Chapter 11:
Roots of multivariate polyomials

This chapter is about the roots of polynomial equations. However, rather than investigate the *computation* of roots, it considers the *analysis* of roots, and the tools used to compute that analysis. In particular, we want to know when the roots to a multivariate system of polynomial equations exists.

A chemist once emailed me about a problem he was studying that involved microarrays. Microarrays measure gene expression, and he was using some data to build a system of equations of this form:

$$\begin{aligned} axy - b_1x - cy + d_1 &= 0 \\ axy - b_2x - cy + d_2 &= 0 \\ axy - b_2x - b_1y + d_3 &= 0 \end{aligned} \tag{34}$$

where $a, b_1, b_2, c, d_1, d_2, d_3 \in \mathbb{N}$ are known constants and $x, y \in \mathbb{R}$ were unknown. A— wanted to find values for x and y that made all the equations true.

This already is an interesting problem, and it is well-studied. In fact, A— had a fancy software program that sometimes solved the system. However, it didn't *always* solve the system, and he didn't understand whether it was because there was something wrong with his numbers, or with the system itself. All he knew is that for some values of the coefficients, the system gave him a solution, but for other values the system turned red, which meant that it found no solution.

The software the chemist was using relied on well-known *numerical techniques* to look for a solution. There are many reasons that numerical techniques can fail; most importantly, they can fail *even when a solution exists.*

Analyzing these systems with an *algebraic* technique, I was able to give him some glum news: the reason the software failed to find a solution is that, in fact, no *real* solution existed. Instead, the solutions were *complex*. So, the problem wasn't with the software's numerical techniques.

This chapter develops and describes the algebraic techniques that allowed me to reach this conclusion. Most of the material in these notes are relatively "old": at least a century old. Gröbner bases, however, are relatively new: they were first described in 1965 [Buc65]. We will develop Gröbner bases, and finally explain how they answer the following important questions for any system of polynomial equations

$$f_1(x_1, x_2, \ldots, x_n) = 0, \quad \cdots \quad f_m(x_1, x_2, \ldots, x_n) = 0$$

whose coefficients are in $\mathbb{R}$:

1. Does the system have any solutions in $\mathbb{C}$?
2. If so,
   (a) Are there infinitely many, or finitely many?
      i. If finitely many, exactly how many?
      ii. If infinitely many, what is the "dimension" of the solution set?
   (b) Are any of the solutions in $\mathbb{R}$?

We will refer to these as ***five natural questions about the roots of a polynomial system***. To answer them, we first review a little linear algebra, then study monomials a bit more, before concluding

with a foray into Hilbert's Nullstellensatz and Gröbner bases, fundamental results and tools of commutative algebra and algebraic geometry.

Remark 11.1. From here on, all rings are polynomial rings over a field $\mathbb{F}$, *unless we say otherwise.*

11.1: Gaussian elimination

Let's look again at the system (34) described in the introduction:

$$\begin{aligned} axy - b_1x - cy + d_1 &= 0 \\ axy - b_2x - cy + d_2 &= 0 \\ axy - b_2x - b_1y + d_3 &= 0. \end{aligned}$$

It is *almost* a linear system, and you've studied linear systems in the past. In fact, you've even studied how to answer the five natural questions about the roots of a linear polynomial system. Let's review that.

A generic system of m linear equations in n variables looks like

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\qquad & \qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

where the a_{ij} and b_i are elements of a field $\mathbb{F}$. Linear algebra can be done over *any* field $\mathbb{F}$, although it is typically taught with $\mathbb{F} = \mathbb{Q}$. Since that's old hat to you, let's try some linear algebra over a finite field!

Example 11.2. A linear system with $m = 3$ and $n = 5$ and coefficients in $\mathbb{Z}_{13}$ is

$$\begin{aligned} 5x_1 + x_2 + 7x_5 &= 7 \\ x_3 + 11x_4 + 2x_5 &= 1 \\ 3x_1 + 7x_2 + 8x_3 &= 2. \end{aligned}$$

An equivalent system, with the same solutions, is

$$\begin{aligned} 5x_1 + x_2 + 7x_5 + 6 &= 0 \\ x_3 + 11x_4 + 2x_5 + 12 &= 0 \\ 3x_1 + 7x_2 + 8x_3 + 11 &= 0. \end{aligned}$$

In these notes, we favor the latter form.

A technique called *Gaussian elimination* obtains a "triangular system" equivalent to the original. By "equivalent", we mean that $(a_1, \ldots, a_n) \in \mathbb{F}^n$ is a solution to the triangular system if and only if it is a solution to the original system as well.

Definition 11.3. Let $G = (g_1, g_2, \ldots, g_m)$ be a list of linear polynomials in n variables. For each $i = 1, 2, \ldots, m$ designate the **leading variable of** g_i, as the smallest-indexed variable of non-zero coefficient. Write $\mathrm{lv}(g_i)$ for this variable.

The leading variable of the zero polynomial is undefined.

Since this ordering guarantees $x_1 > x_2 > \ldots > x_n$, something like a dictionary, we refer to it as the **lexicographic term ordering**.

Example 11.4. Using the example from 11.2,

$$\mathrm{lv}(5x_1 + x_2 + 7x_5 + 8) = x_1,$$
$$\mathrm{lv}(x_3 + 11x_4 + 2x_5 + 12) = x_3.$$

Definition 11.5. A list of linear polynomials F is in **triangular form** if for each $i < j$,

- $f_i = 0$ implies $f_j = 0$, while
- $f_i, f_j \neq 0$ implies $\mathrm{lv}(f_i) > \mathrm{lv}\left(f_j\right)$.

Example 11.6. Using the example from 11.2,the list

$$F \quad = \quad (\quad 5x_1 + x_2 + 7x_5 + 8, \quad x_3 + 11x_4 + 2x_5 + 12, \quad 3x_1 + 7x_2 + 8x_3 + 11 \quad)$$

is not in triangular form, since $\mathrm{lv}(f_1) = \mathrm{lv}(f_3) = x_1$, whereas we want $\mathrm{lv}(f_1) > \mathrm{lv}(f_3)$.

The list

$$G \quad = \quad (\quad x_1 + 6, \quad 0, \quad x_2 + 3x_4 \quad)$$

is also not in triangular form, because g_2 is zero while $g_3 \neq 0$.

However, the list

$$H \quad = \quad (\quad x_1 + 6, \quad x_2 + 3x_4, \quad 0 \quad)$$

is in triangular form, because $h_3 = 0$ and $\mathrm{lv}(h_1) > \mathrm{lv}(h_2)$.

Algorithm 7 describes one way to apply the method.

Theorem 11.7. Algorithm 7 terminates correctly.

Proof. All the loops of the algorithm are explicitly finite, so the algorithm terminates. To show that it terminates correctly, we must show both that G is triangular and that its roots are the roots of F.

That G is triangular: We claim that each iteration of the outer loop terminates with G in *i-subtriangular form*; by this we mean that

- the list $(g_1, \ldots, g_i)$ is in triangular form; *and*
- for each $j = 1, \ldots, i$ if $g_j \neq 0$ then the coefficient of $\mathrm{lv}\left(g_j\right)$ in $g_{i+1}, \ldots, g_m$ is 0.

Algorithm 7. Gaussian elimination

```
1:  inputs
2:    F = (f_1, f_2, ..., f_m), a list of linear polynomials in n variables, with coefficients from a field
      𝔽.
3:  outputs
4:    G = (g_1, g_2, ..., g_m), a list of linear polynomials in n variables, in triangular form, whose
      roots are precisely the roots of F.
5:  do
6:    Let G := F
7:    for i = 1, 2, ..., m − 1
8:      Rearrange g_i, g_{i+1}, ..., g_m so that for each k < ℓ, g_ℓ = 0, or lv(g_k) ≥ lv(g_ℓ)
9:      if g_i ≠ 0
10:       Denote the coefficient of lv(g_i) by a
11:       for j = i + 1, i + 2, ... m
12:         if lv(g_j) = lv(g_i)
13:           Denote the coefficient of lv(g_j) by b
14:           Replace g_j with a g_j − b g_i
15:   return G
```

Note that G is in triangular form if and only if G is in i-subtriangular form for all $i = 1, 2, \ldots, m$. This is fairly straightforward, since line 8 ensures that all the zero polynomials occur at the end of the list, as well as $\mathrm{lv}(g_i) > \mathrm{lv}(g_{i+j})$ for any $j \geq 1$.

Showing that G is equivalent to F is only a little harder. The combinations of F that produce G are all linear; that is, for each $j = 1, \ldots, m$ there exist $c_{i,j} \in \mathbb{F}$ such that

$$g_j = c_{1,j} f_1 + c_{2,j} f_2 + \cdots + a_{m,j} f_m.$$

Hence if $(\alpha_1, \ldots, \alpha_n) \in \mathbb{F}^n$ is a common root of F, it is also a common root of G. For the converse, observe from the algorithm that there exists some i such that $f_i = g_1$; then there exists some $j \in \{1, \ldots, m\} \setminus \{i\}$ and some $a, b \in \mathbb{F}$ such that $f_j = a g_1 - b g_2$; and so forth. Hence the elements of F are also a linear combination of the elements of G, and a similar argument shows that the common roots of G are common roots of F. □

Remark 11.8. There are other ways to define both triangular form and Gaussian elimination. Our method is perhaps stricter than necessary, but we have chosen this definition first to keep matters relatively simple, and second to assist us in the development of Gröbner bases.

Example 11.9. We use Algorithm 7 to illustrate Gaussian elimination for the system of equations described in Example 11.2.

- We start with the input,

$$F \quad = \quad (\quad 5x_1 + x_2 + 7x_5 + 8, \quad x_3 + 11x_4 + 2x_5 + 12, \quad 3x_1 + 7x_2 + 8x_3 + 11 \quad).$$

- Line 6 tells us to set $G = F$, so now

$$G \quad = \quad (\quad 5x_1 + x_2 + 7x_5 + 8, \quad x_3 + 11x_4 + 2x_5 + 12, \quad 3x_1 + 7x_2 + 8x_3 + 11 \quad).$$

- We now enter an *outer* loop:
  - In the first iteration, $i = 1$.
  - We rearrange G, obtaining

$$G \quad = \quad (\quad 5x_1 + x_2 + 7x_5 + 8, \quad 3x_1 + 7x_2 + 8x_3 + 11, \quad x_3 + 11x_4 + 2x_5 + 12 \quad).$$

  - Since $g_i \neq 0$, Line 10 tells us to denote a as the coefficient of $\mathrm{lv}(g_i)$, so $a = 5$.
  - We now enter an *inner* loop:
    - In the first iteration, $j = 2$.
    - Since $\mathrm{lv}\left(g_j\right) = \mathrm{lv}(g_i)$, denote b as the coefficient of $\mathrm{lv}\left(g_j\right)$; since $\mathrm{lv}\left(g_j\right) = x_1$, $b = 3$.
    - Replace g_j with

$$\begin{aligned} a g_j - b g_i &= 5(3x_1 + 7x_2 + 8x_3 + 11) \\ &\quad - 3(5x_1 + x_2 + 7x_5 + 8) \\ &= 32x_2 + 40x_3 - 21x_5 + 31. \end{aligned}$$

      Recall that the field is $\mathbb{Z}_{13}$, so we can rewrite this as

$$6x_2 + x_3 + 5x_5 + 5.$$

      We now have

$$G \quad = \quad (\quad 5x_1 + x_2 + 7x_5 + 8, \quad 6x_2 + x_3 + 5x_5 + 5, \quad x_3 + 11x_4 + 2x_5 + 12 \quad).$$

  - We continue with the inner loop:
    - In the second iteration, $j = 3$.
    - Since $\mathrm{lv}\left(g_j\right) \neq \mathrm{lv}(g_i)$, we do not proceed further.
  - Now $j = 3 = m$, and the inner loop is finished.
- We continue with the outer loop:
  - In the second iteration, $i = 2$.
  - We do not rearrange G, as it is already in the form indicated. (In fact, it is in triangular form already, but the algorithm does not "know" this yet.)
  - Since $g_i \neq 0$, Line 10 tells us to denote a as the coefficient of $\mathrm{lv}(g_i)$; since $\mathrm{lv}(g_i) = x_2$, $a = 6$.
  - We now enter an *inner* loop:
    - In the first iteration, $j = 2$.
    - Since $\mathrm{lv}\left(g_j\right) \neq \mathrm{lv}(g_i)$, we do not proceed with this iteration.
  - Now $j = 3 = m$, and the inner loop is finished.
- Now $i = 2 = m - 1$, and the outer loop is finished.
- We return G, which is in triangular form!

Once we have the triangular form of a linear system, it is easy to answer the five natural questions.

Theorem 11.10. Let $G = (g_1, g_2, \ldots, g_m)$ is a list of nonzero linear polynomials in n variables over a field $\mathbb{F}$. If G is in triangular form, then each of the following holds.

(A) G has common solutions if and only if none of the g_i is a constant.

(B) G has finitely many common solutions if and only if G has a solution and $m = n$. In this case, there is exactly one solution.

(C) G has common solutions of dimension d if and only if G has a solution and $d = n - m$.

A proof of Theorem 11.10 can be found in any textbook on linear algebra, although probably not in one place.

Example 11.11. Continuing with the system that we have used in this section, we found that a triangular form of

$$F = (\; 5x_1 + x_2 + 7x_5 + 8, \quad x_3 + 11x_4 + 2x_5 + 12, \quad 3x_1 + 7x_2 + 8x_3 + 11 \;)$$

is

$$G = (\; 5x_1 + x_2 + 7x_5 + 8, \quad 6x_2 + x_3 + 5x_5 + 5, \quad x_3 + 11x_4 + 2x_5 + 12 \;).$$

Theorem 11.10 implies that

(A) G has a solution, because none of the g_i is a constant.

(B) G has infinitely many solutions, because the number of polynomials ($m = 3$) is *not* the same as the number of variables ($n = 5$).

(C) G has solutions of dimension $d = n - m = 2$.

