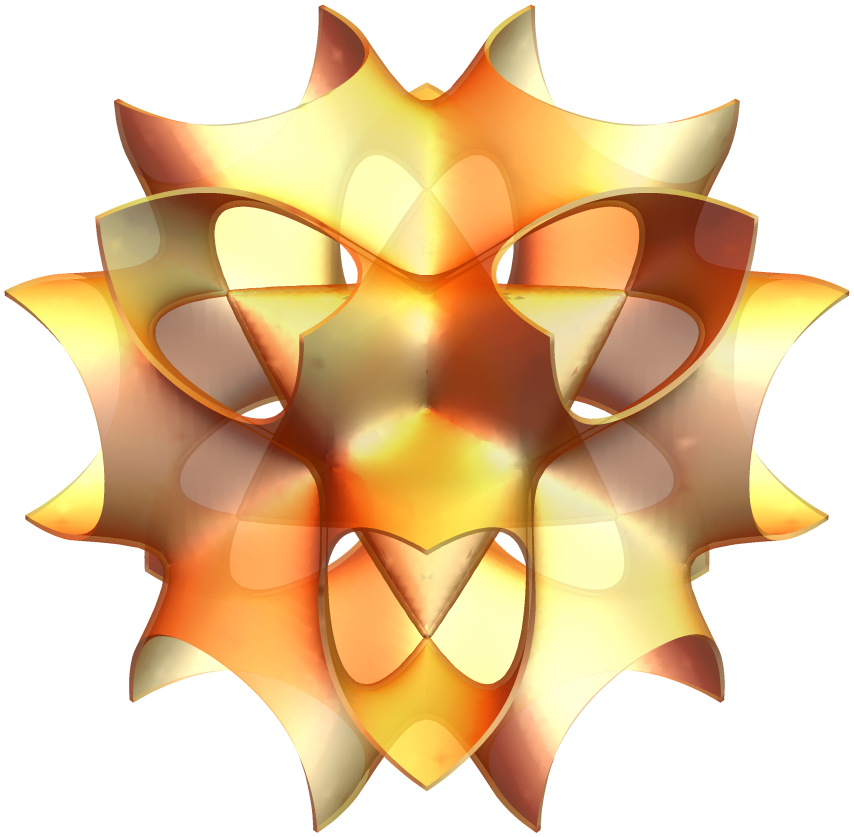
Nonlinear Algebraic Computation

Anton Leykin



[0]You are reading a **draft** compiled on **January 15, 2026**.

This is a collection of notes for several courses taught at Georgia Tech under the names *Introduction to Algebraic Computation* and *Nonlinear Algebra*.

This text may serve as a primer for a mathematician, scientist, or engineer who deals with problems that are intrinsically nonlinear and present themselves in the form of systems of polynomial equations. If nonempty, the solution set to such a problem may be a finite collection of isolated points or it may be a union of points, curves, surfaces, and higher-dimensional geometric objects.

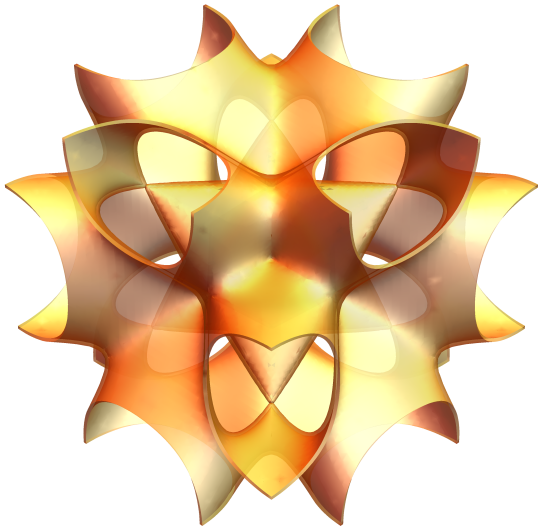


FIGURE 1. A surface defined by a polynomial of degree six in three variables and intersected with the cube $[-1, 1]^3 \subset \mathbb{R}^3$.

We pursue several philosophies of how to describe these solution sets in the frameworks of approximate numerical methods and exact symbolic computation. The reader is exposed to the concepts used in both and a translation from one world to the other.

Prior knowledge: There are no absolute prerequisites to reading this text except for a good understanding of linear algebra.

Notation and decorations: Terms that are being defined by the surrounding sentences are *underlined like this*. Terms *underlined like this* are outside of the immediate scope of these notes. We encourage the reader to read up on them elsewhere.

Contents

Chapter 1. Systems of polynomial equations 1
1.1. Solving univariate polynomial equations 1
1.1.1. What does "solve" mean? 1
1.1.2. Newton's method 2
1.1.3. Subdivision methods 3
1.1.4. Companion matrix 4
1.1.5. Euclidean algorithm, gcd, and resultant 6
1.2. Systems of multivariate polynomials 7
1.2.1. Eigenvalues of multiplication matrices 9
1.2.2. Elimination of variables 11
1.2.3. Rewriting polynomial systems 12

Chapter 2. Numerical homotopy continuation 13
2.1. Polynomial homotopy 13
2.1.1. Constructing a start system 13
2.1.2. Homotopy path tracking: predictor step 14
2.1.3. Homotopy path tracking: correction step 15
2.1.4. Heuristic homotopy tracking 16
2.1.5. Randomization and γ-trick 16
2.1.6. Non-square systems 18
2.1.7. Eigenvalue problem via homotopy continuation 18
2.2. Singular solutions 20
2.2.1. Deflation: univariate case 20
2.2.2. Deflation: general case 21
2.3. Certification 25
2.3.1. Real interval arithmetic 25
2.3.2. Certification of the roots of univariate polynomials 26
2.3.3. Krawczyk's method for multivariate systems 27

Chapter 3. Rings, ideals, and Gröbner bases 28
3.1. Polynomial rings and ideals 28
3.1.1. Ideals 29
3.1.2. Sum, product, and intersection of ideals 30
3.1.3. Ring maps and quotient rings 30
3.2. Gröbner bases 32
3.2.1. Monomial orders 32
3.2.2. Normal form algorithm 33

3.2.3. Initial ideal, Dickson's Lemma, Noetherianity 34
3.2.4. Gröbner bases and their properties 35
3.2.5. Buchberger's algorithm 37
3.3. Basic computations in polynomial rings 38
3.3.1. Computations in a quotient ring 39
3.3.2. Elimination 39

Chapter 4. Algebra-geometry correspondence 41
4.1. Ideal-variety correspondence 41
4.1.1. Hilbert's Nullstellensatz 42
4.1.2. Radical ideals 43
4.1.3. Irreducible varieties and prime ideals 44
4.2. Zariski topology, irreducible decomposition, and dimension 44
4.2.1. Varieties as Zariski closed sets 44
4.2.2. Irreducible decomposition of a variety 45
4.2.3. Dimension 46
4.3. Basic operations with ideals 48
4.3.1. Intersection of varieties: sum of ideals 48
4.3.2. Union of varieties: intersection or multiplication of ideals 48
4.3.3. Difference of varieties: colon ideal 49
4.3.4. Projection of variety: intersection with a subring 50
4.4. Multiplicity 50

Chapter 5. Numerical algebraic geometry 51
5.1. Witness sets 51
5.1.1. Definition 51
5.1.2. Numerical construction 53
5.1.3. Equivalence of witness sets 53
5.1.4. Sampling and the membership test 54
5.2. Numerical irreducible decomposition 54
5.2.1. Irreducible witness sets 54
5.2.2. Monodromy breakup algorithm 55
5.2.3. Linear trace test 56
5.3. Numerical variety 58
5.3.1. Union and difference 58
5.3.2. Intersection with a hypersurface 58
5.3.3. Constructing a numerical variety 60
5.3.4. Intersection 60
5.3.5. Singular witness sets 61
5.4. Trilingual dictionary 62

Chapter 6. Applications 64
6.1. Robotics 64
6.2. Automatic theorem proving 66
6.2.1. Implicitization 66
6.2.2. Proving geometric theorems 66
6.3. Chemical reaction networks 67

6.4. Astrodynamics 68
6.5. Signal processing 70
6.6. Computer vision 71

Appendix A. Homotopy continuation revisited 73
A.1. Singular solutions revisited 73
A.1.1. Puiseux series 73
A.1.2. Endgame 73

Appendix B. Certification 76
B.1. Alpha theory 76
B.1.1. Approximate zeros 76
B.1.2. Smale's α-theorem 76

CHAPTER 1

Systems of polynomial equations

1.1. Solving univariate polynomial equations

Let k be a field. In most examples here we assume that k is $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$: i.e., the field of rational, real, or complex numbers, respectively.

Note: Each subsequent field in the sequence of three is an *extension* of the previous one. One way to continue the sequence is

$$\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C} \subset \mathbb{H} \subset \mathbb{O}.$$

The last two, *quaternions* and *octonions*, are *division algebras* that are not fields (the multiplication is not commutative).

The above fields are of *characteristic* 0. Fields of characteristic p include, for instance, *finite fields*, which are extensively used in *cryptography* and *coding theory*.

A univariate polynomial,

$$f = f(x) = a_d x^d + a_{d-1} x^{d-1} + \ldots + a_1 x + a_0, \tag{1.1.1}$$

can be viewed as a function

$$f : k \to k.$$

We will refer to the k-linear infinite-dimensional space of polynomials as the *ring of polynomials with coefficients in k* and denote it by $k[x]$. We will come back to a formal definition of a *ring* later, for now it is sufficient to know that a ring is a vector space with an operation of multiplication: indeed, in $k[x]$ a product of two polynomials is a polynomial.

The polynomials of degree at most d, i.e., polynomials of the form (1.1.1), form a linear subspace $k[x]_{\leq d}$ of $k[x]$ of dimension $d+1$, but not a subring, since $k[x]_{\leq d}$ is not closed under multiplication.

1.1.1. What does "solve" mean? For equations of degree at most 4, there exist formulas that express all their *roots* in terms of *radicals*; e.g., see *Cardano formulas* for cubics. A crowning achievement of *Galois theory* is showing that a quintic equation can **not** be solved in radicals.

If the demand of exactness of solutions is dropped, then the basic problem of solving equations can be rephrased in the following way.

Problem 1.1.1. *For a polynomial $f \in \mathbb{C}[x]$ and fixed $\delta > 0$ and $\varepsilon > 0$ find $\tilde{x} \in \mathbb{C}$*

1. *such that $\|\tilde{x} - x^*\| < \delta$, where $x^* \in f^{-1}(0)$ is some exact root;*
2. *such that $\|f(\tilde{x})\| < \varepsilon$.*

The chosen numbers δ and ε are called the absolute error tolerance *and the* absolute residual tolerance*, respectively.*

Much of what is being said in this section will be generalized in the multivariate case, but until then the *norm* is simply the absolute value: i.e., $\|x\| = |x|$.

