 1$.



   a. Just as trigonometric functions are used to parametrize the circle, hyperbolic functions are used to parametrize the hyperbola. Show that the point
$$x = \cosh(t),$$
$$y = \sinh(t)$$
always lies on $x^2 - y^2 = 1$. What portion of the hyperbola is covered?

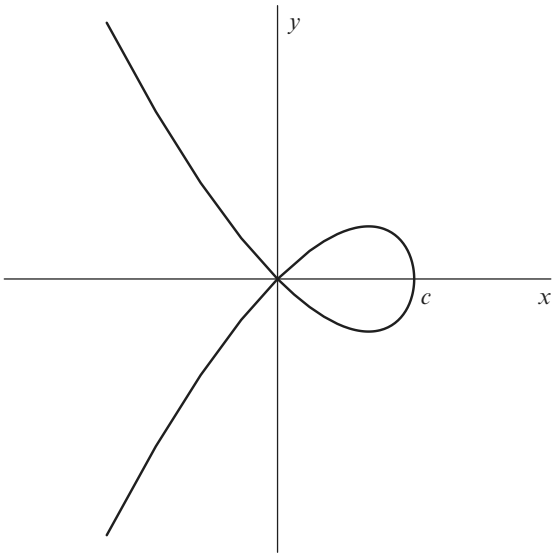
b. Show that a straight line meets a hyperbola in 0,1, or 2 points, and illustrate your answer with a picture. Hint: Consider the cases $x = a$ and $y = mx + b$ separately.
c. Adapt the argument given at the end of the section to derive a parametrization of the hyperbola. Hint: Consider nonvertical lines through the point $(-1,0)$ on the hyperbola.
d. The parametrization you found in part c is undefined for two values of t. Explain how this relates to the asymptotes of the hyperbola.

6. The goal of this problem is to show that the sphere $x^2 + y^2 + z^2 = 1$ in 3-dimensional space can be parametrized by

$$x = \frac{2u}{u^2 + v^2 + 1},$$
$$y = \frac{2v}{u^2 + v^2 + 1},$$
$$z = \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}.$$

The idea is to adapt the argument given at the end of the section to 3-dimensional space.
a. Given a point $(u, v, 0)$ in the xy-plane, draw the line from this point to the "north pole" (0,0,1) of the sphere, and let (x, y, z) be the other point where the line meets the sphere. Draw a picture to illustrate this, and argue geometrically that mapping (u, v) to (x, y, z) gives a parametrization of the sphere minus the north pole.
b. Show that the line connecting (0,0,1) to $(u, v, 0)$ is parametrized by $(tu, tv, 1-t)$, where t is a parameter that moves along the line.
c. Substitute $x = tu$, $y = tv$ and $z = 1 - t$ into the equation for the sphere $x^2 + y^2 + z^2 = 1$. Use this to derive the formulas given at the beginning of the problem.

7. Adapt the argument of the previous exercise to parametrize the "sphere" $x_1^2 + \cdots + x_n^2 = 1$ in n-dimensional affine space. Hint: There will be $n - 1$ parameters.
8. Consider the curve defined by $y^2 = cx^2 - x^3$, where c is some constant. Here is a picture of the curve when $c > 0$:



Our goal is to parametrize this curve.

a. Show that a line will meet this curve at either 0, 1, 2, or 3 points. Illustrate your answer with a picture. Hint: Let the equation of the line be either $x = a$ or $y = mx + b$.
b. Show that a nonvertical line through the origin meets the curve at exactly one other point when $m^2 \neq c$. Draw a picture to illustrate this, and see if you can come up with an intuitive explanation as to why this happens.
c. Now draw the vertical line $x = 1$. Given a point $(1, t)$ on this line, draw the line connecting $(1, t)$ to the origin. This will intersect the curve in a point (x, y). Draw a picture to illustrate this, and argue geometrically that this gives a parametrization of the entire curve.
d. Show that the geometric description from part (c) leads to the parametrization

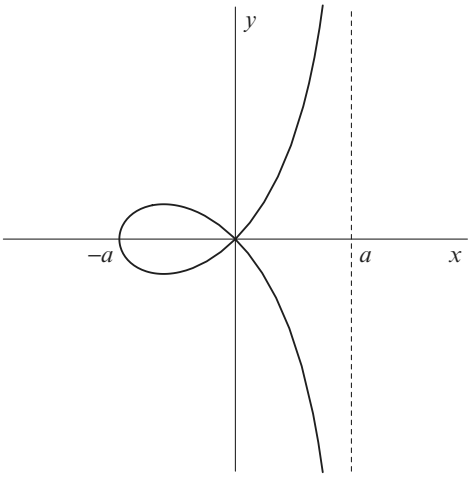
$$x = c - t^2,$$
$$y = t(c - t^2).$$

9. The *strophoid* is a curve that was studied by various mathematicians, including Isaac Barrow (1630–1677), Jean Bernoulli (1667–1748), and Maria Agnesi (1718–1799). A trigonometric parametrization is given by

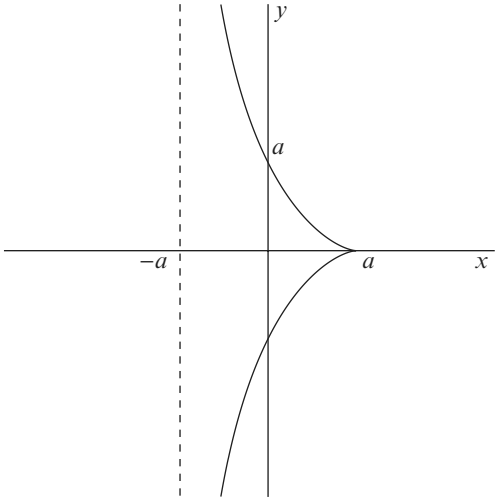
$$x = a\sin(t),$$
$$y = a\tan(t)(1 + \sin(t))$$

where a is a constant. If we let t vary in the range $-4.5 \leq t \leq 1.5$, we get the picture shown below.

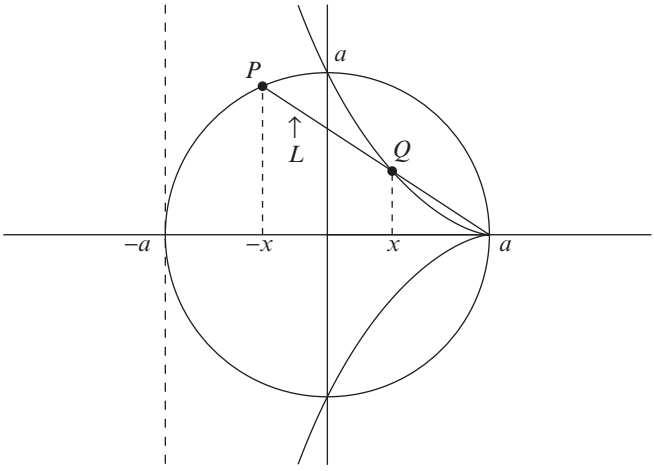
a. Find the equation in x and y that describes the strophoid. Hint: If you are sloppy, you will get the equation $(a^2 - x^2)y^2 = x^2(a + x)^2$. To see why this is not quite correct, see what happens when $x = -a$.
b. Find an algebraic parametrization of the strophoid.



10. Around 180 B.C., Diocles wrote the book *On Burning-Glasses*, and one of the curves he considered was the *cissoid*. He used this curve to solve the problem of the duplication of the cube [see part (c) below]. The cissoid has the equation $y^2(a + x) = (a - x)^3$, where a is a constant. This gives the following curve in the plane:



a. Find an algebraic parametrization of the cissoid.
b. Diocles described the cissoid using the following geometric construction. Given a circle of radius a (which we will take as centered at the origin), pick x between a and $-a$, and draw the line L connecting $(a, 0)$ to the point $P = (-x, \sqrt{a^2 - x^2})$ on the circle. This determines a point $Q = (x, y)$ on L:



Prove that the cissoid is the locus of all such points Q.
c. The duplication of the cube is the classical Greek problem of trying to construct $\sqrt[3]{2}$ using ruler and compass. It is known that this is impossible given just a ruler and compass. Diocles showed that if in addition, you allow the use of the cissoid, then one can construct $\sqrt[3]{2}$. Here is how it works. Draw the line connecting $(-a, 0)$ to $(0, a/2)$. This line will

meet the cissoid at a point (x, y). Then prove that

$$2 = \left(\frac{a-x}{y}\right)^3,$$

which shows how to construct $\sqrt[3]{2}$ using ruler, compass and cissoid.

11. In this problem, we will derive the parametrization

$$\begin{aligned} x &= t(u^2 - t^2), \\ y &= u, \\ z &= u^2 - t^2, \end{aligned}$$

of the surface $x^2 - y^2z^2 + z^3 = 0$ considered in the text.
   a. Adapt the formulas in part (d) of Exercise 8 to show that the curve $x^2 = cz^2 - z^3$ is parametrized by

$$\begin{aligned} z &= c - t^2, \\ x &= t(c - t^2). \end{aligned}$$

   b. Now replace the c in part a by y^2, and explain how this leads to the above parametrization of $x^2 - y^2z^2 + z^3 = 0$.
   c. Explain why this parametrization covers the entire surface $\mathbf{V}(x^2 - y^2z^2 + z^3)$. Hint: See part (c) of Exercise 8.