Lexicographic order allows us to parametrize the solution set easily. Let $s, t \in \mathbb{Z}_{13}$ be arbitrary, and let $x_4 = s$ and $x_5 = t$. Back-substituting in S, we have:

- From $g_3 = 0$, $x_3 = 2s + 11t + 1$.
- From $g_2 = 0$,
$$6x_2 = 12x_3 + 8t + 8. \tag{35}$$
The Euclidean algorithm helps us derive the multiplicative inverse of 6 in $\mathbb{Z}_2$; we get 11. Multiplying both sides of (35) by 11, we have
$$x_2 = 2x_3 + 10t + 10.$$
Recall that we found $x_3 = 2s + 11t + 1$, so
$$x_2 = 2(2s + 11t + 1) + 10t + 10 = 4s + 6t + 12.$$
- From $g_1 = 0$,
$$5x_1 = 12x_2 + 6x_5 + 5.$$
Repeating the process that we carried out in the previous step, we find that
$$x_1 = 7s + 9.$$

We can verify this solution by substituting it into the original system:

$$\begin{aligned}
f_1: \quad & 5(7s+9)+(4s+6t+12)+7t+8 \\
& = (9s+6)+4s+20 \\
& = 0 \\
f_2: \quad & (2s+11t+1)+11s+2t+12 \\
& = 0 \\
f_3: \quad & 3(7s+9)+7(4s+6t+12)+8(2s+11t+1)+11 \\
& = (8s+1)+(2s+3t+6)+(3s+10t+8)+11 \\
& = 0.
\end{aligned}$$

Before proceeding to the next section, study the proof of Theorem 11.7 carefully. Think about how we might relate these ideas to non-linear polynomials.

Exercises.

Exercise 11.12. A *homogeneous linear system* is one where none of the polynomials has a constant term: that is, $b_i = 0$ for $i = 1, \ldots, m$. Explain why homogeneous systems always have at least one solution.

Exercise 11.13. Find the triangular form of the following linear systems, and use it to find the common solutions of the corresponding system of equations (if any).
(a) $f_1 = 3x + 2y - z - 1$, $f_2 = 8x + 3y - 2z$, and $f_3 = 2x + z - 3$; over the field $\mathbb{Z}_7$.
(b) $f_1 = 5a + b - c + 1$, $f_2 = 3a + 2b - 1$, $f_3 = 2a - b - c + 1$; over the same field.
(c) The same system as (a), over the field $\mathbb{Q}$.

Exercise 11.14. In linear algebra you also used matrices to solve linear systems, by rewriting them in echelon (or triangular) form. Do the same with system (a) of the previous exercise.

Exercise 11.15. Does Algorithm 7 also terminate correctly if the coefficients of F are not from a field, but from an integral domain? If so, and if $m = n$, can we then solve the resulting triangular system G for the roots of F as easily as if the coefficients were from a field? Why or why not?

11.2: Monomial orderings

Before looking at how we might analyze systems of nonlinear polynomial equations, we consider the question of identifying the "most important" monomial in this more general setting. With linear polynomials, it was relatively easy; we picked the variable with the smallest index.

But which monomial should be the *leading* monomial of $x + y^3 - 4y$? It seems clear enough that y should not be the leading term, since it divides y^3, and as such does not "lead" even if there were no x's to reckon with. With x and y^3, however, things are less clear.

Recall from Section 7.3 the definition of $\mathbb{M}$, the set of monomials over $x_1, x_2, \ldots, x_n$.

Definition 11.16. Let $t, u \in \mathbb{M}$. The **lexicographic ordering** orders $t > u$ if

- $\deg_{x_1} t > \deg_{x_1} u$, or
- $\deg_{x_1} t = \deg_{x_1} u$ and $\deg_{x_2} t > \deg_{x_2} u$, or
- $\dots$
- $\deg_{x_i} t = \deg_{x_i} u$ for $i = 1, 2, \dots, n-1$ and $\deg_{x_n} t > \deg_{x_n} u$.

Another way of saying this is that $t > u$ iff there exists i such that

- $\deg_{x_j} t = \deg_{x_j} u$ for all $j = 1, 2, \dots, i-1$, and
- $\deg_{x_i} t > \deg_{x_i} u$.

The **leading monomial** of a non-zero polynomial p is any monomial t such that $t > u$ for all other terms u of p. *The leading monomial of 0 is left undefined.*

Notation 11.17. We denote the leading monomial of a polynomial p as $\operatorname{lm}(p)$.

Example 11.18. Using the lexicographic ordering over x, y,

$$\begin{aligned} \operatorname{lm}\left(x^2 + y^2 - 4\right) &= x^2 \\ \operatorname{lm}(xy - 1) &= xy \\ \operatorname{lm}\left(x + y^3 - 4y\right) &= x. \end{aligned}$$

Before proceeding, we should prove a few simple, but important, properties of the lexicographic ordering.

Proposition 11.19. The lexicographic ordering on $\mathbb{M}$

(A) is a linear ordering;
(B) is **compatible with divisibility**: for any $t, u \in \mathbb{M}$, if $t \mid u$, then $t \leq u$;
(C) is **compatible with multiplication**: for any $t, u, v \in \mathbb{M}$, if $t < u$, then for any monomial v over $\mathbf{x}$, $tv < uv$;
(D) orders $1 \leq t$ for any $t \in \mathbb{M}$; and
(E) is a well ordering.

(Recall that we defined a monoid way back in Section 1.1, and used $\mathbb{M}$ as an example.)

Proof. For (A), suppose that $t \neq u$. Then there exists i such that $\deg_{x_i} t \neq \deg_{x_i} u$. Pick the smallest i for which this is true; then $\deg_{x_j} t = \deg_{x_j} u$ for $j = 1, 2, \dots, i-1$. If $\deg_{x_i} t < \deg_{x_i} u$, then $t < u$; otherwise, $\deg_{x_i} t > \deg_{x_i} u$, so $t > u$.

For (B), we know that $t \mid u$ iff $\deg_{x_i} t \leq \deg_{x_i} u$ for all $i = 1, 2, \dots, m$. Hence $t \leq u$.

For (C), assume that $t < u$. Let i be such that $\deg_{x_j} t = \deg_{x_j} u$ for all $j = 1, 2, \dots, i-1$ and

$\deg_{x_i} t < \deg_{x_i} u$. For any $\forall j = 1, 2, \ldots, i-1$, we have

$$\begin{aligned}\deg_{x_j}(tv) &= \deg_{x_j} t + \deg_{x_j} v \\ &= \deg_{x_j} u + \deg_{x_j} v \\ &= \deg_{x_j} uv\end{aligned}$$

and

$$\begin{aligned}\deg_{x_i}(tv) &= \deg_{x_i} t + \deg_{x_i} v \\ &< \deg_{x_i} u + \deg_{x_i} v = \deg_{x_i} uv.\end{aligned}$$

Hence $tv < uv$.

(D) is a special case of (B).

For (E), let $M \subset \mathbb{M}$. We proceed by induction on the number of variables n.

For the inductive base, if $n = 1$ then the monomials are ordered according to the exponent on x_1, which is a natural number. Let E be the set of all exponents of the monomials in M; then $E \subset \mathbb{N}$. Recall that $\mathbb{N}$ is well-ordered. Hence E has a least element; call it e. By definition of E, e is the exponent of some monomial m of M. Since $e \leq \alpha$ for any other exponent $x^\alpha \in M$, m is a least element of M.

For the inductive hypothesis, assume that for all $i < n$, the set of monomials in i variables is well-ordered.

For the inductive step, let N be the set of all monomials in $n-1$ variables such that for each $t \in N$, there exists $m \in M$ such that $m = t \cdot x_n^e$ for some $e \in \mathbb{N}$. By the inductive hypothesis, N has a least element; call it t. Let

$$P = \left\{ t \cdot x_n^e : \ t \cdot x_n^e \in M \ \exists e \in \mathbb{N} \right\}.$$

All the elements of P are equal in the first $n-1$ variables: their exponents are the exponents of t. Let E be the set of all exponents of x_n for any monomial $u \in P$. As before, $E \subset \mathbb{N}$. Hence E has a least element; call it e. By definition of E, there exists $u \in P$ such that $u = t \cdot x_n^e$; since $e \leq \alpha$ for all $\alpha \in E$, u is a least element of P.

Finally, let $v \in M$. Since t is minimal in N, either there exists i such that

$$\deg_{x_j} u = \deg_{x_j} t = \deg_{x_j} v \quad \forall j = 1, \ldots, i-1$$
$$\text{and}$$
$$\deg_{x_i} u = \deg_{x_i} t < \deg_{x_i} v,$$

or

$$\deg_{x_j} u = \deg_{x_j} t = \deg_{x_j} v \quad \forall j = 1, 2, \ldots, n-1$$

In the first case, $u < v$ by definition. Otherwise, since e is minimal in E,

$$\deg_{x_n} u = e \leq \deg_{x_n} v,$$

in which case $u \leq v$. Hence u is a least element of M.

Since M is arbitrary in $\mathbb{M}$, every subset of $\mathbb{M}$ has a least element. Hence $\mathbb{M}$ is well-ordered. □

Before we start looking for a triangular form of non-linear systems, let's observe one more thing.

Proposition 11.20. Let p be a polynomial in the variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. If $\operatorname{lm}(p) = x_i^\alpha$, then every other monomial u of p has the form

$$u = \prod_{j=i}^{n} x_j^{\beta_j}$$

where $\beta_i < \alpha$.

Proof. Assume that $\operatorname{lm}(p) = x_i^\alpha$. Let u be any monomial of p. Write

$$u = \prod_{j=1}^{n} x_j^{\beta_j}$$

for appropriate $\beta_j \in \mathbb{N}$. Since $u < \operatorname{lm}(p)$, the definition of the lexicographic ordering implies that

$$\deg_{x_j} u = \deg_{x_j} \operatorname{lm}(p) = \deg_{x_j} x_i^\alpha \quad \forall j = 1, 2, \ldots, i-1$$

and

$$\deg_{x_i} u < \deg_{x_i} t.$$

Hence u has the form claimed. □

We now identify and generalize the properties of Proposition 11.19 to a generic ordering on monomials.

Definition 11.21. An **admissible ordering** $<$ on $\mathbb{M}$ is a linear ordering which is compatible with divisibility and multiplication.

By definition, properties (A) and (B) of Proposition 11.19 hold for an admissible ordering. What of the others?

Proposition 11.22. The following properties of an admissible ordering all hold.

(A) $1 \leq t$ for all $t \in \mathbb{M}$.

(B) The set $\mathbb{M}$ of all monomials over $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is well-ordered by any admissible ordering. That is, every subset M of $\mathbb{M}$ has a least element.

Proof. Let $<$ be any admissible ordering.

For (A), you do it! See Exercise 11.32.

For (B), let $t, u \in \mathbb{M}$. By (A), we know that $1 \leq u$. By the ordering's compatibility with multiplication, we know that $t \cdot 1 \leq t \cdot u$, or $t \leq tu$, satisfying compatibility with divisibility.

For (C), let $M \subseteq \mathbb{M}$ and let A be the smallest absorbing subset of $\mathbb{M}$ that contains M (you might want to refamiliarize yourself with absorbing subsets, which we studied in Section 1.4). Dickson's Lemma (Theorem on page 60) tells us that A has a finite generating set; call it T. In fact, $T \subseteq M$, as the definition of absorption means that *every* element of A is divisible by an element of M. There are only finitely many elements of T, so the linear ordering property of $<$ implies that that we can identify a smallest element, t. Let $u \in M$; by definition, $u \in A$, so we can find $v \in T$ such that v divides u. Since $t \leq v$, we use compatibility with divisibility to see that $t \leq v \leq u$. We chose u as an arbitrary element of M, so t is minimal in M. We chose M as an arbitrary subset of $\mathbb{M}$, so $\mathbb{M}$ is well-ordered by $<$. □

We can now introduce an ordering that you haven't seen before.

Definition 11.23. For a monomial t, the **total degree** of t is the sum of the exponents, denoted $\operatorname{tdeg}(t)$. For two monomials t, u, a **total-degree ordering** orders $t < u$ whenever $\operatorname{tdeg}(t) < \operatorname{tdeg}(u)$.

Example 11.24. The total degree of x^3y^2 is 5, and $x^3y^2 < xy^5$.

A simple total degree ordering is not itself admissible, because not it is not linear.

Example 11.25. We cannot order x^3y^2 and x^2y^3 by total degree alone, because $\operatorname{tdeg}(x^3y^2) = \operatorname{tdeg}(x^2y^3)$ but $x^3y^2 \neq x^2y^3$.

When there is a tie in the total degree, we need to fall back on another method. An interesting way of doing this is the following.

Definition 11.26. For two monomials t, u the **graded reverse lexicographic ordering**, or **grevlex**, orders $t < u$ whenever

- $\operatorname{tdeg}(t) < \operatorname{tdeg}(u)$, or
- $\operatorname{tdeg}(t) = \operatorname{tdeg}(u)$ *and* there exists $i \in \{1, \ldots, n\}$ such that for all $j = i+1, \ldots, n$
  - $\deg_{x_j} t = \deg_{x_j} u$, and
  - $\deg_{x_i} t > \deg_{x_i} u$.

Notice that to break a total-degree tie, grevlex reverses the lexicographic ordering in a double way: it searches *backwards* for the *smallest* degree, and designates the winner as the larger monomial.

Example 11.27. Under grevlex, $x^3y^2 > x^2y^3$ because the total degrees are the same and $y^2 < y^3$.

Theorem 11.28. The graded reverse lexicographic ordering is an admissible ordering.

Proof. Let $t, u \in \mathbb{M}$.

Linear ordering? Assume $t \neq u$; by definition, there exists $i \in \mathbb{N}^+$ such that $\deg_{x_i} t \neq \deg_{x_i} u$. Choose the largest such i, so that $\deg_{x_j} t = \deg_{x_j} u$ for all $j = i+1, \ldots, n$. Then $t < u$ if $\deg_{x_i} t < \deg_{x_i} u$; otherwise $u < t$.