Note: There are numerous variations of this problem: one may

(1) require one or both conditions above to hold,
(2) restrict the search for a root to a specified region,
(3) replace the word "absolute" with "relative".

In numerical analysis the distance $\Delta x = |\tilde{x} - x^*|$ is sometimes referred to the *backward error*, whereas $|f(\tilde{x})| = |f(\tilde{x}) - f(x^*)|$ is called the *forward error*. The errors that are *normalized* (that requires the presence of a norm in the space of solutions and/or the space of polynomials) are referred to as *relative*, e.g., $\Delta x/|x|$ is the relative error of approximation of a root.

1.1.2. Newton's method. One of the most common methods to solve Problem 1.1.1 is *Newton's method* described below. For a polynomial $f \in k[x]$, define *Newton's operator*

$$N_f : k \to k, \quad N_f(x) = x - \frac{f(x)}{f'(x)}.$$

One step of Newton's method can be thought of graphically (see Figure 1: given x_0, take a tangent line to the graph of the function f at the point $(x_0, f(x_0))$, then the point of its intersection with the x-axis is $x_1 = N_f(x_0)$.

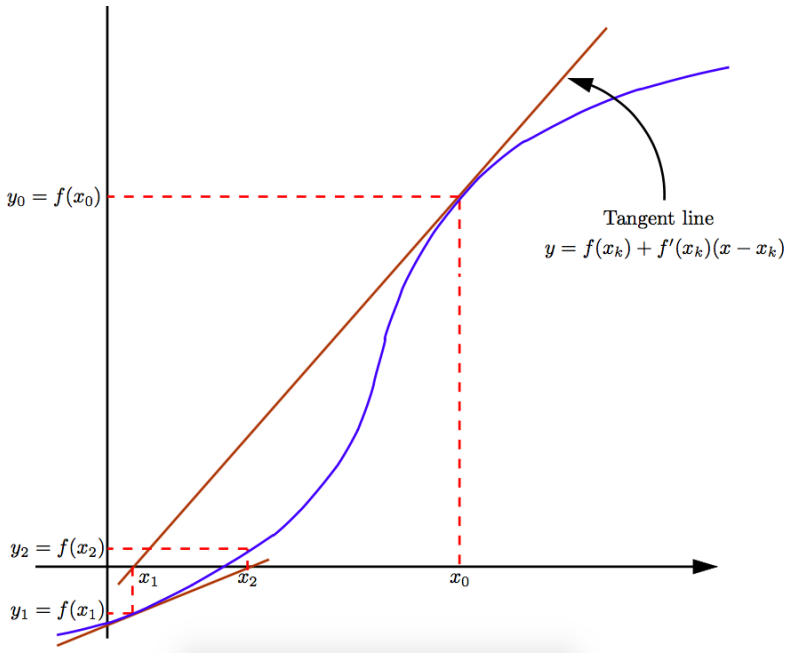


FIGURE 1. Two steps of Newton's method. Here, $f(c) = 0$.

Keep two caveats in mind:

- this algorithm is not guaranteed to terminate;
- for the result $\tilde{x}$ of the algorithm, the error $\|\tilde{x} - x^*\|$ is not necessarily bounded by δ for any (exact) root x^* of f.

Algorithm 1.1.1 $\tilde{x} = \text{NEWTON}(f, x_0, \delta)$

Require: $f \in k[x]$, a polynomial;
$x_0 \in k$, an initial approximation;
δ, the desired absolute error tolerance;
Ensure: $\|x_n - x_{n-1}\| < \delta$.

```
n ← 0
repeat
  n ← n + 1
  x_n ← N_f(x_{n-1})
until ||x_n − x_{n-1}|| < δ
x̃ ← x_n
```

Note: The termination criterion can be easily modified to focus on the desired residual tolerance and can come in both absolute and relative flavor (see remarks following Problem 1.1.1).

The point x^* is called a *regular* root of f if $f'(x^*) \neq 0$. The following proposition is straightforward.

Proposition 1.1.2. *x^* is a regular root of a polynomial f iff $x^* = N_f(x^*)$.*

Provided the sequence $x_0, x_1, \ldots$ converges to a regular root, it converges *quadratically*: roughly speaking, $|x_{n+1} - x^*| \simeq |x_n - x^*|^2$. If the root is *multiple*, i.e., non-regular, this can not be claimed.

The following exercise shows that not all initial points x_0 generate a convergent sequence.

Exercise 1.1.3. *Find all points x such that $x = N_f^2(x) = N_f(N_f(x))$ for $f = x^2 - ax$.*

Note: While classically Newton's method was designed for $k = \mathbb{R}$ and $k = \mathbb{C}$, it may be used looking for roots of polynomials in other fields; for instance, look up *Puiseux series*.

1.1.3. Subdivision methods. Another large class of numerical techniques is *subdivision methods*: given a search domain for a solution such methods proceed by subdividing this domain until

- either all pieces are shown to have no solutions
- or a piece guaranteed to contain a solution is found and this piece is small enough so that every point is an approximation to the solution within a prescribed error tolerance.

We illustrate this class of methods by the well-known *bisection method* to find a root of a real polynomial over a real segment.

Exercise 1.1.4. *Show that for any two distinct regular (real) roots of a polynomial $f \in \mathbb{R}[x]$ there is a (real) root of f' between them.*

Having the values $f(a)$ and $f(b)$ of different signs implies, by continuity, that the segment $[a, b]$ contains at least one root. This algorithm is guaranteed to terminate.

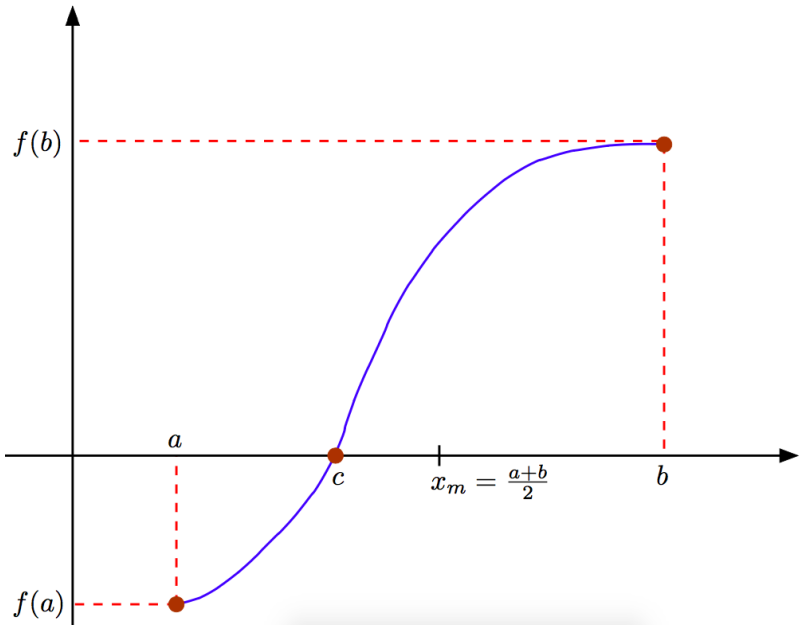


FIGURE 2. Bisection method. Here, $f(c) = 0$.

Algorithm 1.1.2 $\tilde{x} = \text{BISECTION}(f, a, b, \delta)$

Require: $f \in \mathbb{R}[x]$, a polynomial;
$a, b \in \mathbb{R}$, $a < b$, $f(a)f(b) < 0$;
δ, the desired absolute error tolerance;
Ensure: $\tilde{x}$ is an approximation of a root of f with absolute error not exceeding δ.

```
while b − a > 2δ do
  m ← (a+b)/2
  if f(m)f(a) > 0 then
    a ← m
  else
    b ← m
  end if
end while
x̃ ← m
```

1.1.4. Companion matrix. For a polynomial,

$$f = x^d + a_{d-1}x^{d-1} + \ldots + a_1 x + a_0 \in k[x],$$

define the *normal form* with respect to f as the following map:

$$\text{NF}_f : k[x] \to k[x]_{\leq d-1}, \quad \text{NF}_f(g) = \text{remainder of the division } g \text{ by } f,$$

where the remainder $r \in k[x]_{\leq d-1}$ is a unique polynomial of degree less than d satisfying $g = qf + r$, for some $q \in k[x]$.

Now define the "multiplication by x" operator:

$$M_x : k[x]_{\leq d-1} \to k[x]_{\leq d-1}, \quad M_x(g) = \mathrm{NF}_f(xg).$$

The linear map M_x is represented by the square matrix

$$A_x = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -a_{d-1} \end{bmatrix}$$

with respect to the (standard) *monomial basis*,

$$\begin{aligned} e_0 &= 1, \\ e_1 &= x, \\ e_2 &= x^2, \\ &\cdots \\ e_{d-1} &= x^{d-1}, \end{aligned}$$

of $k[x]_{\leq d-1}$.

The *companion matrix* of f is defined as A_x above.

Exercise 1.1.5. (1) *Show that* $A_x^d e_0 = \sum -a_i A_x^i e_0$.

(2) *Suppose the roots* $b_1, \cdots, b_d \in \mathbb{C}$ *of* f *are distinct. Prove that* A_x *is diagonalizable. (Hint:* Vandermonde *matrix provides the eigenvectors of* A_x^T*.)*

(3) *Find the Jordan canonical form of* A_x^T *for* $f = (x-b)^2$.

(4) *Let* $v(x) = (1, x, x^2, \cdots, x^{d-1})^T \in \mathbb{C}^d$. *Let* b *be a root of* f *with multiplicity* m. *Show that*

$$v_\alpha(b) = \frac{1}{\alpha!} \frac{\partial^\alpha v}{\partial x^\alpha}(b)$$

is a generalized eigenvector of A_x^T *for* $\alpha = 0, \ldots, m-1$.

(5) *Find the Jordan canonical form of* A_x^T *and conclude that* f *equals the characteristic and the minimal polynomials of* A_x.

The above exercise leads us to a conclusion that the roots of f and the eigenvalues of A_x are the same set. Therefore, the problem of finding roots of f can be restated as a problem of finding the *spectrum* of the companion matrix A_x.

Exercise 1.1.6. *Consider the linear operator* $M_g : k[x]_{\leq d-1} \to k[x]_{\leq d-1}$ *of multiplication by an arbitrary polynomial* g *constructed in a way similar to* M_x *(i.e., in the case* $g = x$*) above. Show that its* spectrum *is the set of values of* g *taken at the roots of* f.

Note: There is a trove of numerical *iterative methods* for approximating eigenvalues of matrices, including a *homotopy continuation method* that will be discussed in §2.1.7.

1.1.5. Euclidean algorithm, gcd, and resultant. We define the *gcd (greatest common divisor)* of a set of polynomials $S \subset k[x]$ to be the monic polynomial g of the largest degree such that $g|f$ for all $f \in S$.