12. Consider the variety $V = \mathbf{V}(y - x^2, z - x^4) \subset \mathbb{R}^3$.
   a. Draw a picture of V.
   b. Parametrize V in a way similar to what we did with the twisted cubic.
   c. Parametrize the tangent surface of V.

13. The general problem of finding the equation of a parametrized surface will be studied in Chapters 2 and 3. However, when the surface is a plane, methods from calculus or linear algebra can be used. For example, consider the plane in $\mathbb{R}^3$ parametrized by

$$\begin{aligned} x &= 1 + u - v, \\ y &= u + 2v, \\ z &= -1 - u + v. \end{aligned}$$

Find the equation of the plane determined this way. Hint: Let the equation of the plane be $ax + by + cz = d$. Then substitute in the above parametrization to obtain a system of equations for a, b, c, d. Another way to solve the problem would be to write the parametrization in vector form as $(1, 0, -1) + u(1, 1, -1) + v(-1, 2, 1)$. Then one can get a quick solution using the cross product.

14. This problem deals with convex sets and will be used in the next exercise to show that a Bézier cubic lies within its control polygon. A subset $C \subset \mathbb{R}^2$ is *convex* if for all $P, Q \in C$, the line segment joining P to Q also lies in C.
   a. If $P = \binom{x}{y}$ and $Q = \binom{z}{w}$ lie in a convex set C, then show that

$$t\begin{pmatrix} x \\ y \end{pmatrix} + (1-t)\begin{pmatrix} z \\ w \end{pmatrix} \in C$$

when $0 \le t \le 1$.

b. If $P_i = \binom{x_i}{y_i}$ lies in a convex set C for $1 \le i \le n$, then show that

$$\sum_{i=1}^{n} t_i \binom{x_i}{y_i} \in C$$

wherever $t_1, \ldots, t_n$ are nonnegative numbers such that $\sum_{i=1}^{n} t_i = 1$. Hint: Use induction on n.

15. Let a Bézier cubic be given by

$$x = (1-t)^3 x_0 + 3t(1-t)^2 x_1 + 3t^2(1-t)x_2 + t^3 x_3,$$
$$y = (1-t)^3 y_0 + 3t(1-t)^2 y_1 + 3t^2(1-t)y_2 + t^3 y_3.$$

a. Show that the above equations can be written in vector form

$$\binom{x}{y} = (1-t)^3 \binom{x_0}{y_0} + 3t(1-t)^2 \binom{x_1}{y_1} + 3t^2(1-t)\binom{x_2}{y_2} + t^3 \binom{x_3}{y_3}.$$

b. Use the previous exercise to show that a Bézier cubic always lies inside its control polygon. Hint: In the above equations, what is the sum of the coefficients?

16. One disadvantage of Bézier cubics is that curves like circles and hyperbolas cannot be described exactly by cubics. In this exercise, we will discuss a method similar to example (4) for parametrizing conic sections. Our treatment is based on BALL (1987).

A *conic section* is a curve in the plane defined by a second degree equation of the form $ax^2 + bxy + cy^2 + dx + ey + f = 0$. Conic sections include the familiar examples of circles, ellipses, parabolas, and hyperbolas. Now consider the curve parametrized by

$$x = \frac{(1-t)^2 x_1 + 2t(1-t)w x_2 + t^2 x_3}{(1-t)^2 + 2t(1-t)w + t^2},$$
$$y = \frac{(1-t)^2 y_1 + 2t(1-t)w y_2 + t^2 y_3}{(1-t)^2 + 2t(1-t)w + t^2}$$

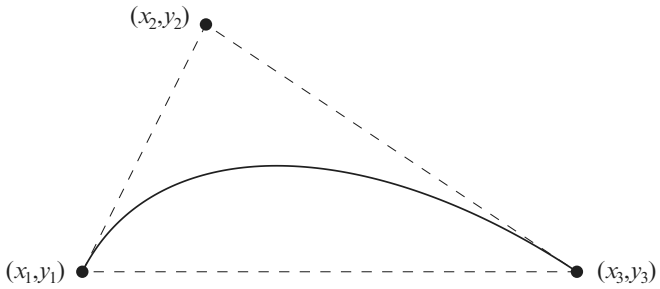
for $0 \le t \le 1$. The constants $w, x_1, y_1, x_2, y_2, x_3, y_3$ are specified by the design engineer, and we will assume that $w \ge 0$. In Chapter 3, we will show that these equations parametrize a conic section. The goal of this exercise is to give a geometric interpretation for the quantities $w, x_1, y_1, x_2, y_2, x_3, y_3$.

a. Show that our assumption $w \ge 0$ implies that the denominator in the above formulas never vanishes.

b. Evaluate the above formulas at $t = 0$ and $t = 1$. This should tell you what x_1, y_1, x_3, y_3 mean.

c. Now compute $(x'(0), y'(0))$ and $(x'(1), y'(1))$. Use this to show that (x_2, y_2) is the intersection of the tangent lines at the start and end of the curve. Explain why (x_1, y_1), (x_2, y_2), and (x_3, y_3) are called the *control points* of the curve.

d. Define the *control polygon* (it is actually a triangle in this case), and prove that the curve defined by the above equations always lies in its control polygon. Hint: Adapt the argument of the previous exercise. This gives the following picture:

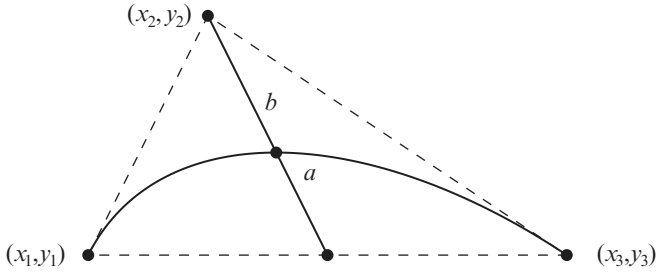


It remains to explain the constant w, which is called the *shape factor*. A hint should come from the answer to part (c), for note that w appears in the formulas for the tangent vectors when $t = 0$ and 1. So w somehow controls the "velocity," and a larger w should force the curve closer to (x_2, y_2). In the last two parts of the problem, we will determine exactly what w does.

e. Prove that

$$\begin{pmatrix} x\left(\frac{1}{2}\right) \\ y\left(\frac{1}{2}\right) \end{pmatrix} = \frac{1}{1+w}\left(\frac{1}{2}\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} x_3 \\ y_3 \end{pmatrix}\right) + \frac{w}{1+w}\begin{pmatrix} x_2 \\ y_2 \end{pmatrix}.$$

Use this formula to show that $\left(x\left(\frac{1}{2}\right), y\left(\frac{1}{2}\right)\right)$ lies on the line segment connecting (x_2, y_2) to the midpoint of the line between (x_1, y_1) and (x_3, y_3).



f. Notice that $\left(x\left(\frac{1}{2}\right), y\left(\frac{1}{2}\right)\right)$ divides this line segment into two pieces, say of lengths a and b as indicated in the above picture. Then prove that

$$w = \frac{a}{b},$$

so that w tells us exactly where the curve crosses this line segment. Hint: Use the distance formula.

17. Use the formulas of the previous exercise to parametrize the arc of the circle $x^2 + y^2 = 1$ from (1, 0) to (0, 1). Hint: Use part (f) of Exercise 16 to show that $w = 1/\sqrt{2}$.

§4 Ideals

We next define the basic algebraic object of the book.

Definition 1. *A subset $I \subset k[x_1, \ldots, x_n]$ is an* **ideal** *if it satisfies:*

(i) $0 \in I$.
(ii) *If $f, g \in I$, then $f + g \in I$.*
(iii) *If $f \in I$ and $h \in k[x_1, \ldots, x_n]$, then $hf \in I$.*

The goal of this section is to introduce the reader to some naturally occurring ideals and to see how ideals relate to affine varieties. The real importance of ideals is that they will give us a language for computing with affine varieties.

The first natural example of an ideal is the ideal generated by a finite number of polynomials.

Definition 2. *Let $f_1, \ldots, f_s$ be polynomials in $k[x_1, \ldots, x_n]$. Then we set*

$$\langle f_1, \ldots, f_s \rangle = \left\{ \sum_{i=1}^{s} h_i f_i : h_1, \ldots, h_s \in k[x_1, \ldots, x_n] \right\}.$$

The crucial fact is that $\langle f_1, \ldots, f_s \rangle$ is an ideal.

Lemma 3. *If $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$, then $\langle f_1, \ldots, f_s \rangle$ is an ideal of $k[x_1, \ldots, x_n]$. We will call $\langle f_1, \ldots, f_s \rangle$ the* **ideal generated by** *$f_1, \ldots, f_s$.*

Proof. First, $0 \in \langle f_1, \ldots, f_s \rangle$ since $0 = \sum_{i=1}^{s} 0 \cdot f_i$. Next, suppose that $f = \sum_{i=1}^{s} p_i f_i$ and $g = \sum_{i=1}^{s} q_i f_i$, and let $h \in k[x_1, \ldots, x_n]$. Then the equations

$$f + g = \sum_{i=1}^{s} (p_i + q_i) f_i,$$
$$hf = \sum_{i=1}^{s} (hp_i) f_i$$

complete the proof that $\langle f_1, \ldots, f_s \rangle$ is an ideal. □

The ideal $\langle f_1, \ldots, f_s \rangle$ has a nice interpretation in terms of polynomial equations. Given $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$, we get the system of equations

$$f_1 = 0,$$
$$\vdots$$
$$f_s = 0.$$

From these equations, one can derive others using algebra. For example, if we multiply the first equation by $h_1 \in k[x_1, \ldots, x_n]$, the second by $h_2 \in k[x_1, \ldots, x_n]$, etc., and then add the resulting equations, we obtain

$$h_1 f_1 + h_2 f_2 + \cdots + h_s f_s = 0,$$

which is a consequence of our original system. Notice that the left-hand side of

this equation is exactly an element of the ideal $\langle f_1, \ldots, f_s \rangle$. Thus, we can think of $\langle f_1, \ldots, f_s \rangle$ as consisting of all "polynomial consequences" of the equations $f_1 = f_2 = \cdots = f_s = 0$.