Compatible with divisibility? Assume $t \mid u$. By definition, $\deg_{x_i} t \leq \deg_{x_i} u$ for all $i = 1, \ldots, n$. If $t = u$, then we're done. Otherwise, $t \neq u$. If $\operatorname{tdeg}(t) > \operatorname{tdeg}(u)$, then the fact that the degrees

are all natural numbers implies (see Exercise) that for some $i = 1, \ldots, n$ we have $\deg_{x_i} t > \deg_{x_i} u$, contradicting the hypothesis that $t \mid u$! Hence $\operatorname{tdeg}(t) = \operatorname{tdeg}(u)$. Since $t \neq u$, there exists $i \in \{1, \ldots, n\}$ such that $\deg_{x_i} t \neq \deg_{x_i} u$. Choose the largest such i, so that $\deg_{x_j} t = \deg_{x_j} u$ for $j = i+1, \ldots, n$. Since $t \mid u$, $\deg_{x_i} t < \deg_{x_i} u$, and $\deg_{x_j} t \leq \deg_{x_j} u$. Hence

$$\begin{aligned}
\operatorname{tdeg}(t) &= \sum_{j=1}^{i-1} \deg_{x_j} t + \deg_{x_i} t + \sum_{j=i+1}^{n} \deg_{x_j} t \\
&= \sum_{j=1}^{i-1} \deg_{x_j} t + \deg_{x_i} t + \sum_{j=i+1}^{n} \deg_{x_j} u \\
&\leq \sum_{j=1}^{i-1} \deg_{x_j} u + \deg_{x_i} t + \sum_{j=i+1}^{n} \deg_{x_j} u \\
&< \sum_{j=1}^{i-1} \deg_{x_j} u + \deg_{x_i} u + \sum_{j=i+1}^{n} \deg_{x_j} u \\
&= \operatorname{tdeg}(u).
\end{aligned}$$

Hence $t < u$.

Compatible with multiplication? Assume $t < u$, and let $v \in \mathbb{M}$. By definition, $\operatorname{tdeg}(t) < \operatorname{tdeg}(u)$ or there exists $i \in \{1, 2, \ldots, n\}$ such that $\deg_{x_i} t > \deg_{x_i} u$ and $\deg_{x_j} t = \deg_{x_j} u$ for all $j = i+1, \ldots, n$. In the first case, you will show in the exercises that

$$\begin{aligned}
\operatorname{tdeg}(tv) &= \operatorname{tdeg}(t) + \operatorname{tdeg}(v) \\
&< \operatorname{tdeg}(u) + \operatorname{tdeg}(v) = \operatorname{tdeg}(uv).
\end{aligned}$$

In the second,

$$\deg_{x_i} tv = \deg_{x_i} t + \deg_{x_i} v > \deg_{x_i} u + \deg_{x_i} v = \deg_{x_i} uv$$

while

$$\deg_{x_j} tv = \deg_{x_j} t + \deg_{x_j} v = \deg_{x_j} u + \deg_{x_j} v = \deg_{x_j} uv.$$

In either case, $tv < uv$ as needed. □

A useful tool when dealing with monomial orderings is a **monomial diagram**. These are most useful for monomials in a bivariate polynomial ring $\mathbb{F}[x, y]$, but we can often imagine important aspects of these diagrams in multivariate rings, as well. We discuss the bivariate case here.

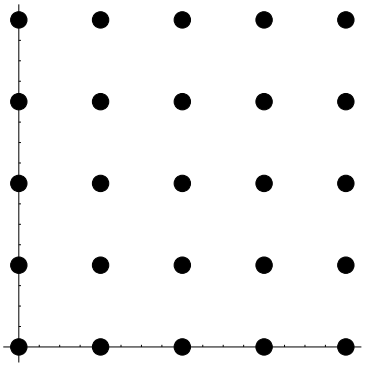
Definition 11.29. Let $t \in \mathbb{M}$. Define the **exponent vector** $(\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ where $\alpha_i = \deg_{x_i} t$.

Let $t \in \mathbb{F}[x, y]$ be a monomial, and (α, β) its exponent vector. That is,

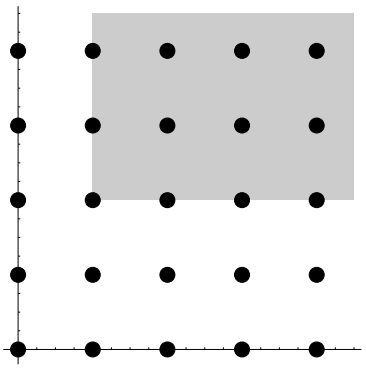
$$t = x^{\alpha} y^{\beta}.$$

We can consider (α, β) as a point in the x-y plane. If we do this with all the monomials of

$\mathbb{M} \subset \mathbb{F}[x, y]$, and we obtain the following diagram:



This diagram is not especially useful, aside from pointing out that the monomial x^2 is the third point on the left in the bottom row, and the monomial 1 is the point in the lower left corner. What does make diagrams like this useful is the fact that if $t \mid u$, then the point corresponding to u lies above and/or to the right of the point corresponding to t, but *never* below or to the left of it. We often shade the points corresponding monomials divisible by a given monomial; for example, the points corresponding to monomials divisible by xy^2 lie within the shaded region of the following diagram:

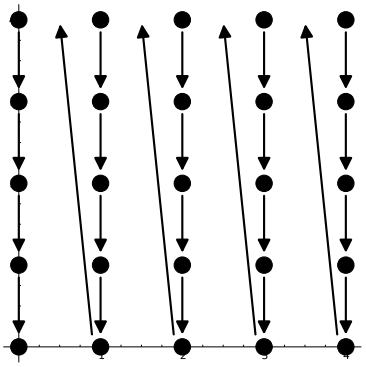


As we will see later, diagrams such as the one above can come in handy when visualizing certain features of an ideal.

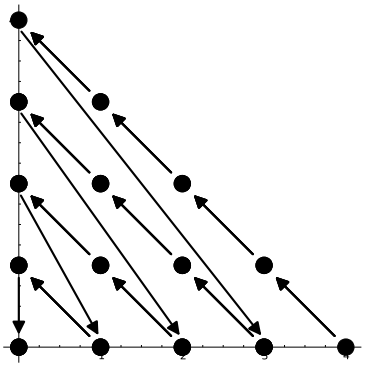
What interests us most for now is that we can sketch vectors on a monomial diagram that show the ordering of the monomials.

Example 11.30. We sketch monomial diagrams that show how lex and grevlex order $\mathbb{M}$. We already know that the smallest monomial is 1. The next smallest will always be y.

For the *lex* order, $y^a < x$ for *every* choice of $a \in \mathbb{N}$, no matter how large. Hence the next largest monomial is y^2, followed by y^3, etc. Once we have marked every power of y, the next largest monomial is x, followed by xy, by xy^2, etc., for $xy^a < x^2$ for all $a \in \mathbb{N}$. Continuing in this fashion, we have the following diagram:



With the *grevlex* order, by contrast, the next largest monomial after y is x, since $\operatorname{tdeg}(x) < \operatorname{tdeg}(y^2)$. After x come y^2, xy, and x^2, in that order, followed by the degree-three monomials y^2, xy^2, x^2y, and x^3, again in that order. This leads to the following monomial diagram:



These diagrams illustrate an important and useful fact.

Theorem 11.31. Let $t \in \mathbb{M}$.

(A) In the lexicographic order, there are infinitely many monomials smaller than t if and only if t is not a power of x_n alone.

(B) In the grevlex order, there are finitely many monomials smaller than t.

Proof. You do it! See Exercise 11.36. □

Exercises.

Exercise 11.32. Show that for any admissible ordering and any $t \in \mathbb{M}$, $1 \leq t$.

Exercise 11.33. The **graded lexicographic order**, which we will denote by **gralex**, orders $t < u$ if

- $\operatorname{tdeg}(t) < \operatorname{tdeg}(u)$, or
- $\operatorname{tdeg}(t) = \operatorname{tdeg}(u)$ and the lexicographic ordering would place $t < u$.

(a) Order x^2y, xy^2, and z^5 by gralex.

(b) Show that gralex is an admissible order.

(d) Sketch a monomial diagram that shows how gralex orders $\mathbb{M}$.

Exercise 11.34. Define $\pi_{\leq i}$ as the map from $\mathbb{M}$ to itself that "projects" a monomial in n variables to a monomial in i variables. For example,

$$\pi_{\leq 3}\left(x_1^5 x_2^4 x_4 x_5^2\right) = x_1^5 x_2^4.$$

We can think of $\pi_{\leq i}$ as "chopping" variables $x_{i+1}, x_{i+2}, \ldots, x_n$ off the monomial. More formally, if $0 < i \leq n$, then

$$\pi_{\leq i} : \mathbb{M}_m \to \mathbb{M}_i \quad \text{by} \quad \pi_{\leq i}\left(x_1^{a_1} \cdots x_n^{a_n}\right) = x_1^{a_1} \cdots x_i^{a_i}.$$

Show that the definition of the grevlex ordering is equivalent to the following:

Definition 11.35 (Alternate definition of grevlex)**.** We say that $t < u$ if $\operatorname{tdeg}\left(\pi_{\leq i}(t)\right) = \operatorname{tdeg}\left(\pi_{\leq i}(u)\right)$ for $i = n, n-1, \ldots, k+1$ but $\operatorname{tdeg}(\pi_i(t)) < \operatorname{tdeg}(\pi_i(u))$.

Exercise 11.36. Prove Theorem 11.31.

11.3: Matrix representations of monomial orderings

In fact, there are limitless ways to design an admissible ordering.

Example 11.37. Consider the matrix

$$M = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ & & & & -1 \\ & & & -1 & \\ & & \cdots & & \\ & -1 & & & \end{pmatrix}$$

where the empty entries are zeroes. We claim that M represents the grevlex ordering, and weighted vectors computed with M can be read from top to bottom, where the first entry that does not tie determines the larger monomial.

Why? The top row of M adds all the elements of the exponent vector, so the top entry of the weighted vector is the total degree of the monomial. If the two monomials have different total degrees, the top entry of the weighted vector determines the larger monomial. In case they have the same total degree, the second entry of $M\mathbf{t}$ contains $-\deg_{x_n} t$, so if they have different degree in the smallest variable, the second entry determines the larger monomial. And so forth.

The monomials $t = x^3y^2$, $u = x^2y^3$, and $v = xy^5$ have exponent vectors $\mathbf{t} = (3,2)$, $\mathbf{u} = (2,3)$, and $\mathbf{v} = (1,5)$, respectively. We have

$$M\mathbf{t} = \begin{pmatrix} 5 \\ -2 \end{pmatrix}, \qquad M\mathbf{u} = \begin{pmatrix} 5 \\ -3 \end{pmatrix}, \qquad M\mathbf{v} = \begin{pmatrix} 6 \\ -5 \end{pmatrix},$$

from which we conclude that $v > t > u$.

Definition 11.38. Let $M \in \mathbb{R}^{n\times n}$. If $\mathbf{t} \in \mathbb{N}^n$, the **weight** of $\mathbf{t}$ is $w(\mathbf{t}) = M\mathbf{t}$. Similarly, if $t \in \mathbb{M}_n$, the **weight** of t is the weight of its exponent vector.

Not all matrices can represent admissible orderings.

Theorem 11.39. Let $M \in \mathbb{R}^{m\times m}$. M represents a admissible ordering if and only if its rows are linearly independent over $\mathbb{Z}$ and the topmost nonzero entry in each column is positive.

To prove the theorem, we need the following lemma.

Lemma 11.40. If M satisfies the criteria of Theorem 11.39, then there exists a matrix N that satisfies (B), whose entries are all nonnegative, and for all $\mathbf{t} \in \mathbb{Z}^n$ comparison from top to bottom implies that $N\mathbf{t} > N\mathbf{u}$ if and only if $M\mathbf{t} > M\mathbf{u}$.

Example 11.41. In Example 11.37, we saw that grevlex could be represented by

$$M = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ & & & & -1 \\ & & & -1 & \\ & & \cdots & & \\ & -1 & & & \end{pmatrix}.$$

However, it can also be represented by

$$N = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 & \\ & \cdots & & & \\ 1 & 1 & & & \\ 1 & & & & \end{pmatrix}$$

where the empty entries are, again, zeroes. Notice that the first row operates exactly the same, while the second row adds all the entries *except the last*. If $t_n < y_n$ then from $t_1 + \cdots + t_n = u_1 + \cdots + u_n$ we infer that $t_1 + \cdots + t_{n-1} > u_1 + \cdots + u_{n-1}$, so the second row of $N\mathbf{t}$ and $N\mathbf{u}$ would break the tie in exactly the same way as the second row of $M\mathbf{t}$ and $M\mathbf{u}$. And so forth.

Remark 11.42.

1. We can obtain N by adding row 1 of M to row 2 of M, then adding the modified row 2 of M to the modified row 3, and so forth. This is the essence of the proof of Lemma 11.40.
2. While M corresponds to our original definition of grevlex ordering, N corresponds to the definition given in Exercise 11.34

Proof of Lemma 11.40. Let $M \in \mathbb{R}^{n\times n}$ satisfy the criteria of Theorem 11.39. Construct N by building matrices $M_0, M_1, \ldots$ in the following way.

Let $M_1 = M$. Suppose that $M_1, M_2, \ldots, M_{i-1}$ all have nonnegative entries in rows 1, 2, …, $i-1$ but M has a negative entry α in row i, column j. By hypothesis, the topmost nonzero entry

β of column j in M_{i-1} is positive; say it is in row k. Use the Archimedean property of $\mathbb{R}$ to find $K \in \mathbb{N}^+$ such that $K\beta \geq |\alpha|$, and add K times row k of M_{i-1} to row j. The entry in row i and column j of M_i is now nonnegative. If there were other negative values in row i of M_i, the fact that row k of M_{i-1} contained nonnegative entries implies that the absolute values of these negative entries are no larger than before. There is a finite number of entries in each row, and a finite number of rows in M, so this process terminates after finitely many additions with a matrix N whose entries are all nonnegative.

In addition, we can write the ith row $N_{(i)}$ of N as

$$N_{(i)} = K_1 M_{(1)} + K_2 M_{(2)} + \cdots + K_i M_{(i)}$$

where $M_{(k)}$ indicates the kth row of M. For any $\mathbf{t} \in \mathbb{M}$, the ith entry of $N\mathbf{t}$ is therefore

$$\begin{aligned} N_{(i)}\mathbf{t} &= \left(K_1 M_{(1)} + K_2 M_{(2)} + \cdots + K_i M_{(i)}\right)\mathbf{t} \\ &= K_1\left(M_{(1)}\mathbf{t}\right) + K_2\left(M_{(2)}\mathbf{t}\right) + \cdots + K_i\left(M_{(i)}\mathbf{t}\right). \end{aligned}$$

We see that if $M_{(1)}\mathbf{t} = \cdots = M_{(i-1)}\mathbf{t} = 0$ and $M_{(i)}\mathbf{t} = \alpha \neq 0$, then $N_{(1)}\mathbf{t} = \cdots = N_{(i-1)}\mathbf{t} = 0$ and $N_{(i)}\mathbf{t} = K_i\alpha \neq 0$. Hence $N\mathbf{t} > N\mathbf{u}$ if and only if $N_{(i)}\mathbf{t} > N_{(i)}\mathbf{u}$ if and only if $K_i\alpha > K_i\left(M_{(i)}\mathbf{u}\right)$ if and only if $M\mathbf{t} > M\mathbf{u}$. □

Now we can prove Theorem 11.39.