The *Euclidean algorithm* can find the gcd for a set of two polynomials. Let $\mathrm{LC}(f)$ denote the leading coefficient of $f \in k[x]$ and let $\textsc{monic}(f) = f/\mathrm{LC}(f)$.

Algorithm 1.1.3 $g = \gcd(f_1, f_2)$

Require: $f_1, f_2 \in k[x]$, nonzero polynomials;
Ensure: g is the gcd of $S = \{f_1, f_2\}$.

while $f_2 \neq 0$ **do**
  $h \leftarrow f_2$
  $f_2 \leftarrow \mathrm{NF}_h(f_1)$
  $f_1 \leftarrow h$
end while
$g \leftarrow \textsc{monic}(f_1)$

Computing a gcd for any finite set of polynomials amounts to applying the above algorithm repetitively.

Note that the command

$$f_2 \leftarrow \mathrm{NF}_h(f_1)$$

finds the polynomial f_2 of the smallest degree such that $f_1 = f_2 + qh$ for a some polynomial q.

Exercise 1.1.7. *Describe algorithms to*

(1) *find the* quotient *q and the* remainder *f_1 in the division-with-remainder procedure described above;*
(2) *find a pair of nonzero polynomials $(c_1, c_2) \in k[x]^2$ of minimal possible degrees*
  (a) *such that $g = \gcd(f_1, f_2) = c_1 f_1 + c_2 f_2$;*
  (b) *such that $c_1 f_1 + c_2 f_2 = 0$.*

Exercise 1.1.8. *Suppose* $\gcd(f_1, f_2) = 1$. *Show that if $(c_1, c_2) \in k[x]^2$ satisfy $c_1 f_1 + c_2 f_2 = 0$, then (c_1, c_2) is a multiple of $(f_2, -f_1)$: i.e., there is $h \in k[x]$ such that $c_1 = h f_2$ and $c_2 = -h f_1$.*

Note that in part (2b) of the Exercise 1.1.7 if $g = \gcd(f_1, f_2) \neq 1$ then $\deg c_1 < d_1 = \deg f_2$ and $\deg c_2 < d_1 = \deg f_1$. Indeed, in that case $f_1 = h_1 g$ and $f_2 = h_2 g$ for some polynomials h_1 and h_2; setting $(c_1, c_2) = (h_2, h_1)$ we have a pair with degrees satisfying the specified inequalities.

This means that the set of polynomials

$$S = \left\{ f_1,\ x f_1,\ \ldots,\ x^{d_2-1} f_1,\quad f_2, x f_2,\ \ldots,\ x^{d_1-1} f_2 \right\} \subset k[x]_{\leq d_1+d_2-1}$$

is linearly dependent. Let

$$f_1 = a_{d_1} x^{d_1} + a_{d_1-1} x^{d_1-1} + \ldots + a_1 x + a_0, \quad a_i \in k, \quad a_{d_1} \neq 0;$$
$$f_2 = b_{d_2} x^{d_2} + b_{d_2-1} x^{d_2-1} + \ldots + b_1 x + b_0, \quad b_i \in k, \quad b_{d_2} \neq 0.$$

The *resultant* $\mathrm{Res}(f_1, f_2)$ is the determinant of the *Sylvester matrix*, a square matrix of size $d_1 + d_2$ with the rows that are the coefficient vectors of polynomials in the set S above. There are two

blocks in this matrix:

$$\begin{array}{c} \\ f_1 \\ xf_1 \\ \cdots \\ x^{d_2-1}f_1 \end{array} \begin{array}{c} \begin{array}{cccccccccc} x^{d_1+d_2-1} & x^{d_1+d_2-2} & \cdots & \cdots & \cdots & \cdots & \cdots & x^2 & x & 1 \end{array} \\ \begin{bmatrix} & & & a_{d_1} & a_{d_1-1} & & \cdots & & a_1 & a_0 \\ & & a_{d_1} & a_{d_1-1} & & \cdots & & a_1 & a_0 & \\ & \iddots & \iddots & & & & \iddots & \iddots & & \\ a_{d_1} & a_{d_1-1} & & \cdots & & a_1 & a_0 & & & \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ f_2 \\ xf_2 \\ \cdots \\ x^{d_1-1}f_2 \end{array} \begin{array}{c} \begin{array}{cccccccccc} x^{d_1+d_2-1} & x^{d_1+d_2-2} & \cdots & \cdots & \cdots & \cdots & \cdots & x^2 & x & 1 \end{array} \\ \begin{bmatrix} & & & b_{d_2} & b_{d_2-1} & & \cdots & & b_1 & b_0 \\ & & b_{d_2} & b_{d_2-1} & & \cdots & & b_1 & b_0 & \\ & \iddots & \iddots & & & & \iddots & \iddots & & \\ b_{d_2} & b_{d_2-1} & & \cdots & & b_1 & b_0 & & & \end{bmatrix} \end{array}$$

THEOREM 1.1.9. $\operatorname{Res}(f_1, f_2) = 0$ *iff* $\gcd(f_1, f_2) \neq 1$.

PROOF. If the two polynomials share a nontrivial common factor, then following the discussion after Exercise 1.1.7 we conclude that the resultant vanishes. If not, Exercise 1.1.8 implies it does not. □

Corollary 1.1.10. *Let* $f_1, f_2 \in \mathbb{C}[x]$. *Then* $\operatorname{Res}(f_1, f_2) = 0$ *iff there exist* $x \in \mathbb{C}$ *such that* $f_1(x) = f_2(x) = 0$.

Exercise 1.1.11. *Consider*

$$f_1 = x^3 + a_2x^2 + a_1x + a_0,$$
$$f_2 = x^3 + b_2x^2 + b_1x + b_0.$$

(1) *Derive constraints on the coefficients of* f_1 *and* f_2 *that are satisfied iff these two polynomials have a common factor of degree at least* 2. *(Hint: What is the rank of the Sylvester matrix in this case?)*
(2) *Find all pairs* (f_1, f_2) *with the above property provided* $a_2 = 2$, $b_2 = 1$, *and* $a_0 = b_1 = 0$.

1.2. Systems of multivariate polynomials

In this section we set the notation for polynomials in n variables $x = (x_1, x_2, \cdots, x_n)$.

A *monomial* is a product of variables written in the following form:

$$x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}, \quad \alpha \in \mathbb{N}^n.$$

The (total) *degree* of a monomial is $\deg x^\alpha = |\alpha| = \sum_{i=1}^n \alpha_i$.

The set of monomials together with the operation of mutiplication forms a *monoid*: in our multiindex notation, the product of two monomials is

$$x^\alpha x^\beta = x^{\alpha+\beta}.$$

The above describes an isomorphism of the monoid of monomials and the monoid $\mathbb{N}^n$ equipped with the operation of addition.

A polynomial with coefficients in k is a linear combination of monomials

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

with all but finitely many coefficients a_α equal to zero. It can be considered as a function from k^n to k.

The infinite-dimensional k-space of polynomials in n variables together with the operation of multiplication is called the *polynomial ring* $k[x] = k[x_1, \ldots, x_n]$. As in the univariate case, the k-subspace $k[x]_{\leq d}$ of polynomials of degree at most d is finite dimensional.

Exercise 1.2.1. *Find* $\dim_k (k[x_1, \ldots, x_n]_{\leq d})$.

A *system of polynomials* is an m-tuple of polynomials $F = (f_1, \ldots, f_m)$. We call the solution set of F a *variety* and denote it by $\mathbb{V}(F)$.

$$\begin{aligned}\mathbb{V}(F) &= \{\, x \in k^n \mid F(x) = 0 \,\} \\ &= \{\, (x_1, \ldots, x_n) \in k^n \mid f_1(x_1, \ldots, x_n) = \cdots = f_n(x_1, \ldots, x_n) = 0 \,\}\end{aligned}$$

For the moment we distinguish two types of varieties: we call $V = \mathbb{V}(F) \subset k^n$

- 0-*dimensional* if V is a finite set;
- *positive-dimensional* if V is infinite.

We will focus on the case of dimension 0 until Chapter 3.

Exercise 1.2.2. *The variety $V = \mathbb{V}(F)$, $F = (f_1, \ldots, f_m)$, remains the same when*

(1) *the polynomials of the system F are permuted;*
(2) *F is replaced by*
$$\left(\sum_j c_{1j} f_j, \ldots, \sum_j c_{mj} f_j \right),$$
where $C = (c_{ij}) \in k^m$ is an invertible matrix.
(3) *another polynomial of the form*
$$\sum_{i=1}^{m} g_i f_i, \quad g_i \in k[x],$$
is appended to F.

A system $F = (f_1, \ldots, f_m)$ can be considered as a map $F : k^n \to k^m$. The matrix of its partial derivatives is called the *Jacobian* (matrix) of F and denoted

$$\frac{\partial F}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

We see that $\frac{\partial F}{\partial x}$ is an $m \times n$ matrix with polynomial entries.

A solution $x^* \in \mathbb{V}(F)$ is called *regular* (or *nonsingular*) if $\frac{\partial F}{\partial x}(x^*)$ is of full rank and *singular*, otherwise.

1.2.1. Eigenvalues of multiplication matrices. Consider the following polynomial system:

$$F = (f_1, f_2, f_3) = \begin{pmatrix} \boxed{x^2} - y^2 \\ \boxed{y^3} - 2xy - y^2 + 2x \\ \boxed{xy^2} - 3xy + 2x \end{pmatrix}.$$

We will refer to the monomials that are highlighted as *leading monomials* of the corresponding polynomials: in this example we order monomials by their degree and see x "heavier" than y to break the ties. A precise definition of a *monomial ordering* will be given in Chapter 3.

We denote by $\mathrm{LM}(f)$ the leading monomial of a polynomial f and by $\mathrm{LT}(f)$ its *leading term*, i.e., the leading monomial together with its coefficient: e.g., for

$$f = \boxed{2x^5} - 3y + 1,$$

we have $\mathrm{LT}(f) = 2\mathrm{LM}(f) = 2x^2$. Note that if f is a part of a polynomial system, we can always replace it with f divided by the *leading coefficient*. This makes the polynomial *monic*: $\mathrm{LT}(f) = \mathrm{LM}(f)$.

The set of *standard monomials* S consists of monomials not divisible by any of the leading monomials. For our example,

$$S = \left\{1, x, y, xy, y^2\right\}.$$

The standard monomials can be seen as the monomials under the *staircase* (see Figure 3) based on the leading monomials of F: the monomials that are not divisible by the leading monomials of F.