To see what this means in practice, consider the example from §3 where we took

$$\begin{aligned} x &= 1 + t, \\ y &= 1 + t^2 \end{aligned}$$

and eliminated t to obtain

$$y = x^2 - 2x + 2$$

[see the discussion following equation (7) in §3]. Let us redo this example using the above ideas. We start by writing the equations as

$$\begin{aligned} x - 1 - t &= 0, \\ y - 1 - t^2 &= 0. \end{aligned} \tag{1}$$

To cancel the t terms, we multiply the first equation by $x - 1 + t$ and the second by -1:

$$\begin{aligned} (x-1)^2 - t^2 &= 0, \\ -y + 1 + t^2 &= 0, \end{aligned}$$

and then add to obtain

$$(x-1)^2 - y + 1 = x^2 - 2x + 2 - y = 0.$$

In terms of the ideal generated by equations (1), we can write this as

$$\begin{aligned} x^2 - 2x + 2 - y &= (x - 1 + t)(x - 1 - t) + (-1)(y - 1 - t^2) \\ &\in \langle x - 1 - t, y - 1 - t^2 \rangle. \end{aligned}$$

Similarly, any other "polynomial consequence" of (1) leads to an element of this ideal.

We say that an ideal I is *finitely generated* if there exist $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$ such that $I = \langle f_1, \ldots, f_s \rangle$, and we say that $f_1, \ldots, f_s$, are a *basis* of I. In Chapter 2, we will prove the amazing fact that *every* ideal of $k[x_1, \ldots, x_n]$ is finitely generated (this is known as the Hilbert Basis Theorem). Note that a given ideal may have many different bases. In Chapter 2, we will show that one can choose an especially useful type of basis, called a Groebner basis.

There is a nice analogy with linear algebra that can be made here. The definition of an ideal is similar to the definition of a subspace: both have to be closed under addition and multiplication, except that, for a subspace, we multiply by scalars, whereas for an ideal, we multiply by polynomials. Further, notice that the ideal generated by polynomials $f_1, \ldots, f_s$ is similar to the span of a finite number of vectors $v_1, \ldots, v_s$. In each case, one takes linear combinations, using field coefficients for the span and polynomial coefficients for the ideal generated. Relations with linear algebra are explored further in Exercise 6.

Another indication of the role played by ideals is the following proposition, which shows that a variety depends only on the *ideal* generated by its defining equations.

Proposition 4. *If $f_1, \dots, f_s$ and $g_1, \dots, g_t$ are bases of the same ideal in $k[x_1, \dots, x_n]$, so that $\langle f_1, \dots, f_s \rangle = \langle g_1, \dots, g_t \rangle$, then we have $\mathbf{V}(f_1, \dots, f_s) = \mathbf{V}(g_1, \dots, g_t)$.*

Proof. The proof is very straightforward and is left as an exercise. □

As an example, consider the variety $\mathbf{V}(2x^2 + 3y^2 - 11, x^2 - y^2 - 3)$. It is easy to show that $\langle 2x^2 + 3y^2 - 11, x^2 - y^2 - 3 \rangle = \langle x^2 - 4, y^2 - 1 \rangle$ (see Exercise 3), so that

$$\mathbf{V}(2x^2 + 3y^2 - 11, x^2 - y^2 - 3) = \mathbf{V}(x^2 - 4, y^2 - 1) = \{(\pm 2, \pm 1)\}$$

by the above proposition. Thus, by changing the basis of the ideal, we made it easier to determine the variety.

The ability to change the basis without affecting the variety is very important. Later in the book, this will lead to the observation that affine varieties are determined by *ideals*, not equations. (In fact, the correspondence between ideals and varieties is the main topic of Chapter 4.) From a more practical point of view, we will also see that Proposition 4, when combined with the Groebner bases mentioned above, provides a powerful tool for understanding affine varieties.

We will next discuss how affine varieties give rise to an interesting class of ideals. Suppose we have an affine variety $V = \mathbf{V}(f_1, \dots, f_s) \subset k^n$ defined by $f_1, \dots, f_s \in k[x_1, \dots, x_n]$. We know that $f_1, \dots, f_s$ vanish on V, but are these the only ones? Are there other polynomials that vanish on V? For example, consider the twisted cubic studied in §2. This curve is defined by the vanishing of $y - x^2$ and $z - x^3$. From the parametrization (t, t^2, t^3) discussed in §3, we see that $z - xy$ and $y^2 - xz$ are two more polynomials that vanish on the twisted cubic. Are there other such polynomials? How do we find them all?

To study this question, we will consider the set of *all* polynomials that vanish on a given variety.

Definition 5. *Let $V \subset k^n$ be an affine variety. Then we set*

$$\mathbf{I}(V) = \{f \in k[x_1, \dots, x_n] : f(a_1, \dots, a_n) = 0 \text{ for all } (a_1, \dots, a_n) \in V\}.$$

The crucial observation is that $\mathbf{I}(V)$ is an ideal.

Lemma 6. *If $V \subset k^n$ is an affine variety, then $\mathbf{I}(V) \subset k[x_1, \dots, x_n]$ is an ideal. We will call $\mathbf{I}(V)$ the* **ideal of** *V.*

Proof. It is obvious that $0 \in \mathbf{I}(V)$ since the zero polynomial vanishes on all of k^n, and so, in particular it vanishes on V. Next, suppose that $f, g \in \mathbf{I}(V)$ and $h \in k[x_1, \dots, x_n]$.

Let $(a_1, \ldots, a_n)$ be an arbitrary point of V. Then

$$f(a_1, \ldots, a_n) + g(a_1, \ldots, a_n) = 0 + 0 = 0,$$
$$h(a_1, \ldots, a_n) f(a_1, \ldots, a_n) = h(a_1, \ldots, a_n) \cdot 0 = 0,$$

and it follows that $\mathbf{I}(V)$ is an ideal. □

For an example of the ideal of a variety, consider the variety $\{(0, 0)\}$ consisting of the origin in k^2. Then its ideal $\mathbf{I}(\{(0, 0)\})$ consists of all polynomials that vanish at the origin, and we claim that

$$\mathbf{I}(\{(0, 0)\}) = \langle x, y \rangle.$$

One direction of proof is trivial, for any polynomial of the form $A(x, y)x + B(x, y)y$ obviously vanishes at the origin. Going the other way, suppose that $f = \sum_{i,j} a_{ij} x^i y^j$ vanishes at the origin. Then $a_{00} = f(0, 0) = 0$ and, consequently,

$$\begin{aligned} f &= a_{00} + \sum_{i,j \neq 0,0} a_{ij} x^i y^j \\ &= 0 + \left(\sum_{\substack{i,j \\ i>0}} a_{ij} x^{i-1} y^j \right) x + \left(\sum_{j>0} a_{0j} y^{j-1} \right) y \in \langle x, y \rangle. \end{aligned}$$

Our claim is now proved.

For another example, consider the case when V is all of k^n. Then $\mathbf{I}(k^n)$ consists of polynomials that vanish everywhere, and, hence, by Proposition 5 of §1, we have

$$\mathbf{I}(k^n) = \{0\} \quad \text{when } k \text{ is infinite.}$$

(Here, "0" denotes the zero polynomial in $k[x_1, \ldots, x_n]$.) Note that Proposition 5 of §1 is equivalent to the above statement. In the exercises, we will discuss what happens when k is a finite field.

For a more interesting example, consider the twisted cubic $V = \mathbf{V}(y - x^2, z - x^3) \subset \mathbb{R}^3$. We claim that

$$\mathbf{I}(V) = \langle y - x^2, z - x^3 \rangle.$$

To prove this, we will first show that given a polynomial $f \in \mathbb{R}[x, y, z]$, we can write f in the form

$$f = h_1(y - x^2) + h_2(z - x^3) + r, \tag{2}$$

where $h_1, h_2 \in \mathbb{R}[x, y, z]$ and r is a polynomial in the variable x alone. First, consider the case when f is a monomial $x^\alpha y^\beta z^\gamma$. Then the binomial theorem tells us that

$$\begin{aligned} x^\alpha y^\beta z^\gamma &= x^\alpha (x^2 + (y - x^2))^\beta (x^3 + (z - x^3))^\gamma \\ &= x^\alpha (x^{2\beta} + \text{terms involving } y - x^2)(x^{3\gamma} + \text{terms involving } z - x^3), \end{aligned}$$

and multiplying this out shows that

$$x^\alpha y^\beta z^\gamma = h_1(y - x^2) + h_2(z - x^3) + x^{\alpha + 2\beta + 3\gamma}$$

for some polynomials $h_1, h_2 \in \mathbb{R}[x, y, z]$. Thus, (2) is true in this case. Since an arbitrary $f \in \mathbb{R}[x, y, z]$ is an $\mathbb{R}$-linear combination of monomials, it follows that (2) holds in general.