Proof of Theorem 11.39. *That (A) implies (B):* Assume that M represents an admissible ordering.

The monomial 1 has the exponent vector $\mathbf{t} = (0,\ldots,0)$, while the monomial x_i has the exponent vector $\mathbf{u}$ with zeroes everywhere except in the ith position. The product $M\mathbf{t} > M\mathbf{u}$ if the ith element of the top row of M is negative, but this contradicts Proposition 11.22(A).

In addition, property of Definition 11.21 implies that no pair of distinct monomials can produce the same weighted vector. Hence the rows of M are linearly independent over $\mathbb{Z}$.

That (B) implies (A): Assume that M satisfies the criteria of the theorem. We need to show that the properties of an admissible order (Definition 11.21) are satisfied.

Linear ordering? Since the rows of M are linearly independent over $\mathbb{Z}$, every pair of monomials t and u produces a pair of distinct weighted vectors $M\mathbf{t}$ and $M\mathbf{u}$ if and only if $t \neq u$. Reading these vectors from top to bottom allows us to decide whether $t > u$, $t < u$, or $t = u$.

Compatible with divisibility? This follows from linear algebra. Let $t, u \in \mathbb{M}$, and assume that $t \mid u$. Then $\deg_{x_i} t \leq \deg_{x_i} u$ for all $i = 1, 2, \ldots, n$. In the exponent vectors $\mathbf{t}$ and $\mathbf{u}$, $t_i \leq u_i$ for each i. Let $\mathbf{v} \in \mathbb{N}^n$ such that $\mathbf{u} = \mathbf{t} + \mathbf{v}$; then

$$M\mathbf{u} = M\left(\mathbf{t} + \mathbf{v}\right) = M\mathbf{t} + M\mathbf{v}.$$

From Lemma 11.40 we can assume that the entries of M are all nonnegative. Thus the entries of $M\mathbf{u}$, $M\mathbf{t}$, and $M\mathbf{v}$ are also nonnegative. Thus the topmost nonzero entry of $M\mathbf{v}$ is positive, and $M\mathbf{u} > M\mathbf{t}$.

Compatible with multiplication? This is similar to compatibility with divisibility, so we omit it. □

In the Exercises you will find other matrices that represent term orderings, some of them somewhat exotic.

Exercises

Exercise 11.43. Find a matrix that represents (a) the lexicographic term ordering, and (b) the gralex ordering.

Exercise 11.44. Explain why the matrix

$$M = \begin{pmatrix} 1 & 1 & & & & & & & \\ 1 & 1 & 1 & & & & & & \\ 1 & 1 & 1 & 1 & & & & & \\ -1 & & & & & & & & \\ & & & & 1 & 1 & 1 & 1 \\ & & & & & 1 & 1 & 1 \\ & & & & & & 1 & 1 \\ & & & & & & & -1 \end{pmatrix}$$

represents an admissible ordering. Use M to order the monomials

$$x_1 x_3^2 x_4 x_6, \quad x_1 x_4^8 x_7, \quad x_2 x_3^2 x_4 x_6, \quad x_8, \quad , x_8^2, \quad x_7 x_8.$$

Exercise 11.45. Suppose you know nothing about an admissible order $<$ on $\mathbb{F}[x,y]$ except that $x > y$ and $x^2 < y^3$. Find a matrix that represents this order.

11.4: The structure of a Gröbner basis

Throughout this section, assume an admissible ordering of monomials.

When we consider the non-linear case, things become a little more complicated. Consider the following system of equations:

$$\begin{aligned} x^2 + y^2 &= 4 \\ xy &= 1. \end{aligned}$$

We can visualize the real solutions to this system; see Figure 11.1 on the following page. The common solutions occur wherever the circle and the hyperbola intersect. We see four intersections in the real plane; one of them is highlighted with a dot.

However, we don't know if complex solutions exist. In addition, plotting equations involving more than two variables is difficult, and more than three is effectively impossible. Finally, while it's relatively easy to solve the system given above, it isn't a "triangular" system in the sense that the last equation is only in one variable. So we can't solve for one variable immediately and then go backwards. We can solve for y in terms of x, but not for an exact value of y.

It gets worse! Although the system is triangular in a "linear" sense, it is not triangular in a non-linear sense: we can multiply the two polynomials above by monomials and obtain a new polynomial that isn't obviously spanned by either of these two:

$$y\left(x^2 + y^2 - 4\right) - x\left(xy - 1\right) = x + y^3 - 4y. \tag{36}$$

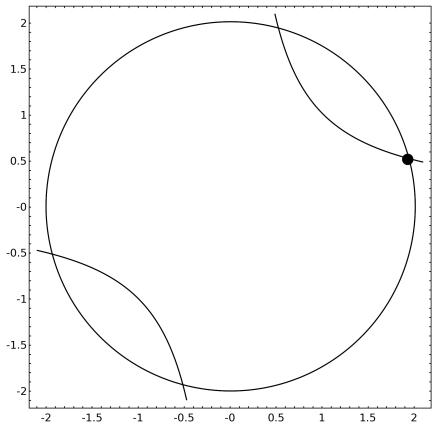


Figure 11.1. Plots of $x^2 + y^2 = 4$ and $xy = 1$

None of the terms of this new polynomial appears in either of the original polynomials. This sort of thing does *not* happen in the linear case, largely because

- cancellation of *variables* can be resolved using *scalar multiplication*, hence in a vector space; but
- cancellation of *terms* cannot be resolved without *monomial multiplication*, hence it requires an ideal.

So we need to find a "triangular form" for non-linear systems.

Let's rephrase this problem in the language of rings and ideals. The primary issue we would like to resolve is the one that we observed immediately after computing the subtraction polynomial of equation (36): we built a polynomial p whose leading term x was not divisible by the leading term of either the hyperbola (xy) or the circle (x^2). When we built p, we used operations of the polynomial ring that allowed us to remain within the ideal generated by the hyperbola and the circle. That is,

$$p = x + y^3 - 4y = y\left(x^2 + y^2 - 4\right) - x\left(xy - 1\right);$$

by Theorem 8.7 ideals absorb multiplication and are closed under subtraction, so

$$p \in \left\langle x^2 + y^2 - 4, xy - 1 \right\rangle.$$

So one problem appears to be that p is in the ideal, but its leading monomial is not divisible by the leading monomials of the ideal's basis. Let's define a special kind of ideal basis that will not give us this problem.

Definition 11.46. Let G be a basis of an ideal I. We call it a **Gröbner basis of** I if for every $p \in I$, we can find $g \in G$ such that $\operatorname{lm}(g) \mid \operatorname{lm}(p)$.

It isn't obvious at the moment how we can decide that any given basis forms a Gröbner basis, because there are infinitely many polynomials that we'd have to check. However, we can certainly determine that the list

$$\left(x^2 + y^2 - 4, xy - 1\right)$$

is *not* a Gröbner basis, because we found a polynomial in its ideal that violated the definition of a Gröbner basis: $x + y^3 - 4y$.

How did we find that polynomial? We built a *subtraction polynomial* that was calculated in such a way as to "raise" the polynomials to the lowest level where their leading monomials would cancel! Let t, u be monomials in the variables $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Write $t = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ and $u = x_1^{\beta_1} x_2^{\beta_2} \cdots x_n^{\beta_n}$. Any common multiple of t and u must have the form

$$v = x_1^{\gamma_1} x_2^{\gamma_2} \cdots x_n^{\gamma_n}$$

where $\gamma_i \geq \alpha_i$ and $\gamma_i \geq \beta_i$ for each $i = 1, 2, \ldots, n$. We can thus identify a **least common multiple**

$$\operatorname{lcm}(t, u) = x_1^{\gamma_1} x_2^{\gamma_2} \cdots x_n^{\gamma_n}$$

where $\gamma_i = \max(\alpha_i, \beta_i)$ for each $i = 1, 2, \ldots, n$. It really is the *least* because no common multiple can have a smaller degree in any of the variables, and so it is smallest by the definition of the lexicographic ordering.

Lemma 11.47. For any two polynomials $p, q \in \mathbb{F}[x_1, x_2, \ldots, x_n]$, with $\operatorname{lm}(p) = t$ and $\operatorname{lm}(q) = u$, we can build a polynomial in the ideal of p and q that would raise the leading terms to the smallest level where they would cancel by computing

$$S \quad = \quad \operatorname{lc}(q) \cdot \frac{\operatorname{lcm}(t, u)}{t} \cdot p \quad - \quad \operatorname{lc}(p) \cdot \frac{\operatorname{lcm}(t, u)}{u} \cdot q.$$

Moreover, for all other monomials τ, μ and $a, b \in \mathbb{F}$, if $a\tau p - b\mu q$ cancels the leading terms of τp and μq, then it is a multiple of S.

Proof. First we show that the leading monomials of the two polynomials in the subtraction cancel. By Proposition 11.19,

$$\begin{aligned} \operatorname{lm}\left(\frac{\operatorname{lcm}(t, u)}{t} \cdot p \right) &= \frac{\operatorname{lcm}(t, u)}{t} \cdot \operatorname{lm}(p) \\ &= \frac{\operatorname{lcm}(t, u)}{t} \cdot t = \operatorname{lcm}(t, u); \end{aligned}$$

likewise

$$\begin{aligned} \operatorname{lm}\left(\frac{\operatorname{lcm}(t, u)}{u} \cdot q \right) &= \frac{\operatorname{lcm}(t, u)}{u} \cdot \operatorname{lm}(q) \\ &= \frac{\operatorname{lcm}(t, u)}{u} \cdot u = \operatorname{lcm}(t, u). \end{aligned}$$

Thus

$$\operatorname{lc}\left(\operatorname{lc}(q) \cdot \frac{\operatorname{lcm}(t, u)}{t} \cdot p \right) = \operatorname{lc}(q) \cdot \operatorname{lc}(p)$$

and

$$\operatorname{lc}\left(\operatorname{lc}(p) \cdot \frac{\operatorname{lcm}(t, u)}{t} \cdot q \right) = \operatorname{lc}(p) \cdot \operatorname{lc}(q).$$

Hence the leading monomials of the two polynomials in S cancel.

Let τ, μ be monomials over $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $a, b \in \mathbb{F}$ such that the leading monomials of the two polynomials in $a\tau p - b\mu q$ cancel. Let $\tau = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and $\mu = x_1^{\beta_1} \cdots x_n^{\beta_n}$ for appropriate α_i and β_i in $\mathbb{N}$. Write $\operatorname{lm}(p) = x_1^{\zeta_1} \cdots x_n^{\zeta_n}$ and $\operatorname{lm}(q) = x_1^{\omega_1} \cdots x_n^{\omega_n}$ for appropriate ζ_i and ω_i in $\mathbb{N}$. The leading monomials of $a\tau p - b\mu q$ cancel, so for each $i = 1, 2, \ldots, n$

$$\alpha_i + \zeta_i = \beta_i + \omega_i.$$

We have

$$\alpha_i = \beta_i + (\omega_i - \zeta_i).$$

Rewrite this as

$$\begin{aligned} \alpha_i - (\max(\zeta_i, \omega_i) - \zeta_i) &= [(\beta_i + (\omega_i - \zeta_i)) - (\max(\zeta_i, \omega_i) - \zeta_i)] \\ &= \beta_i - (\max(\zeta_i, \omega_i) - \omega_i). \end{aligned}$$

Let $\eta_i = \alpha_i - (\max(\zeta_i, \omega_i) - \zeta_i)$ and let

$$v = \prod_{i=1}^{n} x_i^{\eta_i}.$$

Then

$$a\tau p - b\mu q = v\left(a \cdot \frac{\operatorname{lcm}(t,u)}{t} \cdot p - b \cdot \frac{\operatorname{lcm}(t,u)}{u} \cdot q\right),$$

as claimed. □

The subtraction polynomial of Lemma 11.47 is important enough that we give it a special name.

Definition 11.48. Let $p, q \in \mathbb{F}[x_1, x_2, \ldots, x_n]$. We define the ***S*-polynomial of p and q** to be

$$\begin{aligned} \operatorname{Spol}(p,q) = \operatorname{lc}(q) \cdot \frac{\operatorname{lcm}(\operatorname{lm}(p), \operatorname{lm}(q))}{\operatorname{lm}(p)} \cdot p \\ - \operatorname{lc}(p) \cdot \frac{\operatorname{lcm}(\operatorname{lm}(p), \operatorname{lm}(q))}{\operatorname{lm}(q)} \cdot q. \end{aligned}$$

Hopefully, you see that *S*-poly-nomials generalize the cancellation of Gaussian elimination in a natural way.

For some *S*-polynomials, only one of the leading terms needs to change. This merits its own terminology.

Definition 11.49. Let $p, q \in \mathbb{F}[x_1, x_2, \ldots, x_n]$. If $\operatorname{lm}(p)$ divides $\operatorname{lm}(q)$, then we say that p **top-reduces** q.

If p top-reduces q, let $r = \operatorname{Spol}(p, q)$. We say that p **top-reduces** q **to** r.

Finally, let $F = (f_1, f_2, \ldots, f_m)$ be a list of polynomials in $\mathbb{F}[x_1, x_2, \ldots, x_n]$, and $r_1, r_2, \ldots, r_k \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ such that

- some polynomial of F top-reduces p to r_1,
- some polynomial of F top-reduces r_1 to r_2,
- $\ldots$
- some polynomial of F top-reduces r_{k-1} to r_k.

In this case, we say that p **top-reduces to** r_k **with respect to** F.

Example 11.50. Let $p = x + 1$ and $q = x^2 + 1$. We have $\operatorname{lm}(p) = x$ and $\operatorname{lm}(q) = x^2$. Since $\operatorname{lm}(p)$ divides $\operatorname{lm}(q)$, p top-reduces q. Their S-polynomial is

$$r = q - x \cdot p = -x + 1,$$

so q top-reduces to r with respect to $\{p\}$.