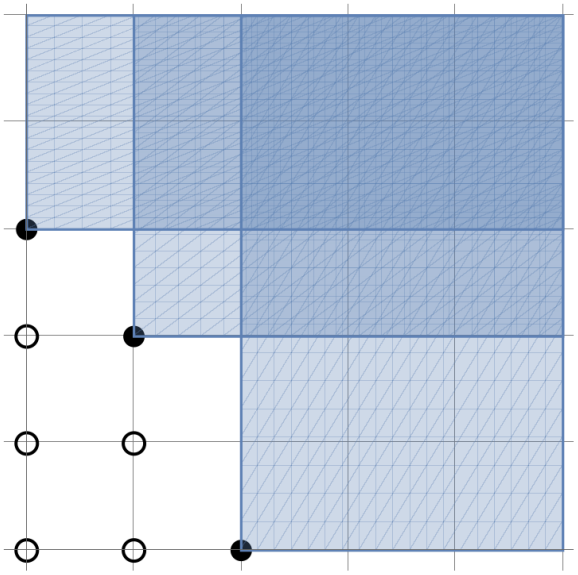


FIGURE 3. The monomial staircase given by $\{x^2, xy^2, y^3\}$; standard monomials $\{1, x, y, xy, y^2\}$.

For the following approach to work we need F to have special properties; these will be made precise when we define a *reduced Gröbner basis* (in Chapter 3). In particular, it is required that S is finite and the "tails" of polynomials in F contain monomials only in S.

Let us generalize the companion matrix method of §1.1.4. Consider the vector space

$$V = \operatorname{Span} S = \{\, f \in k[x,y] \,|\, \operatorname{supp}(f) \subseteq S \,\}$$

of polynomials with support contained in S. Define the operator of multiplication by x:

$$M_x : V \to V, \quad M_x(g) = \operatorname{NF}_F(xg).$$

We postpone the exact definition of *normal form* function NF_F, the generalization of remainder in the univariate polynomial division. However, intuitively, it should "reduce" polynomials "modulo F" until the result is contained in the span of S. This is possible to carry out in our example for the elements of xS, e.g.

$$\begin{aligned}\operatorname{NF}_F(x^2) &= x^2 - (x^2 - y^2) = y^2,\\ \operatorname{NF}_F(xy^2) &= xy^2 - (xy^2 - 3xy + 2x) = 3xy - 2x,\end{aligned}$$

Mor thean one step is necessary for x^2y: we can rewrite

$$x^2y - y(x^2 - y^2) = y^3,$$

but $y^3 \notin S$, so we need to continue the process:

$$\operatorname{NF}_F(x^2y) = y^3 - (y^3 - 2xy - y^2 + 2x) = 2xy + y^2 - 2x.$$

Noting that $\operatorname{NF}_F(xg) = xg$ for $g \in \{1, y\}$ we construct the matrix of M_x with respect to the basis S of V:

$$A_x = \begin{array}{c} \\ 1 \\ x \\ y \\ xy \\ y^2 \end{array} \begin{array}{c} \begin{array}{ccccc} M_x(1) & M_x(x) & M_x(y) & M_x(xy) & M_x(y^2) \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

The eigenvalues of A_x give us the x-coordinates of the points $(x, y) \in \mathbb{V}(F)$. In our example the spectrum of A_x is $\{-1, 0, 1, 2\}$. Substituting one of these values in F produces a univariate problem:

$$F|_{x=1} = \left(-y^2 + 1, y^3 - y^2 - 2y + 2, y^2 - 3y + 2\right)$$

The gcd of these polynomials is $y - 1$, which means that there is only one solution, namely $(1, 1)$, to the system F with $x = 1$.

Exercise 1.2.3. *Construct the matrix A_y for the system F above.*

(1) *Find the eigenvalues of A_y.*
(2) *Find all points in $\mathbb{V}(F)$ that have the y-coordinate equal to 1.*
(3) *Let $g = xy + y + 1$ and let M_g be the operator of multiplication by g constructed similarly to M_x above. Find the eigenvalues of M_g (cf. Exercise 1.1.6)*

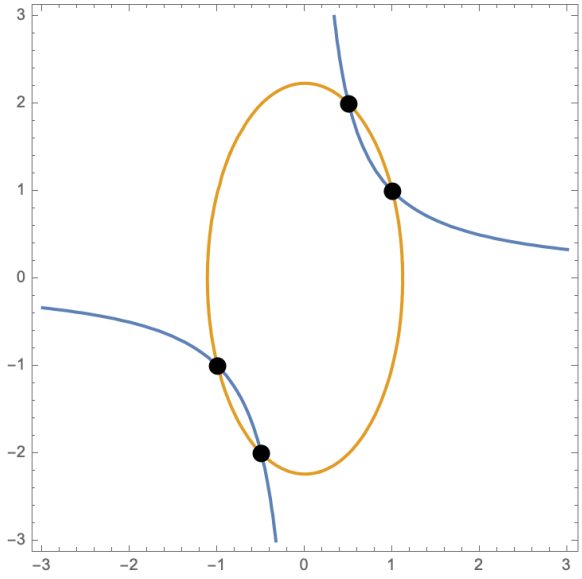


FIGURE 4. Points of intersection of $\mathbb{V}(xy-1)$ and $\mathbb{V}\big(4x^2+y^2-5\big)$.

1.2.2. Elimination of variables. Consider the system of two equations in two variables

$$F = (f_1, f_2) = \begin{pmatrix} xy-1 \\ 4x^2+y^2-5 \end{pmatrix}.$$

The point of the variety $\mathbb{V}(F)$ can be obtained by intersecting the hyberbola $\mathbb{V}(f_1)$ and the ellipse $\mathbb{V}(f_2)$ (see Figure 4).

Let us think of f_1 and f_2 as univariate polynomials of y (with coefficients that depend on x). Then, according to Theorem 1.1.9 and Corollary 1.1.10, they have a common solution iff their resultant vanishes:

$$\operatorname{Res}_x(f_1, f_2) = \begin{vmatrix} & y & -1 \\ y & -1 & \\ 4 & & y^2-5 \end{vmatrix} = -y^4+5y^2-4 = 0.$$

Solving the resulting equation, we get $y \in \{-2,-1,1,2\}$. Substituting these back in F we compute the four intersection points of the hyperbola and the ellipse:

$$\left\{\left(-\frac{1}{2},-2\right),(-1,-1),(1,1),\left(\frac{1}{2},2\right)\right\}.$$

In principle, resultants can be used to solve systems in an arbitrary number of variables simply by eliminating variables one by one. However, the size of polynomial expressions in this method typically grows very rapidly as suggested by the following exercise.

Exercise 1.2.4. *Show that* $\operatorname{Res}(f_1, f_2)$ *as defined in Section 1.1.5 is a (multivariate) polynomial in* d_1+d_2+2 *variables of degree* d_1+d_2*, where* $d_1 = \deg f_1$ *and* $d_2 = \deg f_2$*.*

An alternative technique relying on *Gröbner bases* is proposed in Chapter 3

1.2.3. Rewriting polynomial systems. Now that we touched upon a topic of elimination of variables, one may wonder whether anything can be gained by a reverse procedure: Can we simplify a system by introducing more variables?

The answer depends on what "simpler" means. While elimination leads to a system of equations with fewer variables, it typically increases the degrees of the polynomials. On the other hand, increasing the number of variables one can always obtain a system of quadratic equations.

Example 1.2.5. *Consider a system of two polynomials in two variables:*

$$F = \begin{pmatrix} y^3 - 6x^2 - 2xy + 5y^2 + 2x \\ x^3 - 3x^2 - 3xy + 3y^2 + 2x \end{pmatrix}.$$

If we make substitutions $u = x^2$ and $v = y^2$, we obtain the augmented system *of four quadratic equations*

$$G = \begin{pmatrix} vy - 6u - 2xy + 5v + 2x \\ ux - 3u - 3xy + 3v + 2x \\ u - x^2 \\ v - y^2 \end{pmatrix}$$

in $k[x, y, u, v]$

This transformation can be described by two maps: $\phi : k^2 \to k^4$, $(x, y) \mapsto (x, y, x^2, y^2)$, sending a point in the old coordinates to a point in the new coordinates and the monomial map

$$\begin{aligned} \psi : k[x, y, u, v] &\to k[x, y], \\ u &\mapsto x^2 \\ v &\mapsto y^2 \\ x &\mapsto x \\ y &\mapsto y \end{aligned}$$

sending the first two polynomials of G to F and the other two to zero.

The restrictions of maps ϕ and the projection $\pi : k^4 \to k^2$, $(x, y, u, v) \mapsto (x, y)$, give a bijection of the varieties $\mathbb{V}(G)$ and $\mathbb{V}(F)$.

We call a *binomial* a polynomial with two terms and a *trinomial* a polynomial with three terms.

Exercise 1.2.6. *Show that*

(1) *every system of binomial equations can be transformed into an augmented system of quadratic binomial equations by monomial substitutions;*
(2) *every polynomial system can be transformed into a system of trinomials by binomial substitutions (i.e., by iteratively introducing new variables that are equal to binomials in the old variables).*

CHAPTER 2

Numerical homotopy continuation

a

2.1. Polynomial homotopy

In this chapter we consider $\mathbb{C}[x] = \mathbb{C}[x_1, \ldots, x_n]$, polynomials with complex coefficients. Moreover, we restrict ourselves to a *square* polynomial systems, $F = (f_1, \ldots, f_n) \in \mathbb{C}[x]^n$, that are 0-dimensional, i.e., $\mathbb{V}(F)$ is finite.

The main idea behind solving the system F is to use the homotopy

$$H_t = (1-t)G + tF \in \mathbb{C}[x]^n, \quad t \in [0,1], \tag{2.1.1}$$

that connects a *start system* $G = H_0$ with the *target system* $F = H_1$. Now we need to create a start system G such that

- solutions of G are readily available;
- as the *continuation parameter* t varies from 0 to 1 we get a smooth paths that lead from solutions of G to solutions of F.

2.1.1. Constructing a start system. The following start system results in the so-called *total-degree homotopy*:

$$G = \left(x_1^{d_1} - 1, \ldots, x_n^{d_n} - 1 \right), \tag{2.1.2}$$

where $d_i = \deg f_i$ for $i \in [n]$. The number of solutions equals the *total degree* of the system, $|\mathbb{V}(G)| = d_1 \cdots d_n$. Indeed, the i-th coordinate of a solution is the d_i-th root of unity.