We can now prove $\mathbf{I}(V) = \langle y - x^2, z - x^3 \rangle$. First, by the definition of the twisted cubic V, we have $y - x^2, z - x^3 \in \mathbf{I}(V)$, and since $\mathbf{I}(V)$ is an ideal, it follows that $h_1(y - x^2) + h_2(z - x^3) \in \mathbf{I}(V)$. This proves that $\langle y - x^2, z - x^3 \rangle \subset \mathbf{I}(V)$. To prove the opposite inclusion, let $f \in \mathbf{I}(V)$ and let

$$f = h_1(y - x^2) + h_2(z - x^3) + r$$

be the decomposition given by (2). To prove that r is zero, we will use the parametrization (t, t^2, t^3) of the twisted cubic. Since f vanishes on V, we obtain

$$0 = f(t, t^2, t^3) = 0 + 0 + r(t)$$

(recall that r is a polynomial in x alone). Since t can be any real number, $r \in \mathbb{R}[x]$ must be the zero polynomial by Proposition 5 of §1. But $r = 0$ shows that f has the desired form, and $\mathbf{I}(V) = \langle y - x^2, z - x^3 \rangle$ is proved.

What we did in (2) is reminiscent of the division of polynomials, except that we are dividing by two polynomials instead of one. In fact, (2) is a special case of the generalized division algorithm to be studied in Chapter 2.

A nice corollary of the above example is that given a polynomial $f \in \mathbb{R}[x, y, z]$, we have $f \in \langle y - x^2, z - x^3 \rangle$ if and only if $f(t, t^2, t^3)$ is identically zero. This gives us an algorithm for deciding whether a polynomial lies in the ideal. However, this method is dependent on the parametrization (t, t^2, t^3). Is there a way of deciding whether $f \in \langle y - x^2, z - x^3 \rangle$ without using the parametrization? In Chapter 2, we will answer this question positively using Groebner bases and the generalized division algorithm.

The example of the twisted cubic is very suggestive. We started with the polynomials $y - x^2$ and $z - x^3$, used them to define an affine variety, took all functions vanishing on the variety, and got back the ideal generated by the two polynomials. It is natural to wonder if this happens in general. So take $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$. This gives us

$$\begin{array}{ccccc} \text{polynomials} & & \text{variety} & & \text{ideal} \\ f_1, \ldots, f_s & \to & \mathbf{V}(f_1, \ldots, f_s) & \to & \mathbf{I}(\mathbf{V}(f_1, \ldots, f_s)), \end{array}$$

and the natural question to ask is whether $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s)) = \langle f_1, \ldots, f_s \rangle$? The answer, unfortunately, is not always yes. Here is the best answer we can give at this point.

Lemma 7. *If $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$, then $\langle f_1, \ldots, f_s \rangle \subset \mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$, although equality need not occur.*

Proof. Let $f \in \langle f_1, \ldots, f_s \rangle$, which means that $f = \sum_{i=1}^{s} h_i f_i$ for some polynomials $h_1, \ldots, h_s \in k[x_1, \ldots, x_n]$. Since $f_1, \ldots, f_s$ vanish on $\mathbf{V}(f_1, \ldots, f_s)$, so must $\sum_{i=1}^{s} h_i f_i$. Thus, f vanishes on $\mathbf{V}(f_1, \ldots, f_s)$, which proves $f \in \mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$.

For the second part of the lemma, we need an example where $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$ is strictly larger than $\langle f_1, \ldots, f_s \rangle$. We will show that the inclusion

$$\langle x^2, y^2 \rangle \subset \mathbf{I}(\mathbf{V}(x^2, y^2))$$

is not an equality. We first compute $\mathbf{I}(\mathbf{V}(x^2, y^2))$. The equations $x^2 = y^2 = 0$ imply that $\mathbf{V}(x^2, y^2) = \{(0, 0)\}$. But an earlier example showed that the ideal of $\{(0, 0)\}$ is $\langle x, y \rangle$, so that $\mathbf{I}(\mathbf{V}(x^2, y^2)) = \langle x, y \rangle$. To see that this is strictly larger than $\langle x^2, y^2 \rangle$, note that $x \notin \langle x^2, y^2 \rangle$ since for polynomials of the form $h_1(x, y)x^2 + h_2(x, y)y^2$, every monomial has total degree at least two. □

For arbitrary fields, the relationship between $\langle f_1, \ldots, f_s \rangle$ and $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$ can be rather subtle (see the exercises for some examples). However, over an algebraically closed field like $\mathbb{C}$, there is a straightforward relation between these ideals. This will be explained when we prove the Nullstellensatz in Chapter 4.

Although for a general field, $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$ may not equal $\langle f_1, \ldots, f_s \rangle$, the ideal of a variety always contains enough information to determine the variety uniquely.

Proposition 8. *Let V and W be affine varieties in k^n. Then:*
(i) *$V \subset W$ if and only if $\mathbf{I}(V) \supset \mathbf{I}(W)$.*
(ii) *$V = W$ if and only if $\mathbf{I}(V) = \mathbf{I}(W)$.*

Proof. We leave it as an exercise to show that (ii) is an immediate consequence of (i). To prove (i), first suppose that $V \subset W$. Then any polynomial vanishing on W must vanish on V, which proves $\mathbf{I}(W) \subset \mathbf{I}(V)$. Next, assume that $\mathbf{I}(W) \subset \mathbf{I}(V)$. We know that W is the variety defined by some polynomials $g_1, \ldots, g_t \in k[x_1, \ldots, x_n]$. Then $g_1, \ldots, g_t \in \mathbf{I}(W) \subset \mathbf{I}(V)$, and hence the g_i's vanish on V. Since W consists of *all* common zeros of the g_i's, it follows that $V \subset W$. □

There is a rich relationship between ideals and affine varieties; the material presented so far is just the tip of the iceberg. We will explore this relation further in Chapter 4. In particular, we will see that theorems proved about ideals have strong geometric implications. For now, let us list three questions we can pose concerning ideals in $k[x_1, \ldots, x_n]$:

- (Ideal Description) Can every ideal $I \subset k[x_1, \ldots, x_n]$ be written as $\langle f_1, \ldots, f_s \rangle$ for some $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$?
- (Ideal Membership) If $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$, is there an algorithm to decide whether a given $f \in k[x_1, \ldots, x_n]$ lies in $\langle f_1, \ldots, f_s \rangle$?
- (Nullstellensatz) Given $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$, what is the exact relation between $\langle f_1, \ldots, f_s \rangle$ and $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$?

In the chapters that follow, we will solve these problems completely (and we will explain where the name Nullstellensatz comes from), although we will need to be careful about which field we are working over.

EXERCISES FOR §4

1. Consider the equations

$$x^2 + y^2 - 1 = 0,$$
$$xy - 1 = 0$$

which describe the intersection of a circle and a hyperbola.
   a. Use algebra to eliminate y from the above equations.
   b. Show how the polynomial found in part (a) lies in $\langle x^2 + y^2 - 1, xy - 1\rangle$. Your answer should be similar to what we did in (1). Hint: Multiply the second equation by $xy + 1$.
2. Let $I \subset k[x_1, \ldots, x_n]$ be an ideal, and let $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$. Prove that the following statements are equivalent:
   (i) $f_1, \ldots, f_s \in I$.
   (ii) $\langle f_1, \ldots, f_s\rangle \subset I$.
   This fact is useful when you want to show that one ideal is contained in another.
3. Use the previous exercise to prove the following equalities of ideals in $\mathbb{Q}[x, y]$:
   a. $\langle x + y, x - y\rangle = \langle x, y\rangle$.
   b. $\langle x + xy, y + xy, x^2, y^2\rangle = \langle x, y\rangle$.
   c. $\langle 2x^2 + 3y^2 - 11, x^2 - y^2 - 3\rangle = \langle x^2 - 4, y^2 - 1\rangle$.
   This illustrates that the same ideal can have many different bases and that different bases may have different numbers of elements.
4. Prove Proposition 4.
5. Show that $\mathbf{V}(x + xy, y + xy, x^2, y^2) = \mathbf{V}(x, y)$. Hint: See Exercise 3.
6. The word "basis" is used in various ways in mathematics. In this exercise, we will see that "a basis of an ideal," as defined in this section, is quite different from "a basis of a subspace," which is studied in linear algebra.
   a. First, consider the ideal $I = \langle x\rangle \subset k[x]$. As an ideal, I has a basis consisting of the one element x. But I can also be regarded as a subspace of $k[x]$, which is a vector space over k. Prove that any vector space basis of I over k is infinite. Hint: It suffices to find one basis that is infinite. Thus, allowing x to be multiplied by elements of $k[x]$ instead of just k is what enables $\langle x\rangle$ to have a finite basis.
   b. In linear algebra, a basis must span and be linearly independent over k, whereas for an ideal, a basis is concerned only with spanning—there is no mention of any sort of independence. The reason is that once we allow polynomial coefficients, no independence is possible. To see this, consider the ideal $\langle x, y\rangle \subset k[x, y]$. Show that zero can be written as a linear combination of y and x with nonzero polynomial coefficients.
   c. More generally, suppose that $f_1, \ldots, f_s$ is the basis of an ideal $I \subset k[x_1, \ldots, x_n]$. If $s \geq 2$ and $f_i \neq 0$ for all i, then show that for any i and j, zero can be written as a linear combination of f_i and f_j with nonzero polynomial coefficients.
   d. A consequence of the lack of independence is that when we write an element $f \in \langle f_1, \ldots, f_s\rangle$ as $f = \Sigma_{i=1}^{s} h_i f_i$, the coefficients h_i are not unique. As an example, consider $f = x^2 + xy + y^2 \in \langle x, y\rangle$. Express f as a linear combination of x and y in two different ways. (Even though the h_i's are not unique, one can measure their lack of uniqueness. This leads to the interesting topic of syzygies.)
   e. A basis $f_1, \ldots, f_s$ of an ideal I is said to be *minimal* if no proper subset of $f_1, \ldots, f_s$ is a basis of I. For example, x, x^2 is a basis of an ideal, but not a minimal basis since x generates the same ideal. Unfortunately, an ideal can have minimal bases consisting

of different numbers of elements. To see this, show that x and $x + x^2, x^2$ are minimal bases of the same ideal of $k[x]$. Explain how this contrasts with the situation in linear algebra.