We need the following properties of polynomial operations.

Proposition 11.51. Let $p, q, r \in \mathbb{F}[x_1, x_2, \ldots, x_n]$. Each of the following holds:

(A) $\operatorname{lm}(pq) = \operatorname{lm}(p) \cdot \operatorname{lm}(q)$

(B) $\operatorname{lm}(p \pm q) \leq \max(\operatorname{lm}(p), \operatorname{lm}(q))$

(C) $\operatorname{lm}(\operatorname{Spol}(p, q)) < \operatorname{lcm}(\operatorname{lm}(p), \operatorname{lm}(q))$

(D) If p top-reduces q to r, then $\operatorname{lm}(r) < \operatorname{lm}(q)$.

Proof. For convenience, write $t = \operatorname{lm}(p)$ and $u = \operatorname{lm}(q)$.

(A) Any monomial of pq can be written as the product of two monomials vw, where v is a monomial of p and w is a monomial of q. If $v \neq \operatorname{lm}(p)$, then the definition of a leading monomial implies that $v < t$. Proposition 11.19 implies that

$$vw \leq tw,$$

with equality only if $v = t$. The same reasoning implies that

$$vw \leq tw \leq tu,$$

with equality only if $w = u$. Hence

$$\operatorname{lm}(pq) = tu = \operatorname{lm}(p) \operatorname{lm}(q).$$

(B) Any monomial of $p \pm q$ is a monomial of p or of q. Hence $\operatorname{lm}(p \pm q)$ is a monomial of p or of q. The maximum of these is $\max(\operatorname{lm}(p), \operatorname{lm}(q))$. Hence $\operatorname{lm}(p \pm q) \leq \max(\operatorname{lm}(p), \operatorname{lm}(q))$.

(C) Definition 11.48 and (B) imply $\operatorname{lm}(\operatorname{Spol}(p, q)) < \operatorname{lcm}(\operatorname{lm}(p), \operatorname{lm}(q))$.

(D) By definition, top-reduction is a kind of S-polynomial, so this follows from (C). □

In a triangular linear system, we achieve a triangular form by rewriting all polynomials that share a leading variable. In the *linear* case we can accomplish this using *scalar multiplication*, requiring nothing else. In the non-linear case, we need to check for divisibility of monomials. The following result should, therefore, not surprise you very much.

Theorem 11.52 (Buchberger's characterization)**.** Let $G = \{g_1, g_2, \ldots, g_m\} \subsetneq \mathbb{F}[x_1, x_2, \ldots, x_n]$. It is a Gröbner basis of the ideal $I = \langle g_1, g_2, \ldots, g_m \rangle$ if and only if $\operatorname{Spol}\left(g_i, g_j\right)$ top-reduces to zero with respect to G for each pair i, j with $1 \leq i < j \leq m$.

Example 11.53. Recall two systems considered at the beginning of this chapter,

$$F = \left(x^2 + y^2 - 4, xy - 1\right)$$

and

$$G = \left(x^2 + y^2 - 4, xy - 1, x + y^3 - 4y, -y^4 + 4y^2 - 1\right).$$

Is either of these a Gröbner basis?

- We already showed that F is not, as its one S-polynomial is

$$\begin{aligned} S &= \operatorname{Spol}(f_1, f_2) \\ &= y\left(x^2 + y^2 - 4\right) - x(xy - 1) \\ &= x + y^3 - 4y, \end{aligned}$$

  and $\operatorname{lm}(S) = x$, which neither leading term of F divides.
- On the other hand, G is a Gröbner basis. We will not show all six S-polynomials (you will verify this in Exercise 11.56), but

$$\operatorname{Spol}(g_1, g_2) - g_3 = 0,$$

  so the problem with F does not reappear. It is worth noting that

$$\operatorname{Spol}(g_1, g_4) - \left(4y^2 - 1\right) g_1 + \left(y^2 - 4\right) g_4 = 0.$$

  If we rewrite $\operatorname{Spol}(g_1, g_4) = y^4 g_1 + x^2 g_4$ and substitute it into the above equation, something very interesting turns up:

$$\begin{aligned} \left(y^4 g_1 + x^2 g_4\right) - \left(4y^2 - 1\right) g_1 + \left(y^2 - 4\right) g_4 &= 0 \\ -\left(-y^4 + 4y^2 - 1\right) g_1 + \left(x^2 + y^2 - 4\right) g_4 &= 0 \\ -g_4 g_1 + g_1 g_4 &= 0. \end{aligned}$$

Remark 11.54. Buchberger's characterization suggests a method to compute a Gröbner basis of an ideal: given a basis, use S-polynomials to find elements of the ideal that do not satisfy Definition 11.46. Add these to the basis, repeating until all of them reduce to zero.

This approach has two wrinkles we have to iron out:

- We don't know that a Gröbner basis exists for every ideal. For all we know, there may be ideals for which no Gröbner basis exists.
- We don't know that the proposed method will even terminate! It could be that we can go on forever, adding new polynomials to the ideal without ever stopping.

We resolve these questions in the following section.

It remains to prove Theorem 11.52, but before we can do that we will need the following useful lemma. While small, it has important repercussions later.

Lemma 11.55. Let $p, f_1, f_2, \ldots, f_m \in \mathbb{F}[x_1, x_2, \ldots, x_n]$. Let $F = (f_1, f_2, \ldots, f_m)$. If p top-reduces to zero with respect to F, then there exist $q_1, q_2, \ldots, q_m \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ such that each of the following holds:
(A) $p = q_1 f_1 + q_2 f_2 + \cdots + q_m f_m$; and
(B) for each $k = 1, 2, \ldots, m$, $q_k = 0$ or $\operatorname{lm}(q_k)\operatorname{lm}(g_k) \leq \operatorname{lm}(p)$.

Proof. You do it! See Exercise 11.62. □

You will see in the following that Lemma 11.55 allows us to replace polynomials that are "too large" with smaller polynomials. This allows us to obtain the desired form.

Proof of Theorem 11.52. Assume first that G is a Gröbner basis, and let i, j be such that $1 \leq i < j \leq m$. Then

$$\operatorname{Spol}\left(g_i, g_j\right) \in \left\langle g_i, g_j \right\rangle \subset \langle g_1, g_2, \ldots, g_m \rangle,$$

and the definition of a Gröbner basis implies that there exists $k_1 \in \{1, 2, \ldots, m\}$ such that g_{k_1} top-reduces $\operatorname{Spol}\left(g_i, g_j\right)$ to a new polynomial, say r_1. The definition further implies that if r_1 is not zero, then there exists $k_2 \in \{1, 2, \ldots, m\}$ such that g_{k_2} top-reduces r_1 to a new polynomial, say r_2. Repeating this iteratively, we obtain a chain of polynomials $r_1, r_2, \ldots$ such that r_ℓ top-reduces to $r_{\ell+1}$ for each $\ell \in \mathbb{N}$. From Proposition 11.51, we see that

$$\operatorname{lm}(r_1) > \operatorname{lm}(r_2) > \cdots.$$

Recall that the monomials are well-ordered under an admissible ordering, so any set of monomials has a least element, including the set $R = \{\operatorname{lm}(r_1), \operatorname{lm}(r_2), \ldots\}$. Thus the chain of top-reductions cannot continue indefinitely. It cannot conclude with a non-zero polynomial r_{last}, since:

- top-reduction keeps each r_ℓ in the ideal:
  - subtraction by the subring property, and
  - multiplication by the absorption property; hence
- by the definition of a Gröbner basis, a non-zero r_{last} would be top-reducible by some element of G.

The chain of top-reductions must conclude with zero, so $\operatorname{Spol}\left(g_i, g_j\right)$ top-reduces to zero.

Now assume every S-polynomial top-reduces to zero modulo G. We want to show any element of I is top-reducible by an element of G. So let $p \in I$; by definition, there exist polynomials $h_1, \ldots, h_m \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ such that

$$p = h_1 g_1 + \cdots + h_m g_m.$$

For each i, write $t_i = \operatorname{lm}(g_i)$ and $u_i = \operatorname{lm}(h_i)$. Let $T = \max_{i=1,2,\ldots,m}(u_i t_i)$. We call T the **maximal term of the representation** $h_1, h_2, \ldots, h_m$. If $\operatorname{lm}(p) = T$, then we are done, since

$$\operatorname{lm}(p) = T = u_k t_k = \operatorname{lm}(h_k)\operatorname{lm}(g_k) \quad \exists k \in \{1,2,\ldots,m\}.$$

Otherwise, there must be some cancellation among the leading monomials of each polynomial in the sum on the right hand side. That is,

$$T = \operatorname{lm}\left(h_{\ell_1} g_{\ell_1}\right) = \operatorname{lm}\left(h_{\ell_2} g_{\ell_2}\right) = \cdots = \operatorname{lm}\left(h_{\ell_s} g_{\ell_s}\right)$$

for some $\ell_1, \ell_2, \ldots, \ell_s \in \{1,2,\ldots,m\}$. From Lemma 11.47, we know that we can write the sum of these leading terms as a sum of multiples of a S-polynomials of G. That is,

$$\begin{aligned}\operatorname{lc}\left(h_{\ell_1}\right)\operatorname{lm}\left(h_{\ell_1}\right) g_{\ell_1} + \cdots + \operatorname{lc}\left(h_{\ell_s}\right)\operatorname{lm}\left(h_{\ell_s}\right) g_{\ell_s} = \\ = \sum_{1 \le a < b \le s} c_{a,b} u_{a,b} \operatorname{Spol}\left(g_{\ell_a}, g_{\ell_b}\right)\end{aligned}$$

where for each a, b we have $c_{a,b} \in \mathbb{F}$ and $u_{a,b} \in \mathbb{M}$. Let

$$S = \sum_{1 \le a < b \le s} c_{a,b} u_{a,b} \operatorname{Spol}\left(g_{\ell_a}, g_{\ell_b}\right).$$

Observe that

$$\left[\operatorname{lm}\left(h_{\ell_1}\right) g_{\ell_1} + \operatorname{lm}\left(h_{\ell_2}\right) g_{\ell_2} + \cdots + \operatorname{lm}\left(h_{\ell_s}\right) g_{\ell_s}\right] - S = 0. \tag{37}$$

By hypothesis, each S-polynomial of S top-reduces to zero. This fact, Lemma 11.55 and Proposition 11.51, implies that for each a, b we can find $q_\lambda^{(a,b)} \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ such that

$$\operatorname{Spol}\left(g_{\ell_a}, g_{\ell_b}\right) = q_1^{(a,b)} g_1 + \cdots + g_m^{(a,b)} g_m$$

and for each $\lambda = 1, 2, \ldots, m$ we have $q_\lambda^{(a,b)} = 0$ or

$$\begin{aligned}\operatorname{lm}\left(q_\lambda^{(a,b)}\right)\operatorname{lm}(g_\lambda) &\le \operatorname{lm}\left(\operatorname{Spol}\left(g_{\ell_a}, g_{\ell_b}\right)\right) \\ &< \operatorname{lcm}\left(\operatorname{lm}\left(g_{\ell_a}\right), \operatorname{lm}\left(g_{\ell_b}\right)\right).\end{aligned} \tag{38}$$

Let $Q_1, Q_2, \ldots, Q_m \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ such that

$$Q_k = \begin{cases} \sum_{1 \le a < b \le s} c_{a,b} u_{a,b} q_k^{(a,b)}, & k \in \{\ell_1, \ldots, \ell_s\}; \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$S = Q_1 g_1 + Q_2 g_2 + \cdots + Q_m g_m.$$

In other words,

$$S - (Q_1 g_1 + Q_2 g_2 + \cdots + Q_m g_m) = 0.$$

By equation (38) and Proposition 11.51, for each $k = 1, 2, \ldots, m$ we have $Q_k = 0$ or

$$\begin{aligned}
\operatorname{lm}(Q_k)\operatorname{lm}(g_k) &\leq \max_{1\leq a<b\leq s}\left\{\left[u_{a,b}\operatorname{lm}\left(q_k^{(a,b)}\right)\right]\operatorname{lm}(g_k)\right\} \\
&= \max_{1\leq a<b\leq s}\left\{u_{a,b}\left[\operatorname{lm}\left(q_k^{(a,b)}\right)\operatorname{lm}(g_k)\right]\right\} \\
&\leq \max_{1\leq a<b\leq s}\left\{u_{a,b}\operatorname{lm}\left(\operatorname{Spol}\left(g_{\ell_a}, g_{\ell_b}\right)\right)\right\} \\
&< u_{a,b}\operatorname{lcm}\left(\operatorname{lm}\left(g_{\ell_a}\right), \operatorname{lm}\left(g_{\ell_b}\right)\right) \\
&= T. \qquad (39)
\end{aligned}$$

By substitution,

$$\begin{aligned}
p &= (h_1 g_1 + h_2 g_2 + \cdots + h_m g_m) - \left(S - \sum_{k\in\{\ell_1,\ldots,\ell_s\}} Q_k g_k\right) \\
&= \left[\sum_{k\notin\{\ell_1,\ldots,\ell_s\}} h_k g_k + \sum_{k\in\{\ell_1,\ldots,\ell_s\}} \left(h_k - \operatorname{lc}(h_k)\operatorname{lm}(h_k)\right) g_k\right] \\
&\quad + \cancelto{0}{\left[\sum_{k\in\{\ell_1,\ldots,\ell_s\}} \operatorname{lc}(h_k)\operatorname{lm}(h_k)\, g_k - S\right]} \\
&\quad + \sum_{k\in\{\ell_1,\ldots,\ell_s\}} Q_k g_k.
\end{aligned}$$

Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_m \in \mathbb{F}[x_1, \ldots, x_n]$ such that

$$\mathcal{Q}_k(x) = \begin{cases} h_k, & k \notin \{\ell_1, \ldots, \ell_s\}; \\ h_k - \operatorname{lc}(h_k)\operatorname{lm}(h_k) + Q_k, & \text{otherwise.} \end{cases}$$

By substitution,

$$p = \mathcal{Q}_1 g_1 + \cdots + \mathcal{Q}_m g_m.$$

If $k \notin \{\ell_1, \ldots, \ell_s\}$, then the choice of T as the maximal term of the representation implies that

$$\operatorname{lm}(\mathcal{Q}_k)\operatorname{lm}(g_k) = \operatorname{lm}(h_k)\operatorname{lm}(g_k) < T.$$

Otherwise, Proposition 11.51 and equation (39) imply that

$$\begin{aligned}
\operatorname{lm}(\mathcal{Q}_k)\operatorname{lm}(g_k) &\leq \max\left(\left(\operatorname{lm}\left(h_k - \operatorname{lc}(h_k)\operatorname{lm}(h_k)\right), \operatorname{lm}(Q_k)\right)\operatorname{lm}(g_k)\right) \\
&< \operatorname{lm}(h_k)\operatorname{lm}(g_k) \\
&= T.
\end{aligned}$$

What have we done? We have rewritten the original representation of p over the ideal, which had maximal term T, with another representation, which has maximal term smaller than T.