Example 2.1.1. *For the target system*

$$F = \begin{pmatrix} x_2^3 - 6x_1^2 - 2x_1x_2 + 5x_2^2 + 2x_1 \\ x_1^2 + x_2^2 - 2 \end{pmatrix}$$

the total-degree start system is

$$G = \begin{pmatrix} x_1^3 - 1 \\ x_2^2 - 1 \end{pmatrix}$$

and the start solutions *are*

$$\left\{ \left(-\frac{1}{2} + \frac{\sqrt{3}}{2}\boldsymbol{i}, 1 \right), \quad \left(-\frac{1}{2} - \frac{\sqrt{3}}{2}\boldsymbol{i}, 1 \right), \quad (1,1), \right.$$
$$\left. \left(-\frac{1}{2} + \frac{\sqrt{3}}{2}\boldsymbol{i}, -1 \right), \quad \left(-\frac{1}{2} - \frac{\sqrt{3}}{2}\boldsymbol{i}, -1 \right), \quad (1,-1) \right\}.$$

Exercise 2.1.2. *Show that all points $\mathbb{V}(G)$ for G in (2.1.2) are regular.*

Once the homotopy H_t and start solutions are set up, we follow the *homotopy paths* initiating at the start solutions (at $t = 0$) in hope of getting target solutions at ($t = 1$). A homotopy path, $x(t) \in \mathbb{C}^n$, can be tracked numerically using the so-called *predictor-corrector* technique (see Figure 1). Suppose an approximation $\tilde{x}_0 \approx x(t_0)$ of a solution to H_{t_0} is available for some $t_0 \in [0, 1]$ and

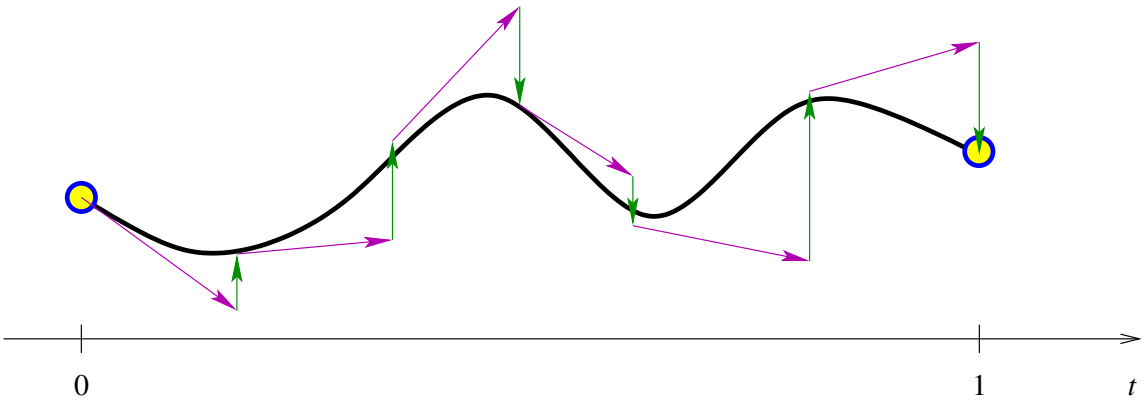


FIGURE 1. Predictor-corrector technique illustration

suppose we are looking to find an approximation $\tilde{x}_1 \approx x(t_1)$ of a solution to H_{t_1} for some $t_1 > t_0$. Two steps are performed: the *predictor* makes a rough approximation of $x(t_1)$, then the *corrector* refines predictor's approximation.

2.1.2. Homotopy path tracking: predictor step. Differentiating $H_t(x(t)) = 0$ with respect to t we get

$$\left(\frac{\partial H_t}{\partial x} x'(t) + \frac{\partial H_t}{\partial t}\right)_{x=x(t)} = 0,$$

solving which for the derivative of the homotopy path $x(t)$ gives the following at the point $x_0 = x(t_0)$ of the path

$$x'(t_0) = c(x_0, t_0) = \left(-\left(\frac{\partial H_t}{\partial x}\right)^{-1} \frac{\partial H_t}{\partial t}\right)_{x=x_0, t=t_0}. \tag{2.1.3}$$

This is a system of *ordinary differential equations (ODEs)* that can be *integrated numerically* using an arbitrary *numerical integration scheme*. We are, however, interested only in one step of a numerical integration procedure and list several popular choices here.

Order 0: This method makes the simplest prediction possible,

$$\tilde{x}_1 = \tilde{x}_0,$$

not using the ODEs (2.1.3) at all and, in particular, not depending on $\Delta t = t_1 - t_0$. As a global numerical integration scheme it is not very useful, but in our case the following corrector step makes even such a simple predictor step meaningful.

Order 1: The *tangent predictor* goes along the tangent line to the path:

$$\tilde{x}_1 = \tilde{x}_0 + c(\tilde{x}_0, t_0)\Delta t,$$

where the coefficient c is an approximation to $x'(t_0)$ derived by plugging in $x = \tilde{x}_0$ and $t = t_0$ in (2.1.3).

Order 2: The *Trapezoid predictor* operates as follows:

$$\tilde{x}_1 = \tilde{x}_0 + \frac{c(\tilde{x}_0, t_0) + c\left(\tilde{x}_0 + c(\tilde{x}_0, t_0)\Delta t,\, t_0 + \Delta t\right)}{2}\Delta t.$$

Note that this formula involves a part that is an exact copy of the tangent method.

Order m: In general, one can construct an integration scheme of an arbitrary order m. We skip the formal definition of the concept of *order*. Roughly speaking, a scheme is of order m if the error of approximatican be bounded from above as a constant multiple of $|\Delta t|^{m+1}$ as $\Delta t \to 0$.

The higher the order, the more accuracy is expected. On the other hand, the higher order methods typically are more involved in comparison to the lower order methods as demonstrated by the schemes of orders 0,1, and 2 above.

One of the most popular techniques is the *classical Runge-Kutta method*, which is or order 4.

2.1.3. Homotopy path tracking: correction step. One can extend *Newton's method* discussed in §1.1.2 to the multivariate setting. Let F be a square polynomial system of size n and assume one has an approximation $\tilde{x}_0$ of an exact solution $x^* \in \mathbb{C}^n$. Then it can be refined by *Newton's operator*

$$N_F : \left\{ x \in \mathbb{C}^n \,\middle|\, \det\left(\frac{\partial F}{\partial x}(x)\right) \neq 0 \right\} \to \mathbb{C}^n,$$
$$N_F(x) = x - \left(\frac{\partial F}{\partial x}(x)\right)^{-1} F(x).$$

using Algorithm 1.1.1 almost word for word.

As long as the solution x^* is regular, which in case of a square system is equivalent to nonvanishing of the determinant of the Jacobian matrix $\frac{\partial F}{\partial x}(x^*)$, Newton's method *converges quadratically*. This, roughly speaking, means that the accuracy of the approximation (the number of correct digits) doubles at every step. We leave the exact definition of quadratic convergence to §2.3.

Exercise 2.1.3. *Design an algorithm*

(1) $\tilde{x} = \textsc{corrector}(H, x_0, t)$ *that takes a homotopy* H*, an approximation* x *to a solution of the system* $F = H_t$ *for a given* t *and outputs* $N_F(x_0)$*.*

(2) $\tilde{x} = \textsc{predictor}(H, \tilde{x}_{t-\Delta t}, t, \Delta t)$ *that takes*
- *a homotopy* H*,*
- *an approximation* $\tilde{x}_{t-\Delta t}$ *to a solution of* $H_{t-\Delta t}$
- *for a given* t
- *and the step size* Δt

and outputs the prediction for a solution of H_t *using trapezoid method.*

2.1.4. Heuristic homotopy tracking. The word *heuristic* suggests that the algorithms described below terminate and produce a correct output for a large number of inputs and parameter settings, however, they do not provide a guarantee of neither termination nor correctness.

The most naïve algorithm of tracking a homotopy path is the following.

Algorithm 2.1.1 $\tilde{x}_1 = \text{NAIVEHOMOTOPYTRACKING}(H, \tilde{x}_0, N)$

Require: $H = H_t$, a polynomial homotopy as in 2.1.1;
$\tilde{x}_0 \in k$, an approximation to a solution of H_0;
N, the number of steps taken on the homotopy path;
Ensure: $\tilde{x}_1$, an approximation of a solution of H_1.

$\Delta t \leftarrow \frac{1}{N}$
for $t = \Delta t$ to 1 with step Δt **do**
  $\hat{x}_t \leftarrow \text{PREDICTOR}(H, \tilde{x}_{t-\Delta t}, t, \Delta t)$
  $\tilde{x}_t \leftarrow \text{CORRECTOR}(H, \hat{x}_t, t)$
end for

The hope is that for a large enough value of N the naïve algorithm does not deviate from the homotopy path. A more sophisticated approach adjusts the size of the step in accordance with the "difficulty" of prediction and correction. Algorithm 2.1.2 presents one popular approach.

> **Note:** The existing practical implementations of the heuristic homotopy tracking algorithms take even more parameters than TRACKHOMOTOPY in Algorithm 2.1.2. For instance, see the function `track` of `NumericalAlgebraicGeometry` package of `Macaulay2`.

Suppose the homotopy path $x(t)$ is *smooth* at every point perhaps with an exception of $t = 1$, i.e.,

$$\det\left(\frac{\partial H_t}{\partial x}(x(t))\right) \neq 0, \quad t \in [0, 1).$$

Above Algorithms 2.1.1 and 2.1.2, in fact, terminate and produce correct results for a sufficiently close approximation $\tilde{x}_0$ of $x(0)$, and sufficiently large N and Δt_0, respectively. However, it is practically impossible to determine the sufficient values of the parameters in the general case.

2.1.5. Randomization and γ-trick. For a path $x(t)$ containing a singular point, i.e., a singular solution $x^* = x(t^*)$ of the system H_{t^*} for some $t^* \in [0, 1)$, all numerical homotopy tracking algorithms are likely to fail: the corrector step becomes *ill-conditioned* close to x^*, since the Jacobian matrix is not invertible at x^*.

> **Note:** A good indicator of how close to singularity in the condition number of the Jacobian of H_{t^*} at the current approximation of $x(t^*)$.
> For a regular $n \times n$ matrix A the *condition number* is defined as
>
> $$\kappa(A) = \|A\| \, \|A^{-1}\|, \quad \|A\| = \max_{x \in \mathbb{C}^n - \{0\}} \frac{\|Ax\|}{\|x\|}.$$
>
> The larger is $\kappa(A)$, the less reliable is a numerical approximate solution of a linear system $Ax = b$.

Algorithm 2.1.2 $\tilde{x}_1 = \text{TRACKHOMOTOPY}(H, \tilde{x}_0, \Delta t_0, c, \delta, max_{corr})$

Require: $H = H_t$, a polynomial homotopy as in 2.1.1;
$\tilde{x}_0 \in k$, an approximation to a solution of H_0;
Δt_0, an initial step size;
c, the step increase factor;
δ, the backward error tolerance;
max_{corr}, the maximal number of corrections;
Ensure: $\tilde{x}_1$, an approximation of a solution of H_1.