7. Show that $\mathbf{I}(\mathbf{V}(x^n, y^m)) = \langle x, y\rangle$ for any positive integers n and m.
8. The ideal $\mathbf{I}(V)$ of a variety has a special property not shared by all ideals. Specifically, we define an ideal I to be *radical* if whenever a power f^m of a polynomial f is in I, then f itself is in I. More succinctly, I is radical when $f \in I$ if and only if $f^m \in I$ for some positive integer m.
   a. Prove that $\mathbf{I}(V)$ is always a radical ideal.
   b. Prove that $\langle x^2, y^2\rangle$ is not a radical ideal. This implies that $\langle x^2, y^2\rangle \neq \mathbf{I}(V)$ for any variety $V \subset k^2$.

   Radical ideals will play an important role in Chapter 4. In particular, the Nullstellensatz will imply that there is a one-to-one correspondence between varieties in $\mathbb{C}^n$ and radical ideals in $\mathbb{C}[x_1, \ldots, x_n]$.
9. Let $V = \mathbf{V}(y - x^2, z - x^3)$ be the twisted cubic. In the text, we showed that $\mathbf{I}(V) = \langle y - x^2, z - x^3\rangle$.
   a. Use the parametrization of the twisted cubic to show that $y^2 - xz \in \mathbf{I}(V)$.
   b. Use the argument given in the text to express $y^2 - xz$ as a combination of $y - x^2$ and $z - x^3$.
10. Use the argument given in the discussion of the twisted cubic to show that $\mathbf{I}(\mathbf{V}(x - y)) = \langle x - y\rangle$. Your argument should be valid for any infinite field k.
11. Let $V \subset \mathbb{R}^3$ be the curve parametrized by (t, t^3, t^4).
   a. Prove that V is an affine variety.
   b. Adapt the method used in the case of the twisted cubic to determine $\mathbf{I}(V)$.
12. Let $V \subset \mathbb{R}^3$ be the curve parametrized by (t^2, t^3, t^4).
   a. Prove that V is an affine variety.
   b. Determine $\mathbf{I}(V)$.

   This problem is quite a bit more challenging than the previous one—figuring out the proper analogue of equation (2) is not easy. Once we study the division algorithm in Chapter 2, this exercise will become much easier.
13. In Exercise 2 of §1, we showed that $x^2y + y^2x$ vanishes at all points of $\mathbb{F}_2^2$. More generally, let $I \subset \mathbb{F}_2[x, y]$ be the ideal of all polynomials that vanish at all points of $\mathbb{F}_2^2$. The goal of this exercise is to show that $I = \langle x^2 - x, y^2 - y\rangle$.
   a. Show that $\langle x^2 - x, y^2 - y\rangle \subset I$.
   b. Show that every $f \in \mathbb{F}_2[x, y]$ can be written as $f = A(x^2 - x) + B(y^2 - y) + axy + bx + cy + d$, where $A, B \in \mathbb{F}_2[x, y]$ and $a, b, c, d \in \mathbb{F}_2$. Hint: Write f in the form $\Sigma_i\, p_i(x)y^i$ and use the division algorithm (Proposition 2 of §5) to divide each p_i by $x^2 - x$. From this, you can write $f = A(x^2 - x) + q_1(y)x + q_2(y)$. Now divide q_1 and q_2 by $y^2 - y$. Again, this argument will become vastly simpler once we know the division algorithm from Chapter 2.
   c. Show that $axy + bx + cy + d \in I$ if and only if $a = b = c = d = 0$.
   d. Using parts (b) and (c), complete the proof that $I = \langle x^2 - x, y^2 - y\rangle$.
   e. Express $x^2y + y^2x$ as a combination of $x^2 - x$ and $y^2 - y$. Hint: Remember that $2 = 1 + 1 = 0$ in $\mathbb{F}_2$.
14. This exercise is concerned with Proposition 8.
   a. Prove that part (ii) of the proposition follows from part (i).
   b. Prove the following corollary of the proposition: if V and W are affine varieties in k^n, then $V \subsetneq W$ if and only if $\mathbf{I}(V) \supsetneq \mathbf{I}(W)$.

15. In the text, we defined $\mathbf{I}(V)$ for a variety $V \subset k^n$. We can generalize this as follows: if $S \subset k^n$ is *any* subset, then we set

$$\mathbf{I}(S) = \{f \in k[x_1, \ldots, x_n] : f(a_1, \ldots, a_n) = 0 \text{ for all } (a_1, \ldots, a_n) \in S\}.$$

   a. Prove that $\mathbf{I}(S)$ is an ideal.
   b. Let $X = \{(a, a) \in \mathbb{R}^2 : a \neq 1\}$. By Exercise 8 of §2, we know that X is not an affine variety. Determine $\mathbf{I}(X)$. Hint: What you proved in Exercise 8 of §2 will be useful. See also Exercise 10 of this section.
   c. Let $\mathbb{Z}^n$ be the points of $\mathbb{C}^n$ with integer coordinates. Determine $\mathbf{I}(\mathbb{Z}^n)$. Hint: See Exercise 6 of §1.

§5 Polynomials of One Variable

In this section, we will discuss polynomials of one variable and study the *division algorithm* from high school algebra. This simple algorithm has some surprisingly deep consequences—for example, we will use it to determine the structure of ideals of $k[x]$ and to explore the idea of a *greatest common divisor*. The theory developed will allow us to solve, in the special case of polynomials in $k[x]$, most of the problems raised in earlier sections. We will also begin to understand the important role played by algorithms.

By this point in their mathematics careers, most students have already seen a variety of algorithms, although the term "algorithm" may not have been used. Informally, an algorithm is a specific set of instructions for manipulating symbolic or numerical data. Examples are the differentiation formulas from calculus and the method of row reduction from linear algebra. An algorithm will have *inputs*, which are objects used by the algorithm, and *outputs*, which are the results of the algorithm. At each stage of execution, the algorithm must specify exactly what the next step will be.

When we are studying an algorithm, we will usually present it in "pseudocode," which will make the formal structure easier to understand. Pseudocode is similar to the computer language Pascal, and a brief discussion is given in Appendix B. Another reason for using pseudocode is that it indicates how the algorithm could be programmed on a computer. We should also mention that most of the algorithms in this book are implemented in computer algebra systems such as AXIOM, Macsyma, Maple, Mathematica, and REDUCE. Appendix C has more details concerning these programs.

We begin by discussing the division algorithm for polynomials in $k[x]$. A crucial component of this algorithm is the notion of the "leading term" of a polynomial in one variable. The precise definition is as follows.

Definition 1. *Given a nonzero polynomial $f \in k[x]$, let*

$$f = a_0 x^m + a_1 x^{m-1} + \cdots + a_m,$$

where $a_i \in k$ and $a_0 \neq 0$ [thus, $m = \deg(f)$]. Then we say that $a_0 x^m$ is the **leading term** *of f, written* $\mathrm{LT}(f) = a_0 x^m$.

For example, if $f = 2x^3 - 4x + 3$, then $\text{LT}(f) = 2x^3$. Notice also that if f and g are nonzero polynomials, then

$$\deg(f) \leq \deg(g) \iff \text{LT}(f) \text{ divides } \text{LT}(g). \tag{1}$$

We can now describe the division algorithm.

Proposition 2 (The Division Algorithm). *Let k be a field and let g be a nonzero polynomial in $k[x]$. Then every $f \in k[x]$ can be written as*

$$f = qg + r,$$

where $q, r \in k[x]$, and either $r = 0$ or $\deg(r) < \deg(g)$. Furthermore, q and r are unique, and there is an algorithm for finding q and r.

Proof. Here is the algorithm for finding q and r, presented in pseudocode:

Input: g, f
Output: q, r
$q := 0$; $r := f$
WHILE $r \neq 0$ AND $\text{LT}(g)$ divides $\text{LT}(r)$ DO
    $q := q + \text{LT}(r)/\text{LT}(g)$
    $r := r - (\text{LT}(r)/\text{LT}(g))g$

The WHILE ... DO statement means doing the indented operations until the expression between the WHILE and DO becomes false. The statements $q := \ldots$ and $r := \ldots$ indicate that we are defining or redefining the values of q and r. Both q and r are *variables* in this algorithm—they change value at each step. We need to show that the algorithm terminates and that the final values of q and r have the required properties. (For a fuller discussion of pseudocode, see Appendix B.)