This was possible because all the S-polynomials reduced to zero; S-polynomials appeared because $T > \mathrm{lm}(p)$, implying cancellation in the representation of p over the ideal. We can repeat this as long as $T > \mathrm{lm}(p)$, generating a list of monomials

$$T_1 > T_2 > \cdots .$$

The well-ordering of $\mathbb{M}$ implies that this cannot continue indefinitely! Hence there must be a representation

$$p = H_1 g_1 + \cdots + H_m g_m$$

such that for each $k = 1, 2, \ldots, m$ $H_k = 0$ or $\mathrm{lm}(H_k)\,\mathrm{lm}(g_k) \leq \mathrm{lm}(p)$. Both sides of the equation must simplify to the same polynomial, with the same leading variable, so at least one k has $\mathrm{lm}(H_k)\,\mathrm{lm}(g_k) = \mathrm{lm}(p)$; that is, $\mathrm{lm}(g_k) \mid \mathrm{lm}(p)$. Since p was arbitrary, G satisfies the definition of a Gröbner basis. □

Exercises.

Exercise 11.56. Show that

$$G \quad = \quad \left(\quad x^2 + y^2 - 4, \quad xy - 1, \quad x + y^3 - 4y, \quad -y^4 + 4y^2 - 1 \right)$$

is a Gröbner basis with respect to the lexicographic ordering.

Exercise 11.57. Show that G of Exercise 11.56 is *not* a Gröbner basis with respect to the grevlex ordering. The Gröbner basis property depends on the choice of term ordering!

Exercise 11.58. Show that any Gröbner basis G of an ideal I is a basis of the same ideal; that is, any $p \in I$ can be written as $p = \sum_{i=1}^{m} h_i g_i$ for appropriate $h_i \in \mathbb{F}[x_1, \ldots, x_n]$.

Exercise 11.59. Show that for any non-constant polynomial f, $F = (f, f+1)$ is not a Gröbner basis.

Exercise 11.60. Show that every list of monomials is a Gröbner basis.

Exercise 11.61. We call a basis G of an ideal a **minimal basis** if no monomial of any $g_1 \in G$ is divisible by the leading monomial of any $g_2 \in G$.

(a) Suppose that a Gröbner basis G is not minimal. Show that we obtain a minimal basis by repeatedly replacing each $g \in G$ by $g - t\,g'$ where $t\,\mathrm{lm}(g')$ is a monomial of g.

(b) Explain why the minimal basis obtained in part (a) is also a Gröbner basis of the same ideal.

Exercise 11.62. Let

$$p = 4x^4 - 3x^3 - 3x^2y^4 + 4x^2y^2 - 16x^2 + 3xy^3 - 3xy^2 + 12x$$

and $F = \left(x^2 + y^2 - 4, xy - 1\right)$.

(a) Show that p reduces to zero with respect to F.

(b) Show that there exist $q_1, q_2 \in \mathbb{F}[x, y]$ such that $p = q_1 f_1 + q_2 f_2$.

(c) Generalize the argument of (b) to prove Lemma 11.55.

Exercise 11.63. For G to be a Gröbner basis, Definition 11.46 requires that every polynomial in the ideal generated by G be top-reducible by some element of G. If polynomials in the basis are top-reducible by other polynomials in the basis, we call them **redundant elements of the basis**.

(a) The Gröbner basis of Exercise 11.56 has redundant elements. Find a subset G_{min} of G that contains no redundant elements, but is still a Gröbner basis.

(b) Describe the method you used to find G_{min}.

(c) Explain why redundant polynomials are not required to satisfy Definition 11.46. That is, if we know that G is a Gröbner basis, then we could remove redundant elements to obtain a smaller list, G_{min}, which is also a Gröbner basis *of the same ideal*.

11.5: Buchberger's algorithm

Algorithm 7 on page 337 shows how to triangularize a linear system. Essentially, it looks for parts of the system that are not triangular (equations with the same leading variable) then adds a new polynomial (an S-polynomial!) to move it closer to the triangular form. The new polynomial replaces one of the older polynomials in the pair.

For non-linear systems, we will try an approach that is similar, not but identical. We *will* look for polynomials in the ideal that do not satisfy the Gröbner basis property, we *will* add a new polynomial to repair this defect. We will not, however, replace the older polynomials, because we may still need their leading monomials, and the S-polynomial may have a very different one. Worse, removing this polynomial could even change the ideal!

Example 11.64. Let $F = \left(xy + xz + z^2, yz + z^2\right)$, and use grevlex with $x > y > z$. The S-polynomial of f_1 and f_2 is

$$S = z\left(xy + xz + z^2\right) - x\left(yz + z^2\right) = z^3.$$

Let $G = \left(xy + xz + z^2, z^3\right)$; that is, G is F with f_2 replaced by S. It turns out that $yz + z^2 \notin \langle G \rangle$. If it were, then

$$yz + z^2 = h_1\left(xy + xz + z^2\right) + h_2 \cdot z^3.$$

Every term of the right hand side will be divisible either by x or by z^2, but yz is divisible by neither. Hence $yz + z^2 \in \langle G \rangle$.

We will have to adapt Algorithm 7 without replacing or discarding any polynomials. With non-linear polynomials, Buchberger's characterization (Theorem 11.52) suggests that we compute the S-polynomials, and top-reduce them. If they all top-reduce to zero, then Buchberger's characterization implies that we have a Gröbner basis already, so there is nothing to do. Otherwise, at least one S-polynomial does *not* top-reduce to zero, so we add its reduced form to the basis and test the new S-polynomials as well. This suggests Algorithm 8.

Theorem 11.65. For any list of polynomials F over a field, Buchberger's algorithm terminates with a Gröbner basis of $\langle F \rangle$.

Correctness isn't hard *if* Buchberger's algorithm terminates, because it discards nothing, adds only polynomials that are already in $\langle F \rangle$, and terminates only if all the S-polynomials of G top-reduce to zero. The problem is termination, which relies on the Ascending Chain Condition.

Algorithm 8. Buchberger's algorithm to compute a Gröbner basis

```
1:  inputs
2:    F = (f_1, f_2, ..., f_m), where each f_i ∈ 𝔽[x_1, ..., x_n].
3:    <, an admissible ordering.
4:  outputs
5:    G, a Gröbner basis of ⟨F⟩ with respect to <.
6:  do
7:    Let G := F
8:    Let P = {(f, g) : ∀ f, g ∈ G such that f ≠ g}
9:    repeat while P ≠ ∅
10:     Choose (f, g) ∈ P
11:     Remove (f, g) from P
12:     Let S be the S-polynomial of f, g
13:     Let r be the top-reduction of S with respect to G
14:     if r ≠ 0
15:       Replace P by P ∪ {(h, r) : h ∈ G}
16:       Append r to G
17:   return G
```

Proof. For *termination*, let $\mathbb{F}$ be a field, and F a list of polynomials over $\mathbb{F}$. Designate

$$\begin{aligned}
I_0 &= \langle \operatorname{lm}(g_1), \operatorname{lm}(g_2), \ldots, \operatorname{lm}(g_m) \rangle \\
I_1 &= \langle \operatorname{lm}(g_1), \operatorname{lm}(g_2), \ldots, \operatorname{lm}(g_m), \operatorname{lm}(g_{m+1}) \rangle \\
I_2 &= \langle \operatorname{lm}(g_1), \ldots, \operatorname{lm}(g_m), \operatorname{lm}(g_{m+1}), \operatorname{lm}(g_{m+2}) \rangle \\
&\vdots \\
I_i &= \langle \operatorname{lm}(g_1), \ldots, \operatorname{lm}(g_{m+i}) \rangle
\end{aligned}$$

where g_{m+i} is the ith polynomial added to G by line 16 of Algorithm 8.

We claim that $I_0 \subseteq I_1 \subseteq I_2 \subseteq \cdots$ is a strictly ascending chain of ideals. After all, a polynomial r is added to the basis only when it is non-zero (line 14); since it has not top-reduced to zero, $\operatorname{lm}(r)$ is not top-reducible by

$$G_{i-1} = (g_1, g_2, \ldots, g_{m+i-1}).$$

Thus for any $p \in G_{i-1}$, $\operatorname{lm}(p)$ does not divide $\operatorname{lm}(r)$. We further claim that this implies that $\operatorname{lm}(r) \notin I_{i-1}$. By way of contradiction, suppose that it is. By Exercise 11.60 on page 360, any list of monomials is a Gröbner basis; hence

$$T = (\operatorname{lm}(g_1), \operatorname{lm}(g_2), \ldots, \operatorname{lm}(g_{m+i-1}))$$

is a Gröbner basis, and by Definition 11.46 every polynomial in I_{i-1} is top-reducible by T. Since r is not top-reducible by T, $\operatorname{lm}(r) \notin I_{i-1}$.

Thus $I_{i-1} \subsetneq I_i$, and $I_0 \subseteq I_1 \subseteq I_2 \subseteq \cdots$ is a strictly ascending chain of ideals in $\mathbb{F}[x_1 x_2, \ldots, x_n]$. By Proposition 8.33 and Definition 8.31, there exists $M \in \mathbb{N}$ such that $I_M = I_{M+1} = \cdots$. This

implies that the algorithm can add at most $M - m$ polynomials to G; after having done so, any remaining elements of P generate S-polynomials that top-reduce to zero! Line 11 removes each pair (i, j) from P, so P decreases after we have added these $M - m$ polynomials. Eventually P decreases to $\emptyset$, and the algorithm terminates.

For *correctness,* we have to show two things: first, that G is a basis of the same ideal as F, and second, that G satisfies the Gröbner basis property. For the first, observe that every polynomial added to G is by construction an element of $\langle F \rangle$, and we removed no elements from the basis, so the ideal does not change. For the second, observe that the very construction of G ensures that Buchberger's characterization of a Gröbner basis is satisfied. □

Exercises

Exercise 11.66. Using G of Exercise 11.56, compute a Gröbner basis with respect to the grevlex ordering.

Exercise 11.67. Following up on Exercises 11.57 and 11.66, a simple diagram will help show that it is usually "faster" to compute a Gröbner basis in any total degree ordering than it is in the lexicographic ordering. We can diagram the monomials in x and y on the x-y plane by plotting $x^\alpha y^\beta$ at the point (α, β).

(a) Shade the region of monomials that are smaller than x^2y^3 with respect to the lexicographic ordering.

(b) Shade the region of monomials that are smaller than x^2y^3 with respect to the graded reverse lexicographic ordering.

(c) Explain why the diagram implies that top-reduction of a polynomial with leading monomial x^2y^3 will *probably* take less effort in grevlex than in the lexicographic ordering.

Exercise 11.68. However, it is not *always* faster to use the grevlex ordering. To see this, consider the system

$$\begin{aligned} C_4 \quad = \quad (\quad & x_1 + x_2 + x_3 + x_4, \\ & x_1x_2 + x_2x_3 + x_3x_4 + x_4x_1, \\ & x_1x_2x_3 + x_2x_3x_4 + x_3x_4x_1 + x_4x_1x_2, \\ & x_1x_2x_3x_4 - 1 \quad). \end{aligned}$$

Compute the size of the Gröbner basis of C_4 over the field $\mathbb{Z}_2$ with respect to grevlex ordering, then with respect to lex ordering.

Exercise 11.69. Let $g_1, g_2, \ldots, g_m \in \mathbb{F}[x_1, x_2, \ldots, x_n]$. We say that a non-linear polynomial is *homogeneous* if every term is of the same total degree. For example, $xy - 1$ is not homogeneous, but $xy - h^2$ is. As you may have guessed, we can homogenize any polynomial by multiplying every term by an appropriate power of a *homogenizing variable* h. When $h = 1$, we have the original polynomial.

(a) Homogenize the following polynomials.
   (i) $x^2 + y^2 - 4$
   (ii) $x^3 - y^5 + 1$

(iii) $xz + z^3 - 4x^5y - xyz^2 + 3x$

(b) Explain the relationship between solutions to a system of nonlinear polynomials G and solutions to the system of homogenized polynomials H.

(c) With homogenized polynomials, we usually use a variant of the lexicographic ordering. Although h comes *first* in the dictionary, we pretend that it comes last. So $x > yh^2$ and $y > h^{10}$. Use this modified lexicographic ordering to determine the leading monomials of your solutions for part (a).

(d) Does homogenization preserve leading monomials?

Exercise 11.70. Assume that the $g_1, g_2, \ldots, g_m$ are homogeneous; in this case, we can build the *ordered Macaulay matrix of G of degree D* in the following way.

- Each row of the matrix represents a monomial multiple of some g_i. If g_i is of degree $d \leq D$, then we compute all the monomial multiples of g_i that have degree D.
- Each column represents a monomial. Column 1 corresponds to the largest monomial with respect to the lexicographic ordering; column 2 corresponds to the next-largest polynomial; etc.
- Each entry of the matrix is the coefficient of a monomial for a monomial multiple of some g_i.

(a) The homogenization of the circle and the hyperbola gives us the system

$$F = \left(x^2 + y^2 - 4h^2, xy - h^2\right).$$

Verify that its ordered Macaulay matrix of degree 3 is

$$\begin{pmatrix}
x^3 & x^2y & xy^2 & y^3 & x^2h & xyh & y^2h & xh^2 & yh^2 & h^3 & \\
1 & & 1 & & & & & -4 & & & xf_1 \\
& 1 & & 1 & & & & & -4 & & yf_1 \\
& & & & 1 & & 1 & & & -4 & hf_1 \\
& 1 & & & & & & -1 & & & xf_2 \\
& & 1 & & & & & & -1 & & yf_2 \\
& & & & & 1 & & & & -1 & hf_2
\end{pmatrix}.$$

Show that if you triangularize this matrix *without swapping columns*, the row corresponding to xf_2 now contains coefficients that correspond to the homogenization of $x + y^3 - 4y$.