$\tilde{x} \leftarrow \tilde{x}_0$, $t \leftarrow 0$
repeat
  repeat
    if $t + \Delta t > 1$ **then**
      $\Delta t = 1 - t$
    end if
    $\tilde{x}' \leftarrow \text{PREDICTOR}(H, \tilde{x}, t, \Delta t)$
    success $\leftarrow$ **false**, $i \leftarrow 0$
    while not *success* and $i < max_{corr}$ **do**
      $\tilde{x}'' \leftarrow \text{CORRECTOR}(H, \tilde{x}', t + \Delta t)$
      if $|\tilde{x}'' - \tilde{x}'| < \delta$ **then**
        success $\leftarrow$ **true**
      end if
      $\tilde{x}' \leftarrow \tilde{x}''$, $i \leftarrow i + 1$
    end while
    if *success* **then**
      $\Delta t \leftarrow \min(c\Delta t, 1 - t)$ - increase the step size
    else
      $\Delta t \leftarrow c^{-1}\Delta t$ - decrease the step size
    end if
  until *success*
  $t \leftarrow t + \Delta t$, $\tilde{x} \leftarrow \tilde{x}'$
until $t = 1$
$\tilde{x}_1 \leftarrow \tilde{x}$

Exercise 2.1.4. *Consider a homotopy*

$$H_t(x) = (1 - t)G(x) + \gamma t F(x). \tag{2.1.4}$$

Let $G = x^2 - 1$ *and* $F = x^2 + 2x - 3$. *Find all* $\gamma \in \mathbb{C}$ *such that there exists* $t \in \mathbb{R}$ *with a singular solution* x *to* H_t.

Exercise 2.1.4 suggests a way to avoid such singularities replacing a homotopy of (2.1.1) with that of (2.1.4) picking the value of $\gamma \in \mathbb{C}$ at random. This is known as the *γ-trick*.

Note: In practice, γ is picked on the unit circle in the complex plane with uniform distribution.

The following result allows us to compute *all* solutions of the target system via homotopy continuation.

THEOREM 2.1.5. *Let F be a square polynomial system.*
Then for the homotopy (2.1.4) using the total-degree start system G and a generic $\gamma \in \mathbb{C}$

(1) *for every start solution $x_0 \in \mathbb{V}(G)$ the homotopy path $x(t)$ starting at $x_0 = x(0)$ is regular for $t \in [0, 1)$;*
(2) *every isolated target solution $x_1 \in \mathbb{V}(F)$ is at the end of some homotopy path $x(t)$, i.e., $x_1 = x(1)$.*

PROOF. Postponed. □

2.1.6. Non-square systems. Consider a 0-dimensional *overdetermined* system

$$F = \begin{pmatrix} f_1 \\ \dots \\ f_m \end{pmatrix} \in \mathbb{C}[x_1, \dots, x_n]^m, \quad m > n.$$

We shall describe two ways of *squaring-up* an overdetermined system.

Take a matrix $A \in \mathbb{C}^{n \times m}$ and consider the square system $AF \in \mathbb{C}[x]^n$. Since the polynomials of AF are linear combinations of f_i,

$$\mathbb{V}(F) \subseteq \mathbb{V}(AF).$$

Proposition 2.1.6. *For a 0-dimensional system F and a generic A as above, the system AF is 0-dimensional.*

PROOF. Postponed until §4.2.3. □

Exercise 2.1.7. *If $x^* \in \mathbb{V}(F) \subset \mathbb{C}^n$ is a regular solution of an overdetermined system $F \in \mathbb{C}[x]^m$, then it is regular with respect to AF for a generic matrix $A \in \mathbb{C}^{n \times m}$.*

Exercise 2.1.8. *Construct an example of an overdetermined 0-dimensional system F in n variables such that every subset of n polynomials in F forms a positive-dimensional system.*

Proposition 2.1.6 suggests a simple method of finding $\mathbb{V}(F)$:

(1) pick a random A;
(2) find $\mathbb{V}(AF)$ for the square system AF;
(3) $\mathbb{V}(F) = \{\, x \in \mathbb{V}(AF) \,|\, F(x) = 0 \,\}$, which amounts to a finite number of checks for polynomial vanishing.

2.1.7. Eigenvalue problem via homotopy continuation. The following problem is central in linear algebra; here we use numerical homotopy continuation technique to solve it approximately.

Problem 2.1.9. *Given $A \in \mathbb{C}^{n \times n}$ find the eigenvalues and eigenvectors of A.*

Consider the case $n = 2$; the approach outlined here works for all n. Let

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad v = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

There are three unknowns in the eigenproblem: the eigenvalue λ and the coordinates of the vector v. We seek solutions of the system

$$(Av - \lambda v) = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 - \lambda x_1 \\ a_{21}x_1 + a_{22}x_2 - \lambda x_2 \end{pmatrix}.$$

This is a system of two quadratic polynomials in $\mathbb{C}[\lambda, x_1, x_2]$. The system is positive-dimensional: there are fewer equations than unknowns. Indeed, if v is an eigenvector, then there is the whole line of solutions corresponding to Span (v).

Exercise 2.1.10. *For the 2×2 matrix A find a condition (involving the entries a_{ij}) that ensures that the characteristic polynomial of A has a double root.*

Augment the system with one linear equation:

$$F = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 - \lambda x_1 \\ a_{21}x_1 + a_{22}x_2 - \lambda x_2 \\ b_1 x_1 + b_2 x_2 + 1 \end{pmatrix}.$$

Assume that the eigenspaces of A are one-dimensional (this is the case, in particular, when the eigenvalues are not repeated; cf. Exercise 2.1.10). If $b_1, b_2 \in \mathbb{C}$ are generic, the last equation in the system picks out one nonzero vector in each eigenspace. We conclude that F is a 0-dimensional system generically, since a randomly picked matrix has distinct eigenvalues.

Now if we solve the system F, consisting of two quadratic and one linear polynomials, using the total-degree homotopy, we would have to track $4 (= 2 \cdot 2 \cdot 1)$ paths. However, $|\mathbb{V}(F)| = 2$, meaning that this homotopy is not *optimal*.

Instead, let us take a start system that arises from the eigenproblem that we know a solution to: take a diagonal matrix D with distinct entries on the diagonal,

$$D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}, \quad d_1 \neq d_2$$

In a similar way as above we translate D into the 0-dimensional system

$$G = (Dv - \lambda v) = \begin{pmatrix} d_1 x_1 - \lambda x_1 \\ d_2 x_2 - \lambda x_2 \\ b_1 x_1 + b_2 x_2 + 1 \end{pmatrix}. \tag{2.1.5}$$

The two start points $(\lambda, x_1, x_2) \in \mathbb{V}(G)$ are $(d_1, \frac{-1}{b_1}, 0)$ and $(d_2, 0, \frac{-1}{b_2})$.

Exercise 2.1.11. *Show that the solutions of the system G in (2.1.5) are regular (for nonzero b_1 and b_2).*

The homotopy connecting G to F is

$$\begin{aligned} H_t = (1-t)G + tF &= \begin{pmatrix} (tD + (1-t)A)\, v - \lambda v \\ b_1 x_1 + b_2 x_2 + 1 \end{pmatrix} \\ &= \begin{pmatrix} ((1-t)d_1 + ta_{11}) & x_1 & + & ta_{12} & x_2 & - & \lambda x_1 \\ ta_{21} & x_1 & + & ((1-t)d_2 + ta_{22}) & x_2 & - & \lambda x_2 \\ b_1 & x_1 & + & b_2 & x_2 & + & 1 \end{pmatrix}. \end{aligned} \tag{2.1.6}$$

For all $t \in [0, 1]$ the system H_t represents the eigenproblem for the matrix

$$M_t = (1-t)D + tA. \tag{2.1.7}$$

For almost all choices of the entries d_1 and d_2 of D, the eigenvalues of M_t are distinct and the solutions $\mathbb{V}(H_t)$ are regular for all $t \in [0, 1)$.

Exercise 2.1.12. *Consider the homotopy M_t in (2.1.7) for*

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} \gamma & 0 \\ 0 & 0 \end{bmatrix}, \quad \gamma \in \mathbb{C}$$

Find the locus of values of γ that produce M_t with a repeated eigenvalue for some $t \in \mathbb{R}$. (Hint: See Exercise 2.1.10.)

2.2. Singular solutions

The γ-trick of Section 2.1.5 ensures that homotopy H_t is such that the points of variety $\mathbb{V}(H_t)$ are regular for all $t \in [0, 1)$. This guarantees fast convergence of the corrector step at every point of the path with a possible exception at the end $t = 1$.

If system F has a singular solution x^* then the straightforward implementation of a numerical homotopy path tracking algorithm would track the path until some $\widetilde{t} < 1$ and, hopefully, $\widetilde{t} \approx 1$ and it produces $\widetilde{x}$ with $H_{\widetilde{t}}(\widetilde{x}) \approx 0$ serving as a rough approximation of x^*.

Some advanced techniques for computing better approximation to singular isolated solutions of the target system are covered in Section A.1. This section focuses on one simple way to transform F into another system with a regular solution corresponding to x^*. If this is accomplished then, using Newton's method, $\widetilde{x}$ can be refined to a much better approximation.

2.2.1. Deflation: univariate case. First, let us consider the univariate case: suppose that $f \in \mathbb{C}[x]$ has a multiple root x^*. Then $f = (x - x^*)^m g$ for some $m > 1$ where g is not divisible by $x - x^*$.

Differentiating f we get

$$f' = m(x - x^*)^{m-1} g + (x - x^*)^m g' = (x - x^*)^{m-1}(mg + (x - x^*)g').$$

Since $mg + (x - x^*)g'$ does not have x^* as a root, we conclude that f' has x^* as a root with multiplicity $m - 1$.

Newton's method converges much slower around a multiple root. One way to restore fast quadratic convergence is to approximate x^* as a root of $f^{(m-1)}$.

However, in practice, the multiplicity m may be unknown. We may need to rely on a heuristic subroutine that gives up if the convergence is (heuristically) deemed to be slow.

Algorithm 2.2.1 $(\tilde{x}, fast) = \text{NewtonFast}(f, x_0, \delta)$

Require: $f \in k[x]$, a polynomial;
$x_0 \in k$, an initial approximation;
δ, the desired absolute error tolerance;

Ensure: either $fast =$ **true** (convergence is quadratic) and $\tilde{x}$ is an approximate root with backward error estimated to be at most δ
or $fast =$ **false** (convergence is not quadratic) and $\tilde{x}$ is a possibly finer approximation with no estimate on the error.

As mentioned above, to implement this algorithm, one would need to determine heuristically whether "convergence is quadratic". One practical way is to compute a fixed number (at least two) of Newton iterations and see if the error estimate from the last step is (roughly) the square of the error estimate from the step before last.

Then it is possible to design a heuristic regularization algorithm, Algorithm 2.2.2, that takes one derivative at a time. It is dubbed *deflation* as the multiplicity of the singular root decreases as we replace the original polynomial with its derivatives.