To see why this algorithm works, first note that $f = qg + r$ holds for the initial values of q and r, and that whenever we redefine q and r, the equality $f = qg + r$ remains true. This is because of the identity

$$f = qg + r = (q + \text{LT}(r)/\text{LT}(g))g + (r - (\text{LT}(r)/\text{LT}(g))g).$$

Next, note that the WHILE ... DO statement terminates when "$r \neq 0$ and $\text{LT}(g)$ divides $\text{LT}(r)$" is false, i.e., when either $r = 0$ or $\text{LT}(g)$ does not divide $\text{LT}(r)$. By (1), this last statement is equivalent to $\deg(r) < \deg(g)$. Thus, when the algorithm terminates, it produces q and r with the required properties.

We are not quite done; we still need to show that the algorithm terminates, i.e., that the expression between the WHILE and DO eventually becomes false (otherwise, we would be stuck in an infinite loop). The key observation is that $r - (\text{LT}(r)/\text{LT}(g))g$ is either 0 or has smaller degree than r. To see why, suppose that

$$\begin{aligned} r &= a_0x^m + \cdots + a_m, \quad \text{LT}(r) = a_0x^m, \\ g &= b_0x^k + \cdots + b_k, \quad \text{LT}(g) = b_0x^k, \end{aligned}$$

and suppose that $m \geq k$. Then

$$r - (\mathrm{LT}(r)/\mathrm{LT}(g))g = (a_0x^m + \cdots) - (a_0/b_0)x^{m-k}(b_0x^k + \cdots),$$

and it follows that the degree of r must drop (or the whole expression may vanish). Since the degree is finite, it can drop at most finitely many times, which proves that the algorithm terminates.

To see how this algorithm corresponds to the process learned in high school, consider the following partially completed division:

$$\begin{array}{r} \frac{1}{2}x^2 \\ 2x+1 \;\overline{)\; x^3+2x^2+x+1} \\ \underline{x^3+\frac{1}{2}x^2} \\ \frac{3}{2}x^2+x+1. \end{array}$$

Here, f and g are given by $f = x^3 + 2x^2 + x + 1$ and $g = 2x + 1$, and more importantly, the current (but *not* final) values of q and r are $q = \frac{1}{2}x^2$ and $r = \frac{3}{2}x^2 + x + 1$. Now notice that the statements

$$\begin{aligned} q &:= q + \mathrm{LT}(r)/\mathrm{LT}(g), \\ r &:= r - (\mathrm{LT}(r)/\mathrm{LT}(g))g \end{aligned}$$

in the WHILE ... DO loop correspond exactly to the next step in the above division.

The final step in proving the proposition is to show that q and r are unique. So suppose that $f = qg + r = q'g + r'$ where both r and r' have degree less than g (unless one or both are 0). If $r \neq r'$, then $\deg(r' - r) < \deg(g)$. On the other hand, since

$$(q - q')g = r' - r, \tag{2}$$

we would have $q - q' \neq 0$, and consequently,

$$\deg(r' - r) = \deg((q - q')g) = \deg(q - q') + \deg(g) \geq \deg(g).$$

This contradiction forces $r = r'$, and then (2) shows that $q = q'$. This completes the proof of the proposition. □

Most computer algebra systems implement the above algorithm [with some modifications—see DAVENPORT, SIRET, and TOURNIER (1993)] for dividing polynomials.

A useful corollary of the division algorithm concerns the number of roots of a polynomial in one variable.

Corollary 3. *If k is a field and $f \in k[x]$ is a nonzero polynomial, then f has at most* $\deg(f)$ *roots in k.*

Proof. We will use induction on $m = \deg(f)$. When $m = 0$, f is a nonzero constant, and the corollary is obviously true. Now assume that the corollary holds for all polynomials of degree $m - 1$, and let f have degree m. If f has no roots in k, then we are done. So suppose a is a root in k. If we divide f by $x - a$, then Proposition 2 tells

us that $f = q(x-a)+r$, where $r \in k$ since $x-a$ has degree one. To determine r, evaluate both sides at $x=a$, which gives $0 = f(a) = q(a)(a-a)+r = r$. It follows that $f = q(x-a)$. Note also that q has degree $m-1$.

We claim that any root of f other than a is also a root of q. To see this, let $b \neq a$ be a root of f. Then $0 = f(b) = q(b)(b-a)$ implies that $q(b) = 0$ since k is a field. Since q has at most $m-1$ roots by our inductive assumption, f has at most m roots in k. This completes the proof. □

Corollary 3 was used to prove Proposition 5 in §1, which states that $\mathbf{I}(k^n) = \{0\}$ whenever k is infinite. This is an example of how a geometric fact can be the consequence of an algorithm.

We can also use Proposition 2 to determine the structure of all ideals of $k[x]$.

Corollary 4. *If k is a field, then every ideal of $k[x]$ can be written in the form $\langle f \rangle$ for some $f \in k[x]$. Furthermore, f is unique up to multiplication by a nonzero constant in k.*

Proof. Take an ideal $I \subset k[x]$. If $I = \{0\}$, then we are done since $I = \langle 0 \rangle$. Otherwise, let f be a nonzero polynomial of minimum degree contained in I. We claim that $\langle f \rangle = I$. The inclusion $\langle f \rangle \subset I$ is obvious since I is an ideal. Going the other way, take $g \in I$. By division algorithm (Proposition 2), we have $g = qf + r$, where either $r = 0$ or $\deg(r) < \deg(f)$. Since I is an ideal, $qf \in I$ and, thus, $r = g - qf \in I$. If r were not zero, then $\deg(r) < \deg(f)$, which would contradict our choice of f. Thus, $r = 0$, so that $g = qf \in \langle f \rangle$. This proves that $I = \langle f \rangle$.

To study uniqueness, suppose that $\langle f \rangle = \langle g \rangle$. Then $f \in \langle g \rangle$ implies that $f = hg$ for some polynomial h. Thus,

$$\deg(f) = \deg(h) + \deg(g), \tag{3}$$

so that $\deg(f) \geq \deg(g)$. The same argument with f and g interchanged shows $\deg(f) \leq \deg(g)$, and it follows that $\deg(f) = \deg(g)$. Then (3) implies $\deg(h) = 0$, so that h is a nonzero constant. □

In general, an ideal generated by one element is called a *principal ideal*. In view of Corollary 4, we say that $k[x]$ is a *principal ideal domain*, abbreviated PID.

The proof of Corollary 4 tells us that the generator of an ideal in $k[x]$ is the nonzero polynomial of minimum degree contained in the ideal. This description is not useful in practice, for it requires that we check the degrees of all polynomials (there are infinitely many) in the ideal. Is there a better way to find the generator? For example, how do we find a generator of the ideal

$$\langle x^4 - 1, x^6 - 1 \rangle \subset k[x]?$$

The tool needed to solve this problem is the greatest common divisor.

Definition 5. A greatest common divisor *of polynomials $f, g \in k[x]$ is a polynomial h such that:*

(i) *h divides f and g.*
(ii) *If p is another polynomial which divides f and g, then p divides h. When h has these properties, we write $h = \mathrm{GCD}(f, g)$.*

Here are the main properties of GCDs.

Proposition 6. *Let $f, g \in k[x]$. Then:*
(i) GCD*(f, g) exists and is unique up to multiplication by a nonzero constant in k.*
(ii) GCD*(f, g) is a generator of the ideal $\langle f, g\rangle$.*
(iii) *There is an algorithm for finding* GCD*(f, g).*

Proof. Consider the ideal $\langle f, g\rangle$. Since every ideal of $k[x]$ is principal (Corollary 4), there exists $h \in k[x]$ such that $\langle f, g\rangle = \langle h\rangle$. We claim that h is the GCD of f, g. To see this, first note that h divides f and g since $f, g \in \langle h\rangle$. Thus, the first part of Definition 5 is satisfied. Next, suppose that $p \in k[x]$ divides f and g. This means that $f = Cp$ and $g = Dp$ for some $C, D \in k[x]$. Since $h \in \langle f, g\rangle$, there are A, B such that $Af + Bg = h$. Substituting, we obtain

$$h = Af + Bg = ACp + BDp = (AC + BD)p,$$

which shows that p divides h. Thus, $h = \mathrm{GCD}(f, g)$.

This proves the existence of the GCD. To prove uniqueness, suppose that h' was another GCD of f and g. Then, by the second part of Definition 5, h and h' would each divide the other. This easily implies that h is a nonzero constant multiple of h'. Thus, part (i) of the corollary is proved, and part (ii) follows by the way we found h in the above paragraph.

The existence proof just given is not useful in practice. It depends on our ability to find a generator of $\langle f, g\rangle$. As we noted in the discussion following Corollary 4, this involves checking the degrees of infinitely many polynomials. Fortunately, there is a classic algorithm, known as the *Euclidean Algorithm*, which computes the GCD of two polynomials in $k[x]$. This is what part (iii) of the proposition is all about.