(b) Compute the ordered Macaulay matrix of F of degree 4, then triangularize it. Be sure *not to swap columns*, nor to destroy rows that provide new information. Show that

- the entries of at least one row correspond to the coefficients of a multiple of the homogenization of $x + y^3 - 4y$, and
- the entries of at least one other row are the coefficients of the homogenization of $\pm\left(y^4 - 4y^2 + 1\right)$.

(c) Explain the relationship between triangularizing the ordered Macaulay matrix and Buchberger's algorithm.

Sage programs

The following programs can be used in Sage to help make the amount of computation involved in the exercises less burdensome. Use

- `M, mons = sylvester_matrix(F,d)` to make an ordered Macaulay matrix of degree d for the list of polynomials F,
- `N = triangularize_matrix(M)` to triangularize M in a way that respects the monomial order, and
- `extract_polys(N,mons)` to obtain the polynomials of N.

```
def make_monomials(xvars,d,p=0,order="lex"):
  result = set([1])
  for each in range(d):
    new_result = set()
    for each in result:
      for x in xvars:
        new_result.add(each*x)
    result = new_result
  result = list(result)
  result.sort(lambda t,u:  monomial_cmp(t,u))
  n = sage.rings.integer.Integer(len(xvars))
  return result

def monomial_cmp(t,u):
  xvars = t.parent().gens()
  for x in xvars:
    if t.degree(x) != u.degree(x):
      return u.degree(x) - t.degree(x)
  return 0

def homogenize_all(polys):
  for i in range(len(polys)):
    if not polys[i].is_homogeneous():
      polys[i] = polys[i].homogenize()

def sylvester_matrix(polys,D,order="lex"):
  L = [ ]
  homogenize_all(polys)
  xvars = polys[0].parent().gens()
  for p in polys:
    d = D - p.degree()
    R = polys[0].parent()
    mons = make_monomials(R.gens(),d,order=order)
    for t in mons:
      L.append(t*p)
  mons = make_monomials(R.gens(),D,order=order)
```

```
  mons_dict = {}
  for each in range(len(mons)):
    mons_dict.update({mons[each]:each})
  M = matrix(len(L),len(mons))
  for i in range(len(L)):
    p = L[i]
    pmons = p.monomials()
    pcoeffs = p.coefficients()
    for j in range(len(pmons)):
      M[i,mons_dict[pmons[j]]] = pcoeffs[j]
  return M, mons

def triangularize_matrix(M):
  N = M.copy()
  m = N.nrows()
  n = N.ncols()
  for i in range(m):
    pivot = 0
    while pivot < n and N[i,pivot] == 0:
      pivot = pivot + 1
    if pivot < n:
      a = N[i,pivot]
      for j in range(i+1,m):
        if N[j,pivot] != 0:
          b = N[j,pivot]
          for k in range(pivot,n):
            N[j,k] = a * N[j,k] - b * N [i,k]
  return N

def extract_polys(M, mons):
  L = [ ]
  for i in range(M.nrows()):
    p = 0 for j in range(M.ncols()):
    if M[i,j] != 0:
      p = p + M[i,j]*mons[j]
    L.append(p)
  return L
```

11.6: Nullstellensatz

The German word *Nullstellensatz* means "Theorem (*satz*) on the locations (*stellen*) of zero (*null*)." There are two different theorems; a *weak* Nullstellensatz, and a "*not-so-weak*" Nullstellensatz. In this section, we consider only the weak version. Throughout this section,

- $\mathbb{F}$ is an *algebraically closed* field—that is, all nonconstant polynomials over $\mathbb{F}$ have all their roots in $\mathbb{F}$;
- $\mathcal{R} = \mathbb{F}[x_1, x_2, \ldots, x_n]$ is a polynomial ring;
- $F \subseteq \mathcal{R}$;
- $V_F \subseteq \mathbb{F}^n$ is the set of common roots of elements of F;[19] and
- $I = \langle F \rangle$.

Note that $\mathbb{C}$ is algebraically closed, but $\mathbb{R}$ is not, since the roots of $x^2 + 1 \in \mathbb{R}[x]$ are not in $\mathbb{R}$.

An interesting and useful consequence of algebraic closure is the following.

Lemma 11.71. $\mathbb{F}$ is infinite.

Proof. Let $n \in \mathbb{N}^+$, and $a_1, \ldots, a_n \in \mathbb{F}$. Obviously, $f = (x - a_1) \cdots (x - a_n)$ satisfies $f(x) = 0$ for all $x = a_1, \ldots, a_n$. Let $g = f + 1$; then $g(x) \neq 0$ for all $x = a_1, \ldots, a_n$. Since $\mathbb{F}$ is closed, g has a root $b \in \mathbb{F} \setminus \{a_1, \ldots, a_n\}$. Thus, no finite list of elements enumerates $\mathbb{F}$, which means $\mathbb{F}$ must be infinite. □

Theorem 11.72 (Hilbert's Weak Nullstellensatz)**.** If $V_F = \emptyset$, then $I = \mathcal{R}$.

Proof. We proceed by induction on n, the number of variables.

Inductive base: Let $n = 1$. Recall that in this case, $\mathcal{R} = \mathbb{F}[x]$ is a Euclidean domain, and hence a principal ideal domain. Thus $I = \langle f \rangle$ for some $f \in \mathcal{R}$. If $V_F = \emptyset$, then f has no roots in $\mathbb{F}$. Theorem 10.18 tells us that every principal ideal domain is a unique factorization domain, so if f is non-constant, it has a unique factorization into irreducible polynomials. Theorem 10.42 tells us that any irreducible p extends $\mathcal{R}$ to a field $\mathbb{E} = R/\langle p \rangle$ containing both $\mathbb{F}$ and a root α of p. Since $\mathbb{F}$ is algebraically closed, $\alpha \in \mathbb{F}$ itself; that is, $\mathbb{E} = \mathbb{F}$. But then $x - \alpha \in \mathcal{R}$ is a factor of p, contradicting the assumption that p is irreducible. Since p was an arbitrary factor, f itself has no irreducible factors, which (since we are in a unique factorization domain) means that f is a nonzero constant; that is, $f \in \mathbb{F}$. By the inverse property of fields, $f^{-1} \in \mathbb{F} \subseteq \mathbb{F}[x]$, and absorption implies that $1 = f \cdot f^{-1} \in I$.

Inductive hypothesis: Let $k \in \mathbb{N}^+$, and suppose that in any polynomial ring over a closed field with $n = k$ variables, $V_F = \emptyset$ implies $I = \mathcal{R}$.

Inductive step: Let $n = k + 1$. Assume $V_F = \emptyset$. If F contains a constant polynomial, then we are done; thus, let $f \in F$. Let d be the maximum degree of a term of f. Rewrite f by substituting

$$\begin{aligned} x_1 &= y_1, \\ x_2 &= y_2 + a_2 y_1, \\ &\vdots \\ x_n &= y_n + a_n y_1, \end{aligned}$$

[19] The notation V_F comes from the term **variety** in algebraic geometry.

for some $a_1,\ldots,a_n \in \mathbb{F}$. (We make the choice of which $a_1,\ldots,a_n$ specific below.) This can be a little confusing, so let's take an example.

Example 11.73. Suppose $f = x_1 + x_2^2 x_3$. We rewrite f as

$$y_1 + (y_2 + a_2 y_1)^2 (y_3 + a_3 y_1)^3 = \\ y_1 + \left(y_2^2 + 2a_2 y_1 y_2 + a_2^2 y_1^2\right)\left(y_3^4 + 3a_3 y_1 y_3^3 + 3a_3^2 y_1^2 y_3^2 + a_3^3 y_1^3 y_3\right).$$

Take note of the forms within the parentheses.

Observe that if $i \neq 1$, then we rewrite x_i^d as $y_i^d + a_2 y_1 y_i^{d-1} \cdots + a_i^d y_1^d$, so if both $1 < i < j$ and $b + c = d$, then

$$\begin{aligned} x_i^b x_j^c &= \left(y_i^b + \cdots + a_i^b y_1^b\right)\left(y_j^c + \cdots + a_j^c y_1^c\right) \\ &= a_i^b a_j^c y_1^{b+c} + g\left(y_1, y_i, y_j\right) \\ &= a_i^b a_j^c y_1^d + g\left(y_1, y_i, y_j\right), \end{aligned}$$

where $\deg_{y_1} g < d$. Thus, we can collect terms containing y_1^d as

$$f = c y_1^d + g\left(y_1,\ldots,y_n\right)$$

where $c \in \mathbb{F}$ and $\deg_y g < d$. Since $\mathbb{F}$ is infinite, we can find $a_2,\ldots,a_n$ such that $c \neq 0$.

Let $\varphi : \mathcal{R} \longrightarrow \mathbb{F}[y_1,\ldots,y_n]$ by

$$\varphi\left(f\left(x_1,\ldots,x_n\right)\right) = f\left(y_1, y_2 + a_2 y_1,\ldots,y_n + a_n y_1\right);$$

that is, φ substitutes *every* element of $\mathcal{R}$ with the values that we obtained so that f_1 would have the special form above. This is a ring isomomorphism (Exercise 11.76), so $J = \varphi(I)$ is an ideal of $\mathbb{F}[y_1,\ldots,y_n]$. If $V_J \neq \emptyset$, then any $b \in V_J$ can be transformed into an element of V_F (see Exercise 11.77); hence $V_J = \emptyset$ as well.

Now let $\eta : \mathbb{F}[y_1,\ldots,y_n] \longrightarrow \mathbb{F}[y_2,\ldots,y_n]$ by $\eta(g) = g(0, y_2,\ldots,y_n)$.

Example 11.74. For instance, $\eta\left(x_1^3 + x_1 x_3^2 + x_2^2 x_3 + x_4\right) = x_2^2 x_3 + x_4$.

Again, $K = \eta(J)$ is an ideal, though the proof is different (Exercise 11.79). We claim that if $V_K \neq \emptyset$, then likewise $V_J \neq \emptyset$. To see why, let $h \in \eta(\mathbb{F}[y_1,\ldots,y_n])$, and suppose $b \in \mathbb{F}^{n-1}$ satisfies $h(b) = 0$. Let g be any element of $\mathbb{F}[y_1,\ldots,y_n]$ such that $\eta(g) = h$; then

$$g\left(0, b_1,\ldots,b_{n-1}\right) = h\left(b_1,\ldots,b_{n-1}\right) = 0,$$

so that we can prepend 0 to any element of V_K and obtain an element of V_J. Since $V_J = \emptyset$, this is impossible, so $V_K = \emptyset$.

Since $V_K = \emptyset$ and $K \subseteq \mathbb{F}[y_2,\ldots,y_n]$, the inductive hypothesis finally helps us see that $K = \mathbb{F}[y_2,\ldots,y_n]$. In other words, $1 \in K$. Since $K \subset J$ (see Exercise 11.79), $1 \in J$. Since $\varphi(f) \in \mathbb{F}$ if and only if $f \in \mathbb{F}$ (Exercise 11.78), there exists some $f \in \langle F \rangle$ such that $f \in \mathbb{F}$. □

Exercises

Exercise 11.75. Show that the intersection of two radical ideals is also radical.

Exercise 11.76. Show that φ in the proof of Theorem 11.72 is a ring isomomorphism.

Exercise 11.77. Show that in the proof of Theorem 11.72, any $b \in V_{\varphi(F)}$ can be rewritten to obtain an element of V_F. *Hint:* Reverse the translation that defines φ.

Exercise 11.78. Show that in the proof of Theorem 11.72, $\varphi(f) \in \mathbb{F}$ if and only if $f \in \mathbb{F}$.

Exercise 11.79. Show that η in the proof of Theorem 11.72, if J is an ideal of $\mathbb{F}[y_1, \ldots, y_n]$, then $\eta(J)$ is an ideal of $\mathbb{F}[y_2, \ldots, y_n]$. *Hint:* $\mathbb{F}[y_2, \ldots, y_n] \subsetneq \mathbb{F}[y_1, \ldots, y_n]$ and $\eta(J) = J \cap \mathbb{F}[x_2, \ldots, y_n]$ is an ideal of $\mathbb{F}[y_2, \ldots, y_n]$.

11.7: Elementary applications

We now turn our attention to posing, and answering, questions that make Gröbner bases interesting. As in Section 11.6,

- $\mathbb{F}$ is an *algebraically closed* field—that is, all polynomials over $\mathbb{F}$ have their roots in $\mathbb{F}$;
- $\mathcal{R} = \mathbb{F}[x_1, x_2, \ldots, x_n]$ is a polynomial ring;
- $F \subset \mathcal{R}$;
- $V_F \subset \mathbb{F}^n$ is the set of common roots of elements of F;
- $I = \langle F \rangle$; and
- $G = (g_1, g_2, \ldots, g_m)$ is a Gröbner basis of I with respect to an admissible ordering.

Note that $\mathbb{C}$ is algebraically closed, but $\mathbb{R}$ is not, since the roots of $x^2 + 1 \in \mathbb{R}[x]$ are not in $\mathbb{R}$.

Our first question regards membership in an ideal.

Theorem 11.80 (The Ideal Membership Problem)**.** Let $p \in \mathcal{R}$. The following are equivalent:
(A) $p \in I$, and
(B) p top-reduces to zero with respect to G.

Proof. That (A) $\Longrightarrow$ (B): Assume that $p \in I$. If $p = 0$, then we are done. Otherwise, the definition of a Gröbner basis implies that $\mathrm{lm}(p)$ is top-reducible by some element of G; let r be the result of this top-reduction. By Proposition 11.51, $\mathrm{lm}(r_1) < \mathrm{lm}(p)$. By the definition of an ideal, $r_1 \in I$. If $r_1 = 0$, then we are done; otherwise the definition of a Gröbner basis implies that $\mathrm{lm}(p)$ is top-reducible by some element of G. Continuing as above, we generate a list of polynomials $p, r_1, r_2, \ldots$ such that

$$\mathrm{lm}(p) > \mathrm{lm}(r_1) > \mathrm{lm}(r_2) > \cdots.$$

By the well-ordering of $\mathbb{M}$, this list cannot continue indefinitely, so eventually top-reduction must be impossible. As long as $r_i \neq 0$, we can continue this indefinitely, so the chain must terminate with $r_i = 0$.