Algorithm 2.2.2 $\tilde{x} = \text{UNIVARIATEDEFLATION}(f, x_0, \delta)$

Require: $f \in k[x]$, a polynomial;
$x_0 \in k$, an initial approximation to a root of f;
δ, the absolute error tolerance;
Ensure: $\tilde{x}$ is an approximate root with backward error estimated to be at most δ.

```
x̃ ← x_0
repeat
  (x̃, fast) = NewtonFast(f, x̃, δ)
  if not fast then
    f ← f'
  end if
until fast
```

2.2.2. Deflation: general case. Now suppose that we have a system of equations $F = (f_1, \ldots, f_m)$ of multivariate polynomials $f_i \in \mathbb{C}[x] = \mathbb{C}[x_1, \ldots, x_n]$ with an isolated solution x^*. This implies $m \geq n$.

If x^* is singular, then the Jacobian $J = \frac{\partial F}{\partial x}(x^*)$ is rank deficient and has a nonzero kernel. That means that

$$\begin{aligned} r &= \operatorname{rank} J < n \text{ and} \\ c &= \operatorname{corank} J = n - r = \dim \ker J > 0. \end{aligned}$$

Pick a generic constant matrix $A = (a_{ij}) \in \mathbb{C}^{c \times n}$ and consider the following augmented system of polynomials in $\mathbb{C}[x, \lambda] = \mathbb{C}[x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_n]$:

$$\begin{pmatrix} F \\ \frac{\partial F}{\partial x}\,\lambda \\ A\lambda + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_n \\ \frac{\partial f_1}{\partial x_1}\lambda_1 + \cdots + \frac{\partial f_1}{\partial x_n}\lambda_n \\ \vdots \\ \frac{\partial f_m}{\partial x_1}\lambda_1 + \cdots + \frac{\partial f_m}{\partial x_n}\lambda_n \\ a_{11}\lambda_1 + \cdots + a_{1n}\lambda_n + 1 \\ \vdots \\ a_{c1}\lambda_1 + \cdots + a_{cn}\lambda_n + 1 \end{pmatrix} \tag{2.2.1}$$

The three blocks of polynomials in the system correspond to

(1) the original equations;
(2) equations that place the column vector of indeterminates λ in the kernel of the Jacobian;

(3) equations that describe a random r-plane (of dimension $r = n-c$) in the space of parameters λ (of dimension n).

Proposition 2.2.1. *For a generic choice of A the system (2.2.1) has an isolated solution $(x^*, \lambda^*) \in \mathbb{C}^{2n}$ for some $\lambda^* \in \mathbb{C}^n$.*

PROOF. Set $x = x^*$, then the second block of equations in the augmented system describes $K = \ker J$, where $J = \frac{\partial F}{\partial x}(x^*)$. Intersecting K, which is of dimension $c = n - r$, with a random r-plane cut out by the third block we get one point λ^*.

This leads to the conclusion that the (x^*, λ^*) solves the augmented system and is isolated. □

Taking a closer look at the system (2.2.1) one may notice that c parameters can be easily eliminated using the third block of equations. Alternatively, one can reformulate the augmentation procedure: take a generic constant matrix $B \in \mathbb{C}^{n\times(r+1)}$ and consider the system

$$D_B F = \begin{pmatrix} F \\ \frac{\partial F}{\partial x}\, B \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_r \\ 1 \end{bmatrix} \end{pmatrix} \in \mathbb{C}[x, \lambda]^{2m}, \tag{2.2.2}$$

where $\mathbb{C}[x, \lambda] = [x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_r]$. Note that the deflation procedure, denoted by D (we write D_B if the choice of a generic matrix B needs to be emphasized), produces DF, a system of m equations in $n + r$ unknowns, while the system (2.2.1) has $2m + c$ equations in $2n$ unknowns.

Exercise 2.2.2. *Consider the system*

$$F = \begin{pmatrix} x_1^2 - x_2^4 \\ x_1^2 - x_2^6 \end{pmatrix}.$$

(1) *Show that the origin $x^* = (0, 0)$ is a singular isolated solution of F.*
(2) *Find the rank r of the Jacobian at x^*.*
(3) *Construct the deflated system $DF = D_B F$ picking entries of B to be nonzero integers.*
(4) *Is the lifted solution $(x^*, \lambda^*) \in \mathbb{V}(DF)$ regular?*

Proposition 2.2.3. *For a generic choice of B the system (2.2.2) has an isolated solution $(x^*, \lambda^*) \in \mathbb{C}^{n+r}$ for some $\lambda^* \in \mathbb{C}^r$.*

PROOF. The argument is similar to that of Proposition 2.2.1.

Note that $B(\lambda, 1)^T$ parametrizes an r-plane and the the second block of equations in 2.2.2 implies that $(x^*, \lambda) \in \mathbb{V}(DF)$ iff

$$B \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \in \ker J, \quad J = \frac{\partial F}{\partial x}(x^*).$$

There is a unique point λ^* in the intersection of the $(n - r)$-dimensional kernel and a generic r-plane, which leads to the conclusion. □

The following algorithm can be seen as a generalization of Algorithm 2.2.2 to the multivariate setting. It relies on two subroutines:

- a multivariate generalization of NEWTONFAST(F, x_0, δ), which implements the multivariate Newton's method as in §2.1.3 (if given an overdetermined system F, Newton's method is applied to a square system formed using one of the approaches in §2.1.6);
- a heuristic function NUMERICALRANK$(\tilde{A})$, which given an approximation $\tilde{A}$ of a matrix A attempts to recover the (exact) rank of A.

Algorithm 2.2.3 $\tilde{x} =$ DEFLATION(F, x_0, δ)

Require: $F \in \mathbb{C}[x]^m$, a polynomial system;
$x_0 \in \mathbb{C}^n$, an initial approximation to an isolated solution of F;
δ, the absolute error tolerance;
Ensure: $\tilde{x}$ is an approximate solution with the estimated backward error at most δ.

$(\tilde{x}, fast) \leftarrow$ NEWTONFAST(F, x_0, δ)
if not $fast$ **then**
  $J \leftarrow \dfrac{\partial F}{\partial x}(\tilde{x})$
  $r \leftarrow$ NUMERICALRANK(J)
  $B \leftarrow$ a random $n \times (r+1)$ matrix
  $\tilde{\lambda} \leftarrow$ the least-squares approximate solution of $JB \begin{bmatrix} \lambda \\ 1 \end{bmatrix} = 0$
  $\tilde{x} \leftarrow$ the first n coordinates of DEFLATION$(D_B F, (\tilde{x}, \tilde{\lambda}), \delta)$
end if

Unlike in the univariate case it is not clear whether the recursion in DEFLATION terminates. In Section 4.4 we will define *multiplicity* $\mu_{x^*}(F)$ for an isolated solution x^*, such that $\mu(x^*) = 1$ iff x^* is a regular solution of F.

THEOREM 2.2.4. *Let $x^* \in \mathbb{V}(F)$ be an isolated solution of a polynomial system F and $(x^*, \lambda^*) \in \mathbb{V}(DF)$ be the corresponding solution of the deflated system DF.*
Then $\mu_{(x^*,\lambda^*)}(DF) < \mu_{x^*}(F)$.

PROOF. Give a reference or a sketch of proof in Section 4.4 □

Corollary 2.2.5. *Algorithm 2.2.3 terminates.*

Example 2.2.6. *Consider*

$$F = \begin{pmatrix} x_1^3 - x_2^2 \\ x_2^3 \end{pmatrix}.$$

This system has one solution, $x^* = (0, 0)$. However, x^* is singular:

$$\frac{\partial F}{\partial x} = \begin{bmatrix} 3x_1^2 & -2x_2 \\ 0 & 3x_2^2 \end{bmatrix}, \quad \frac{\partial F}{\partial x}(x^*) = 0, \quad r = 0.$$

The first deflation step is not using any new variables: take a generic $B_1 \in \mathbb{C}^{2\times 1}$, for this example $B_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ works, then

$$F_1 = D_{B_1} F = \begin{pmatrix} F \\ \frac{\partial F}{\partial x} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \end{pmatrix} = \begin{pmatrix} x_1^3 - x_2^2 \\ x_2^3 \\ 3x_1^2 - 2x_2 \\ 3x_2^2 \end{pmatrix} \in \mathbb{C}[x_1, x_2]^4.$$

However, x^* is singular for this system as well:

$$\frac{\partial F_1}{\partial x} = \begin{bmatrix} 3x_1^2 & -2x_2 \\ 0 & 3x_2^2 \\ \hline 6x_1 & -2 \\ 0 & 6x_2 \end{bmatrix}, \quad \frac{\partial F_1}{\partial x}(x^*) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \hline 0 & -2 \\ 0 & 0 \end{bmatrix}, \quad r = 1.$$

In the second deflation step we take a generic $B_2 \in \mathbb{C}^{2\times 2}$, here $B_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ works, to construct

$$F_2 = D_{B_2} D_{B_1} F = \begin{pmatrix} F_1 \\ \hline \frac{\partial F_1}{\partial x} \begin{bmatrix} 1 \\ \lambda_1 \end{bmatrix} \end{pmatrix} = \begin{pmatrix} x_1^3 - x_2^2 \\ x_2^3 \\ 3x_1^2 - 2x_2 \\ 3x_2^2 \\ \hline 3x_1^2 - 2x_2\lambda_1 \\ 3x_2^2\lambda_1 \\ 6x_1 - 2\lambda_1 \\ 6x_2\lambda_1 \end{pmatrix} \in \mathbb{C}[x_1, x_2, \lambda_1]^8.$$

The above system has one solution $(x^*, \lambda^*) = (0, 0, 0)$, which is still singular:

$$\frac{\partial F_2}{\partial (x,\lambda)} = \begin{bmatrix} & \frac{\partial F_1}{\partial x} & \\ \hline 6x_1 & -2\lambda_1 & -2x_2 \\ 0 & 6x_2\lambda_1 & 3x_2^2 \\ 6 & 0 & -2 \\ 0 & 6\lambda_2 & 6x_2 \end{bmatrix}, \quad \frac{\partial F_2}{\partial (x,\lambda)}(x^*, \lambda^*) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -2 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \\ 6 & 0 & -2 \\ 0 & 0 & 0 \end{bmatrix}, \quad r = 2.$$

In the second deflation step we take a generic $B_3 \in \mathbb{C}^{3\times 3}$*; we shall show that* $B = \mathrm{id}$ *does the trick. Construct*

$$F_3 = D_{B_3} D_{B_2} D_{B_1} F = \begin{pmatrix} F_2 \\ \hline \frac{\partial F_2}{\partial(x,\lambda)} \begin{bmatrix} \lambda_2 \\ \lambda_3 \\ 1 \end{bmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} x_1^3 - x_2^2 \\ x_2^3 \\ 3x_1^2 - 2x_2 \\ 3x_2^2 \\ 3x_1^2 - 2x_2\lambda_1 \\ 3x_2^2\lambda_1 \\ 6x_1 - 2\lambda_1 \\ 6x_2\lambda_1 \\ 3x_1^2\lambda_2 - 2x_2\lambda_3 \\ 3x_2^2\lambda_3 \\ 6x_1\lambda_2 - 2\lambda_3 \\ 6x_2\lambda_3 \\ 6x_1\lambda_2 - 2\lambda_1\lambda_3 - 2x_2 \\ 6x_2\lambda_1\lambda_3 + 3x_2^2 \\ 6\lambda_2 - 2 \\ 6\lambda_1\lambda_3 + 6x_2 \end{pmatrix} \in \mathbb{C}[x_1, x_2, \lambda_1, \lambda_2, \lambda_3]^{16}.$$

The above system has one solution $(x^*, \lambda^*) = (0, 0, 0, \frac{1}{3}, 0, 1)$*, which one can check is regular.*

Note: In practice, to avoid doubling the number of equations at each deflation step, one may want to square-up the deflated after each deflation step: use the first approach of §2.1.6.