We will need the following notation. Let $f, g \in k[x]$, where $g \neq 0$, and write $f = qg + r$, where q and r are as in Proposition 2. Then we set $r = \mathrm{remainder}(f, g)$. We can now state the Euclidean Algorithm for finding GCD(f, g):

Input: f, g
Output: h
$h := f$
$s := g$
WHILE $s \neq 0$ DO
    $rem := \mathrm{remainder}(h, s)$
    $h := s$
    $s := rem$

To see why this algorithm computes the GCD, write $f = qg + r$ as in Proposition 2. We claim that

$$\text{GCD}(f, g) = \text{GCD}(f - qg, g) = \text{GCD}(r, g). \tag{4}$$

To prove this, by part (ii) of the proposition, it suffices to show that the ideals $\langle f, g\rangle$ and $\langle f - qg, g\rangle$ are equal. We will leave this easy argument as an exercise.

We can write (4) in the form

$$\text{GCD}(f, g) = \text{GCD}(g, r).$$

Notice that $\deg(g) > \deg(r)$ or $r = 0$. If $r \neq 0$, we can make things yet smaller by repeating this process. Thus, we write $g = q'r + r'$ as in Proposition 2, and arguing as above, we obtain

$$\text{GCD}(g, r) = \text{GCD}(r, r'),$$

where $\deg(r) > \deg(r')$ or $r = 0$. Continuing in this way, we get

$$\text{GCD}(f, g) = \text{GCD}(g, r) = \text{GCD}(r, r') = \text{GCD}(r', r'') = \cdots, \tag{5}$$

where either the degrees drop

$$\deg(g) > \deg(r) > \deg(r') > \deg(r'') > \cdots,$$

or the process terminates when one of $r, r', r'', \ldots$ becomes 0.

We can now explain how the Euclidean Algorithm works. The algorithm has variables h and s, and we can see these variables in equation (5): the values of h are the first polynomial in each GCD, and the values of s are the second. You should check that in (5), going from one GCD to the next is exactly what is done in the WHILE ... DO loop of the algorithm. Thus, at every stage of the algorithm, $\text{GCD}(h, s) = \text{GCD}(f, g)$.

The algorithm must terminate because the degree of s keeps dropping, so that at some stage, $s = 0$. When this happens, we have $\text{GCD}(h, 0) = \text{GCD}(f, g)$, and since $\langle h, 0\rangle$ obviously equals $\langle h\rangle$, we have $\text{GCD}(h, 0) = h$. Combining these last two equations, it follows that $h = \text{GCD}(f, g)$ when $s = 0$. This proves that h is the GCD of f and g when the algorithm terminates, and the proof of Proposition 6 is now complete. □

We should mention that there is also a version of the Euclidean Algorithm for finding the GCD of two integers. Most computer algebra systems have a command for finding the GCD of two polynomials (or integers) that uses a modified form of the Euclidean Algorithm [see DAVENPORT, SIRET, and TOURNIER (1993) for more details].

For an example of how the Euclidean Algorithm works, let us compute the GCD of $x^4 - 1$ and $x^6 - 1$. First, we use the division algorithm:

$$\begin{aligned} x^4 - 1 &= 0(x^6 - 1) + x^4 - 1, \\ x^6 - 1 &= x^2(x^4 - 1) + x^2 - 1, \\ x^4 - 1 &= (x^2 + 1)(x^2 - 1) + 0. \end{aligned}$$

Then, by equation (5), we have

$$\begin{aligned}\mathrm{GCD}(x^4-1, x^6-1) &= \mathrm{GCD}(x^6-1, x^4-1)\\ &= \mathrm{GCD}(x^4-1, x^2-1) = \mathrm{GCD}(x^2-1, 0) = x^2-1.\end{aligned}$$

Note that this GCD computation answers our earlier question of finding a generator for the ideal $\langle x^4-1, x^6-1\rangle$. Namely, Proposition 6 and $\mathrm{GCD}(x^4-1, x^6-1) = x^2-1$ imply that

$$\langle x^4-1, x^6-1\rangle = \langle x^2-1\rangle.$$

At this point, it is natural to ask what happens for an ideal generated by three or more polynomials. How do we find a generator in this case? The idea is to extend the definition of GCD to more than two polynomials.

Definition 7. A greatest common divisor *of polynomials $f_1, \dots, f_s \in k[x]$ is a polynomial h such that:*
(i) *h divides $f_1, \dots, f_s$.*
(ii) *If p is another polynomial which divides $f_1, \dots, f_s$, then p divides h.*
When h has these properties, we write $h = \mathrm{GCD}(f_1, \dots, f_s)$.

Here are the main properties of these GCDs.

Proposition 8. *Let $f_1, \dots, f_s \in k[x]$, where $s \geq 2$. Then:*
(i) *$\mathrm{GCD}(f_1, \dots, f_s)$ exists and is unique up to multiplication by a nonzero constant in k.*
(ii) *$\mathrm{GCD}(f_1, \dots, f_s)$ is a generator of the ideal $\langle f_1, \dots, f_s\rangle$.*
(iii) *If $s \geq 3$, then $\mathrm{GCD}(f_1, \dots, f_s) = \mathrm{GCD}(f_1, \mathrm{GCD}(f_2, \dots, f_s))$.*
(iv) *There is an algorithm for finding $\mathrm{GCD}(f_1, \dots, f_s)$.*

Proof. The proofs of parts (i) and (ii) are similar to the proofs given in Proposition 6 and will be omitted. To prove part (iii), let $h = \mathrm{GCD}(f_2, \dots, f_s)$. We leave it as an exercise to show that

$$\langle f_1, h\rangle = \langle f_1, f_2, \dots, f_s\rangle.$$

By part (ii) of this proposition, we see that

$$\langle \mathrm{GCD}(f_1, h)\rangle = \langle \mathrm{GCD}(f_1, \dots, f_s)\rangle.$$

Then $\mathrm{GCD}(f_1, h) = \mathrm{GCD}(f_1, \dots, f_s)$ follows from the uniqueness part of Corollary 4, which proves what we want.

Finally, we need to show that there is an algorithm for finding $\mathrm{GCD}(f_1, \dots, f_s)$. The basic idea is to combine part (iii) with the Euclidean Algorithm. For example, suppose that we wanted to compute the GCD of four polynomials f_1, f_2, f_3, f_4. Using part (iii) of the proposition twice, we obtain

$$\begin{aligned}\mathrm{GCD}(f_1, f_2, f_3, f_4) &= \mathrm{GCD}(f_1, \mathrm{GCD}(f_2, f_3, f_4))\\ &= \mathrm{GCD}(f_1, \mathrm{GCD}(f_2, \mathrm{GCD}(f_3, f_4))).\end{aligned} \tag{6}$$

Then if we use the Euclidean Algorithm three times [once for each GCD in the second line of (6)], we get the GCD of f_1, f_2, f_3, f_4. In the exercises, you will be asked to write pseudocode for an algorithm that implements this idea for an arbitrary number of polynomials. Proposition 8 is proved. □

The GCD command in most computer algebra systems only handles two polynomials at a time. Thus, to work with more than two polynomials, you will need to use the method described in the proof of Proposition 8. For an example, consider the ideal

$$\langle x^3 - 3x + 2, x^4 - 1, x^6 - 1\rangle \subset k[x].$$

We know that $\mathrm{GCD}(x^3 - 3x + 2, x^4 - 1, x^6 - 1)$ is a generator. Furthermore, you can check that

$$\begin{aligned}\mathrm{GCD}(x^3 - 3x + 2, x^4 - 1, x^6 - 1) &= \mathrm{GCD}(x^3 - 3x + 2, \mathrm{GCD}(x^4 - 1, x^6 - 1))\\ &= \mathrm{GCD}(x^3 - 3x + 2, x^2 - 1) = x - 1.\end{aligned}$$

It follows that

$$\langle x^3 - 3x + 2, x^4 - 1, x^6 - 1\rangle = \langle x - 1\rangle.$$

More generally, given $f_1, \ldots, f_s \in k[x]$, it is clear that we now have an algorithm for finding a generator of $\langle f_1, \ldots, f_s\rangle$.

For another application of the algorithms developed here, consider the *ideal membership problem* from §4: given $f_1, \ldots, f_s \in k[x]$, is there an algorithm for deciding whether a given polynomial $f \in k[x]$ lies in the ideal $\langle f_1, \ldots, f_s\rangle$? The answer is yes, and the algorithm is easy to describe. The first step is to use GCDs to find a generator h of $\langle f_1, \ldots, f_s\rangle$. Then, since $f \in \langle f_1, \ldots, f_s\rangle$ is equivalent to $f \in \langle h\rangle$, we need only use the division algorithm to write $f = qh + r$, where $\deg(r) < \deg(h)$. It follows that f is in the ideal if and only if $r = 0$. For example, suppose we wanted to know whether

$$x^3 + 4x^2 + 3x - 7 \in \langle x^3 - 3x + 2, x^4 - 1, x^6 - 1\rangle.$$

We saw above that $x-1$ is a generator of this ideal so that our question can be rephrased as to whether

$$x^3 + 4x^2 + 3x - 7 \in \langle x - 1\rangle.$$

Dividing, we find that

$$x^3 + 4x^2 + 3x - 7 = (x^2 + 5x + 8)(x - 1) + 1$$

and it follows that $x^3 + 4x^2 + 3x - 7$ is *not* in the ideal $\langle x^3 - 3x + 2, x^4 - 1, x^6 - 1\rangle$. In Chapter 2, we will solve the ideal membership problem for polynomials in $k[x_1, \ldots, x_n]$ using a similar strategy: we will first find a nice basis of the ideal (called a Groebner basis) and then we will use a generalized division algorithm to determine whether or not a polynomial is in the ideal.