That (B) $\Longrightarrow$ (A): Assume that p top-reduces to zero with respect to G. By Lemma 11.55, $p \in I$. □

Now that we have ideal membership, let us return to a topic we considered briefly in Chapter 7. In Exercise 8.24 you showed that

> …the common roots of $f_1, f_2, \ldots, f_m$ are common roots of all polynomials in the ideal I.

Since $I = \langle G \rangle$, the common roots of $g_1, g_2, \ldots, g_m$ are common roots of all polynomials in I. Thus if we start with a system F, and we want to analyze its polynomials, we can do so by analyzing the roots of any Gröbner basis G of $\langle F \rangle$. This might seem unremarkable, except that like triangular linear systems, *it is easy to analyze the roots of Gröbner bases!* Our next result gives an easy test for the existence of common roots.

Theorem 11.81. The following both hold.
(A) $V_F = V_G$; that is, common roots of F are common roots of G, and vice versa.
(B) F has no common roots if and only if G contains a nonzero constant polynomial.

Proof. (A) Let $\alpha \in V_F$. By definition, $f_i(\alpha_1, \ldots, \alpha_n) = 0$ for each $i = 1, \ldots, m$. By construction, $G \subseteq \langle F \rangle$, so $g \in G$ implies that $g = h_1 f_1 + \cdots + h_m f_m$ for certain $h_1, \ldots, h_m \in \mathcal{R}$. By substitution,

$$\begin{aligned} g(\alpha_1, \ldots, \alpha_n) &= \sum_{i=1}^{m} h_i(\alpha_1, \ldots, \alpha_n) f_i(\alpha_1, \ldots, \alpha_n) \\ &= \sum_{i=1}^{m} h_i(\alpha_1, \ldots, \alpha_n) \cdot 0 \\ &= 0. \end{aligned}$$

That is, α is also a common root of G. In other words, $V_F \subseteq V_G$.

On the other hand, $F \subseteq \langle F \rangle = \langle G \rangle$ by Exercise 11.58, so a similar argument shows that $V_F \supseteq V_G$. We conclude that $V_F = V_G$.

(B) Let g be a nonzero constant polynomial, and observe that $g(\alpha_1, \ldots, \alpha_n) \neq 0$ for any $\alpha \in \mathbb{F}^n$. Thus, if $g \in G$, then $V_G = \emptyset$. By (A), $V_F = V_G = \emptyset$, so F has no common roots if G contains a nonzero constant polynomial.

For the converse, we need the Weak Nullstellensatz, Theorem 11.72 on page 367. If F has no common roots, then $V_F = \emptyset$, and by the Weak Nullstellensatz, $I = \mathcal{R}$. In this case, $1_{\mathcal{R}} \in I$. By definition of a Gröbner basis, there is some $g \in G$ such that $\operatorname{lm}(g) \mid \operatorname{lm}(1_{\mathcal{R}})$. This requires g to be a constant. □

Once we know common solutions exist, we want to know how many there are.

Theorem 11.82. There are finitely many complex solutions if and only if for each $i = 1, \ldots, n$ we can find $g \in G$ and $a \in \mathbb{N}$ such that $\operatorname{lm}(g) = x_i^a$.

Remark 11.83. Theorem 11.82 is related to the strong Nullstellensatz.

Proof. We can find $g \in G$ and $\alpha \in \mathbb{N}$ such that $\operatorname{lm}(g) = x_i^a$ for each $i = 1, 2, \ldots, n$ if and only if $\mathcal{R}/I$ is finite; see Figure 11.2. The definition $\mathcal{R}/I$ is independent of any monomial ordering,

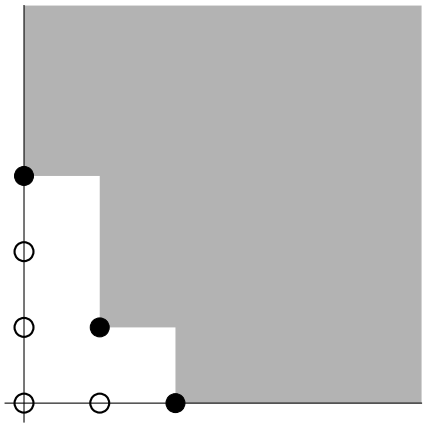


Figure 11.2. This monomial diagram shades the monomials divisible by the leading monomials of a Gröbner basis of I. If $\mathcal{R}/I$ is finite, then we cannot find infinitely many polynomials in $\mathcal{R}$ and outside I. This includes the axes of the monomial diagram, which consist of the monomials x, x^2, x^3, ... and y, y^2, y^3, ... They *must* reduce into a finite $\mathcal{R}/I$, so the Gröbner basis must have polynomials whose leading monomials divide them: in this case, x^2 and y^3.

so we can assume the ordering is lexicographic without loss of generality.

Assume first that for each $i = 1,\ldots,n$ we can find $g \in G$ and $a \in \mathbb{N}$ such that $\operatorname{lm}(g) = x_i^a$. Since x_n is the smallest variable, even $x_{n-1} > x_n$, so g must be a polynomial in x_n *alone;* any other variable in a non-leading monomial would contradict the assumption that $\operatorname{lm}(g) = x_n^a$. The Fundamental Theorem of Algebra implies that g has a complex solution. We can back-substitute these solutions into the remaining polynomials, using similar logic. Each back-substitution yields only finitely many solutions. There are finitely many polynomials, so G has finitely many complex solutions.

Conversely, assume G has finitely many solutions; call them $\alpha^{(1)},\ldots,\alpha^{(\ell)} \in \mathbb{F}^n$. Let

$$J = \left\langle x_1 - \alpha_1^{(1)},\ldots,x_n - \alpha_n^{(1)} \right\rangle \bigcap \cdots \bigcap \left\langle x_1 - \alpha_1^{(\ell)},\ldots,x_n - \alpha_n^{(\ell)} \right\rangle.$$

Recall that J is an ideal. You will show in the exercises that I and J have the same common solutions; that is, $V_I = V_J$.

For any $f \in \sqrt{I}$, the fact that $\mathcal{R}$ is an integral domain implies that

$$f(\alpha) = 0 \qquad \Longleftrightarrow \qquad f^a(\alpha) = 0 \; \exists a \in \mathbb{N}^+,$$

so $V_I = V_{\sqrt{I}}$. Let K be the ideal of polynomials that vanish on V_I. Notice that $I \subseteq \sqrt{I} \subseteq K$ by definition. We claim that $\sqrt{I} \supseteq K$ as well. Why? Let $p \in K$ be nonzero. Consider the polynomial ring $\mathbb{F}[x_1,\ldots,x_n,y]$ where y is a new variable. Let $A = \langle f_1,\ldots,f_m,1-yp\rangle$. Notice that $V_A = \emptyset$, since $f_i = 0$ for each i implies that $p = 0$, but then $1 - yp \neq 0$. By Theorem 11.81, any Gröbner basis of A has a nonconstant polynomial, call it c. By definition of A, there exist $H_1,\ldots,H_{m+1} \in \mathbb{F}[x_1,\ldots,x_n,y]$ such that

$$c = H_1 f_1 + \cdots + H_m f_m + H_{m+1}(1 - yp).$$

Let $h_i = c^{-1}H_i$ and

$$1 = h_1 f_1 + \cdots + h_m f_m + h_{m+1}(1 - yp).$$

Put $y = \frac{1}{p}$ and we have

$$1 = h_1 f_1 + \cdots + h_m f_m + h_{m+1} \cdot 0$$

where each h_i is now in terms of $x_1, \ldots, x_n$ and $1/p$. Clear the denominators by multiplying both sides by a suitable power a of p, and we have

$$p^a = h_1' f_1 + \cdots + h_m' f_m$$

where each $h_i' \in \mathcal{R}$. Since $I = \langle f_1, \ldots, f_m \rangle$, we see that $p^a \in I$. Thus $p \in \sqrt{I}$. Since p was abitrary in K, we have $\sqrt{I} \supseteq K$, as claimed.

We have shown that $K = \sqrt{I}$. Since K is the ideal of polynomials that vanish on V_I, and by construction, $V_{\sqrt{I}} = V_I = V_J$. You will show in the exercises that $J = \sqrt{J}$, so $V_{\sqrt{I}} = V_{\sqrt{J}}$. Hence $\sqrt{I} = \sqrt{J}$. By definition of J,

$$q_j = \prod_{i=1}^{\ell} \left(x_j - a_j^{(i)} \right) \in J$$

for each $j = 1, \ldots, n$. Since $\sqrt{I} = J$, suitable choices of $a_1, \ldots, a_n \in \mathbb{N}^+$ give us

$$q_1 = \prod_{i=1}^{\ell} \left(x_1 - \alpha_1^{(i)} \right)^{a_1}, \ldots, q_n = \prod_{i=1}^{\ell} \left(x_n - \alpha_n^{(i)} \right)^{a_n} \in I.$$

Notice that $\operatorname{lm}(q_i) = x_i^{a_i}$ for each i. Since G is a Gröbner basis of I, the definition of a Gröbner basis implies that for each i there exists $g \in G$ such that $\operatorname{lm}(g) \mid \operatorname{lm}(q_i)$. In other words, for each i there exists $g \in G$ and $a \in \mathbb{N}$ such that $\operatorname{lm}(g) = x_i^a$. □

Example 11.84. Recall the system from Example 11.53,

$$F = \left(x^2 + y^2 - 4, xy - 1 \right).$$

In Exercise 11.56 you computed a Gröbner basis in the lexicographic ordering. You probably obtained a superset of

$$G = \left(x + y^3 - 4y, y^4 - 4y^2 + 1 \right).$$

G is also a Gröbner basis of $\langle F \rangle$. Since G contains no constants, we know that F has common roots. Since $x = \operatorname{lm}(g_1)$ and $y^4 = \operatorname{lm}(g_2)$, we know that there are finitely many common roots.

We conclude by pointing in the direction of how to find the common roots of a system.

Theorem 11.85 (The Elimination Theorem)**.** Suppose the ordering is lexicographic with $x_1 > x_2 > \cdots > x_n$. For all $i = 1, 2, \ldots, n$, each of the following holds.

(A) $\widehat{I} = I \cap \mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$ is an ideal of $\mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$. (If $i = n$, then $\widehat{I} = I \cap \mathbb{F}$.)

(B) $\widehat{G} = G \cap \mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$ is a Gröbner basis of the ideal $\widehat{I}$.

Proof. For (A), let $f, g \in \widehat{I}$ and $h \in \mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$. Now $f, g \in I$ as well, we know that $f - g \in I$, and subtraction does not add any terms with factors from $x_1, \ldots, x_{i-1}$, so $f - g \in \mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$ as well. By definition of $\widehat{I}$, $f - g \in \widehat{I}$. Similarly, $h \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ as well, so $fh \in I$, and multiplication does not add any terms with factors from $x_1, \ldots, x_{i-1}$, so $fh \in \mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$ as well. By definition of $\widehat{I}$, $fh \in \widehat{I}$.

For (B), let $p \in \widehat{I}$. Again, $p \in I$, so there exists $g \in G$ such that $\operatorname{lm}(g)$ divides $\operatorname{lm}(p)$. The ordering is lexicographic, so g cannot have *any* terms with factors from $x_1, \ldots, x_{i-1}$. Thus $g \in \mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$. By definition of $\widehat{G}$, $g \in \widehat{G}$. Thus $\widehat{G}$ satisfies the definition of a Gröbner basis of $\widehat{I}$. □

The ideal $\widehat{I}$ is important enough to merit its own terminology.

Definition 11.86. For $i = 1, 2, \ldots, n$ the ideal $\widehat{I} = I \cap \mathbb{F}\left[x_i, x_{i+1}, \ldots, x_n\right]$ is called the **ith elimination ideal of I**.

Theorem 11.85 suggests that to find the common roots of F, we use a lexicographic ordering, then:

- find common roots of $G \cap \mathbb{F}[x_n]$;
- back-substitute to find common roots of $G \cap \mathbb{F}\left[x_{n-1}, x_n\right]$;
- ...
- back-substitute to find common roots of $G \cap \mathbb{F}[x_1, x_2, \ldots, x_n]$.

This is *exactly* how Gaussian elimination worked: reducing a matrix to row-echelon form gives us a polynomial in the bottom row whose solutions we can calculate easily, then back-substitute into previous rows.

Example 11.87. We can find the common solutions of the circle and the hyperbola in Figure 11.1 on page 352 using the Gröbner basis computed in Example 372 on page 11.84. Since

$$G = \left(x + y^3 - 4y, y^4 - 4y^2 + 1\right),$$

we have

$$\widehat{G} = G \cap \mathbb{C}[y] = \left\{y^4 - 4y^2 + 1\right\}.$$

It isn't hard to find the roots of this polynomial. Let $u = y^2$; the resulting substitution gives us the quadratic equation $u^2 - 4u + 1$ whose roots are

$$u = \frac{4 \pm \sqrt{(-4)^2 - 4 \cdot 1 \cdot 1}}{2} = 2 \pm \sqrt{3}.$$

Back-substituting u into $\widehat{G}$,

$$y = \pm\sqrt{u} = \pm\sqrt{2 \pm \sqrt{3}}.$$

We can now back-substitute y into G to find that

$$\begin{aligned} x &= -y^3 + 4y \\ &= \mp\left(\sqrt{2 \pm \sqrt{3}}\right)^3 \pm 4\sqrt{2 \pm \sqrt{3}}. \end{aligned}$$

Thus there are four common roots, all of them real, illustrated by the four intersections of the circle and the hyperbola.

Exercises.

Exercise 11.88. Determine whether $x^6 + x^4 + 5y - 2x + 3xy^2 + xy + 1$ is an element of the ideal $\left\langle x^2 + 1, xy + 1 \right\rangle$.

Exercise 11.89. Refer back to Exercise 11.68. How many solutions does this system have? If infinitely many, what is the dimension?

Exercise 11.90. Consider the system

$$F = \left(\begin{array}{l} xyz + xz + 3y + 3, \\ x^2yz^2 + x^2z^2 - y - 1 \end{array} \right).$$

Exercise 11.91. Suppose A, B are ideals of $\mathcal{R}$.

(a) Show that $V_{A \cap B} = V(A) \cup V(B)$.

(b) Explain why this shows that for the ideals I and J defined in the proof of Theorem 11.82, $V_I = V_J$.