In the exercises, we will see that in the one-variable case, other problems posed in earlier sections can be solved algorithmically using the methods discussed here.

EXERCISES FOR §5

1. Over the complex numbers $\mathbb{C}$, Corollary 3 can be stated in a stronger form. Namely, prove that if $f \in \mathbb{C}[x]$ is a polynomial of degree $n > 0$, then f can be written in the form $f = c(x - a_1) \cdots (x - a_n)$, where $c, a_1, \ldots, a_n \in \mathbb{C}$ and $c \neq 0$. Hint: Use Theorem 7 of §1. Note that this result holds for *any* algebraically closed field.
2. Although Corollary 3 is simple to prove, it has some interesting consequences. For example, consider the $n \times n$ Vandermonde determinant determined by $a_1, \ldots, a_n$ in a field k:
$$\det \begin{pmatrix} 1 & a_1 & a_1^2 & \ldots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \ldots & a_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & a_n^2 & \ldots & a_n^{n-1} \end{pmatrix}.$$
Prove that this determinant is nonzero when the a_i's are distinct. Hint: If the determinant is zero, then the columns are linearly dependent. Show that the coefficients of the linear relation determine a polynomial of degree $\leq n-1$ which has n roots. Then use Corollary 3.
3. The fact that every ideal of $k[x]$ is principal (generated by one element) is special to the case of polynomials in one variable. In this exercise we will see why. Namely, consider the ideal $I = \langle x, y \rangle \subset k[x, y]$. Prove that I is not a principal ideal. Hint: If $x = fg$, where $f, g \in k[x, y]$, then prove that f or g is a constant. It follows that the treatment of GCDs given in this section applies only to polynomials in one variable. GCDs can be computed for polynomials of ≥ 2 variables, but the theory involved is more complicated [see DAVENPORT, SIRET, and TOURNIER (1993), §4.1.2].
4. If h is the GCD of $f, g \in k[x]$, then prove that there are polynomials $A, B \in k[x]$ such that $Af + Bg = h$.
5. If $f, g \in k[x]$, then prove that $\langle f - qg, g \rangle = \langle f, g \rangle$ for any q in $k[x]$. This will prove equation (4) in the text.
6. Given $f_1, \ldots, f_s \in k[x]$, let $h = \mathrm{GCD}(f_2, \ldots, f_s)$. Then use the equality $\langle h \rangle = \langle f_2, \ldots, f_s \rangle$ to show that $\langle f_1, h \rangle = \langle f_1, f_2, \ldots, f_s \rangle$. This equality is used in the proof of part (iii) of Proposition 8.
7. If you are allowed to compute the GCD of only two polynomials at a time (which is true for most computer algebra systems), give pseudocode for an algorithm that computes the GCD of polynomials $f_1, \ldots, f_s \in k[x]$, where $s > 2$. Prove that your algorithm works. Hint: See (6). This will complete the proof of part (iv) of Proposition 8.
8. Use a computer algebra system to compute the following GCDs:
   a. $\mathrm{GCD}(x^4 + x^2 + 1, x^4 - x^2 - 2x - 1, x^3 - 1)$.
   b. $\mathrm{GCD}(x^3 + 2x^2 - x - 2, x^3 - 2x^2 - x + 2, x^3 - x^2 - 4x + 4)$.
9. Use the method described in the text to decide whether $x^2 - 4 \in \langle x^3 + x^2 - 4x - 4, x^3 - x^2 - 4x + 4, x^3 - 2x^2 - x + 2 \rangle$.
10. Give pseudocode for an algorithm that has input $f, g \in k[x]$ and output $h, A, B \in k[x]$ where $h = \mathrm{GCD}(f, g)$ and $Af + Bg = h$. Hint: The idea is to add variables A, B, C, D to the algorithm so that $Af + Bg = h$ and $Cf + Dg = s$ remain true at every step of the algorithm. Note that the initial values of A, B, C, D are $1, 0, 0, 1$, respectively. You may find it useful to let $\mathrm{quotient}(f, g)$ denote the quotient of f on division by g, i.e., if the division algorithm yields $f = qg + r$, then $q = \mathrm{quotient}(f, g)$.
11. In this exercise we will study the one-variable case of the *consistency problem* from §2. Given $f_1, \ldots, f_s \in k[x]$, this asks if there is an algorithm to decide whether $\mathbf{V}(f_1, \ldots, f_s)$ is nonempty. We will see that the answer is yes when $k = \mathbb{C}$.

a. Let $f \in \mathbb{C}[x]$ be a nonzero polynomial. Then use Theorem 7 of §1 to show that $\mathbf{V}(f) = \emptyset$ if and only if f is constant.
b. If $f_1, \ldots, f_s \in \mathbb{C}[x]$, prove $\mathbf{V}(f_1, \ldots, f_s) = \emptyset$ if and only if $\mathrm{GCD}(f_1, \ldots, f_s) = 1$.
c. Describe (in words, not pseudocode) an algorithm for determining whether or not $\mathbf{V}(f_1, \ldots, f_s)$ is nonempty.

When $k = \mathbb{R}$, the consistency problem is much more difficult. It requires giving an algorithm that tells whether a polynomial $f \in \mathbb{R}[x]$ has a real root.

12. This exercise will study the one-variable case of the *Nullstellensatz problem* from §4, which asks for the relation between $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$ and $\langle f_1, \ldots, f_s \rangle$ when $f_1, \ldots, f_s \in \mathbb{C}[x]$. By using GCDs, we can reduce to the case of a single generator. So, in this problem, we will explicitly determine $\mathbf{I}(\mathbf{V}(f))$ when $f \in \mathbb{C}[x]$ is a nonconstant polynomial. Since we are working over the complex numbers, we know by Exercise 1 that f factors completely, i.e.,

$$f = c(x - a_1)^{r_1} \cdots (x - a_l)^{r_l},$$

where $a_1, \ldots, a_l \in \mathbb{C}$ are distinct and $c \in \mathbb{C} - \{0\}$. Define the polynomial

$$f_{\mathrm{red}} = c(x - a_1) \cdots (x - a_l).$$

Note that f and f_{red} have the same roots, but their *multiplicities* may differ. In particular, all roots of f_{red} have multiplicity one. It is common to call f_{red} the *reduced* or *square-free* part of f. To explain the latter name, notice that f_{red} is the square-free factor of f of largest degree.

a. Show that $\mathbf{V}(f) = \{a_1, \ldots, a_l\}$.
b. Show that $\mathbf{I}(\mathbf{V}(f)) = \langle f_{\mathrm{red}} \rangle$.

Whereas part (b) describes $\mathbf{I}(\mathbf{V}(f))$, the answer is not completely satisfactory because we need to factor f completely to find f_{red}. In Exercises 13, 14, and 15 we will show how to determine f_{red} without *any* factoring.

13. We will study the formal derivative of $f = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n \in \mathbb{C}[x]$. The *formal derivative* is defined by the usual formulas from calculus:

$$f' = n a_0 x^{n-1} + (n-1) a_1 x^{n-2} + \cdots + a_{n-1} + 0.$$

Prove that the following rules of differentiation apply:

$$\begin{aligned} (af)' &= af' \quad \text{when } a \in \mathbb{C}, \\ (f+g)' &= f' + g', \\ (fg)' &= f'g + fg'. \end{aligned}$$

14. In this exercise we will use the differentiation properties of Exercise 13 to compute $\mathrm{GCD}(f, f')$ when $f \in \mathbb{C}[x]$.
a. Suppose $f = (x-a)^r h$ in $\mathbb{C}[x]$, where $h(a) \neq 0$. Then prove that $f' = (x-a)^{r-1} h_1$, where $h_1 \in \mathbb{C}[x]$ does not vanish at a. Hint: Use the product rule.
b. Let $f = c(x-a_1)^{r_1} \cdots (x-a_l)^{r_l}$ be the factorization of f, where $a_1, \ldots, a_l$ are distinct. Prove that f' is a product $f' = (x-a_1)^{r_1-1} \cdots (x-a_l)^{r_l-1} H$, where $H \in \mathbb{C}[x]$ is a polynomial vanishing at none of $a_1, \ldots, a_l$.
c. Prove that $\mathrm{GCD}(f, f') = (x-a_1)^{r_1-1} \cdots (x-a_l)^{r_l-1}$.

15. This exercise is concerned with the square-free part f_{red} of a polynomial $f \in \mathbb{C}[x]$, which is defined in Exercise 12.
a. Use Exercise 14 to prove that f_{red} is given by the formula

$$f_{\mathrm{red}} = \frac{f}{\mathrm{GCD}(f, f')}.$$

The virtue of this formula is that it allows us to find the square-free part without factoring f. This allows for much quicker computations.

b. Use a computer algebra system to find the square-free part of the polynomial

$$x^{11} - x^{10} + 2x^8 - 4x^7 + 3x^5 - 3x^4 + x^3 + 3x^2 - x - 1.$$

16. Use Exercises 12 and 15 to describe (in words, not pseudocode) an algorithm whose input consists of polynomials $f_1, \ldots, f_s \in \mathbf{C}[x]$ and whose output consists of a basis of $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$. It is *much* more difficult to construct such an algorithm when dealing with polynomials of more than one variable.
17. Find a basis for the ideal $\mathbf{I}(\mathbf{V}(x^5 - 2x^4 + 2x^2 - x, x^5 - x^4 - 2x^3 + 2x^2 + x - 1))$.