Exercise 2.2.7. *For the system* F *of Exercise 2.2.2 construct a sequence of* s *deflations picking matrices* $B_1, \ldots, B_s$ *(take* s *to be as large as needed) such that the resulting system* $D_{B_s} \cdots D_{B_1} F$ *has a regular solution projecting to the origin.*

2.3. Certification

It is possible to rigorously verify that an approximate solution $\widetilde{x}$ is close to a true regular solution x^* of a polynomial system. The technique used here is based on interval arithmetic: based on $\widetilde{x}$, it produces an interval that includes both $\widetilde{x}$ and x^* while excluding any other solutions. It enables the certification of real solutions of real systems; however, albeit nontrivially, the technique extends to the complex case.

For an alternative approach, *Smale's α-theory*, based on exact values of maps and their derivatives at an approximate solution (or certifiable bounds on certain quantities derived from those), refer to Section B.1.

2.3.1. Real interval arithmetic.

Consider the set of compact real intervals,

$$\mathbb{IR} = \{[a, b] \mid a, b \in \mathbb{R}, a \leq b\}.$$

For $A, B \in \mathbb{IR}$ and *implementation* of interval arithmetic for a binary operation $\circ \in \{+, -, \cdot, \div\}$ is an algorithm that ensures

$$A \circ B \supseteq \{a \circ b \,|\, a \in A, b \in B\}$$

if it produces $A \circ B \in \mathbb{IR}$. It also is allowed to fail: for instance, $\div$ should fail when $0 \in B$.

For instance,

$$[a_1, a_2] + [b_1, b_2] := [a_1 + b_1, a_2 + b_2]$$

is a reasonable algorithm for addition, but even this implementation needs to be adjusted in practice if the only intervals one can store are the ones with endpoints at floating point numbers.

Also, note that even simple properties with the simplest implementation may fail.

$$\begin{aligned} [0,1] \cdot ([-1,0] + [1,1]) &= [0,1] \cdot [0,1] = [0,1] \quad \text{but} \\ [0,1] \cdot [-1,0] + [0,1] \cdot [1,1] &= [-1,0] + [0,1] = [-1,1]\,. \end{aligned}$$

This shows that there is no distributive law for interval arithmetic.

2.3.2. Certification of the roots of univariate polynomials. Let $f \in \mathbb{R}[x]$, which could be viewed as $f : \mathbb{R} \to \mathbb{R}$. We call $\Box f : \mathbb{IR} \to \mathbb{IR}$ an *interval enclosure* if for all intervals $A \in \mathbb{IR}$ the interval $\Box f(a)$ contains the set $\{f(a) \mid a \in A\}$.

Note: Given some implementation of interval arithmetic, one can construct an implementation of an interval enclosure for a polynomial function by building an *arithmetic circuit*, i.e. an algorithm for evaluating a polynomial that uses additions and multiplications in a particular fixed order.

Exercise 2.3.1. *Let $f(x) = x^2 - 5x + 6$. Construct two distinct interval enclosures $\Box_1 f$ and $\Box_2 f$ such that $\Box_1 f([2,3]) \subseteq [-5,5]$ and $\Box_2 f([2,3]) \subseteq [-1,-1]$.*

We can define interval enclosures for arbitrary real functions as long as we allow them to be undefined for some intervals in $\mathbb{IR}$. For example, $\Box\frac{1}{x}$ need not be defined for $A \in \mathbb{IR}$ containing 0.

Now, we can talk about *interval Newton operator*, represented by some interval enclosure

$$\Box N_f(A) = A - \frac{\Box f(A)}{\Box f'(A)}.$$

Given an approximation $\widetilde{x}$ such that $\lim_{m\to\infty} N_f^m(\widetilde{x}) = x^*$ is a regular root of f, a natural question arises: Is it possible to construct an interval $\mathbf{I} \ni \widetilde{x}$ such that

$$\bigcap_{m\to\infty} \Box N_f^m(\mathbf{I}) = [x^*, x^*] \in \mathbb{IR},$$

that is $\mathbf{I} \supset \Box N_f(\mathbf{I}) \supset \Box N_f(\Box N_f(\mathbf{I})) \supset \ldots$ is a sequence of nested intervals contracting to a point, which is an exact root?

Fix f and define another operator on $\mathbb{IR}$

$$K_{f,\widetilde{x},y}(\mathbf{I}) = \widetilde{x} - y \cdot \Box f(\widetilde{x}) + (1 - y\Box f'(\mathbf{I}))(\mathbf{I} - \widetilde{x})$$

that depends on $\widetilde{x} \in \mathbb{R}$ (think: an approximate root) and $y \in \mathbb{R}$ (think: an approximation of the value of $1/f'$ at the root).

Exercise 2.3.2. *Show that for $g(x) = x - y\,f(x)$ for every $\mathbf{I} \in \mathbb{IR}$ we have $g(\mathbf{I}) \subset K_{f,\widetilde{x},y}(\mathbf{I})$ for $\widetilde{x} \in \mathbf{I}$.*
(Hint: Noting $K_{f,\widetilde{x},y}(I) = g(\widetilde{x}) + (1 - y\,\Box f'(\mathbf{I}))(\mathbf{I} - \widetilde{x})$, it is enough to show that $g(x) - g(\widetilde{x}) \in (1 - y\,\Box f'(\mathbf{I}))(\mathbf{I} - \widetilde{x})$.)

THEOREM 2.3.3. *Let $K = K_{f,\widetilde{x},y} : \mathbb{IR} \to \mathbb{IR}$ be the operator defined above for f and some choice of $\widetilde{x} \in \mathbb{R}$ and nonzero $y \in \mathbb{R}$.*

Then

(1) *if $K_{\widetilde{x},y}(\mathbf{I}) \subset \mathbf{I}$, there is a root of f in $\mathbf{I}$;*
(2) *if $|1 - y \cdot \Box f'(\mathbf{I})| < 1$, then F has a unique root in $\mathbf{I}$.*

PROOF. The prove of (1) is based on the application of *Brouwer's fixed-point theorem* to $g_(x) = x - y\,f(x)$ Indeed, $g(x) \subset K_{\widetilde{x},y}(\mathbf{I}) \subset \mathbf{I}$ for any $x \in \mathbf{I}$, thus $g(\mathbf{I}) \subset \mathbf{I}$. The above-mentioned theorem concludes the existence of $x^* \in I$ such that $g(x^*) = x^*$, which implies that $f(x^*) = 0$.

We leave it as an exercise to the reader to argue (2). (Hint: Construct a sequence of intervals that decrease in length.) □

2.3.3. Krawczyk's method for multivariate systems. Let F be a system of m polynomials in n variables. We call a map

$$\Box F(\mathbb{IR})^n \to (\mathbb{IR})^m$$

such that $\{F(a) \mid a \in A\} \subseteq \Box F(A)$ for every $A \in (\mathbb{IR})^n$ an *interval enclosure* of F.

Given $\Box F$ and $\Box J$, interval enclosures of F and its Jacobian $J : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, an interval $\mathbf{I} \in (\mathbb{IR})^n$, a point $\widetilde{x} \in \mathbb{R}^n$, and an invertible matrix $Y \in \mathbb{R}^{n\times n}$, define the *Krawczyk operator*

$$K_{F,\widetilde{x},Y}(\mathbf{I}) = \widetilde{x} - Y \cdot \Box F(\widetilde{x}) + (\mathbf{1}_n - Y \cdot \Box \mathrm{J}F(I))(\mathbf{I} - \widetilde{x})\,,$$

where $\mathbf{1}_n$ is the $n \times n$ identity matrix. The norm of a matrix interval $A \in (\mathbb{IR})^{n\times n}$ is

$$\|A\|_\infty = \max_{B\in A} \max_{\widetilde{x}\in\mathbb{R}^n} \|B\widetilde{x}\|_\infty / \|\widetilde{x}\|_\infty,$$

where $\|(x_1, \ldots, x_n)\|_\infty = \max_{1\le i\le n} |x_i|$ for $\widetilde{x} \in \mathbb{R}^n$.

THEOREM 2.3.4. *Let $F = (f_1, \ldots, f_n)$ be a system of n polynomials in n variables, $\mathbf{I} \in (\mathbb{IR})^n$, $\widetilde{x} \in \mathbf{I}$, and let $Y \in \mathbb{R}^{n\times n}$ be invertible.*

(1) *If $K_{F,\widetilde{x},Y}(\mathbf{I}) \subset \mathbf{I}$, there is a solution of F in $\mathbf{I}$.*
(2) *If $\|\mathbf{1}_n - Y \cdot \Box JF(\mathbf{I})\|_\infty < 1$, then F has a unique solution in $\mathbf{I}$.*

CHAPTER 3

Rings, ideals, and Gröbner bases

3.1. Polynomial rings and ideals

The main object of study in this section is a polynomial ring in a finite number of variables $R = k[x_1, \ldots, x_n]$, where k is an arbitrary field.

The abstract concept of a *ring* $(R, +, \cdot)$ assumes that

(1) operations + (addition) and $\cdot$ (multiplication) are defined for pairs of ring elements,
(2) both $(R, +)$ and $(R, \cdot)$ are *abelian groups*, i.e., both addition and multiplication are commutative,
(3) multiplication *distributes* over addition:
$$(a + b)c = ac + bc, \quad a, b, c \in R,$$
(4) there exist an *additive identity*, denoted by 0, and a *multiplicative identity*, denoted by 1, such that
$$1 \cdot a = a,$$
(5) there exists an *additive inverse* $-a$ for every $a \in R$:
$$a + (-a) = 0.$$

The ring of polynomials possesses a natural addition and multiplication satisfying the above ring axioms. Moreover, it enjoys many other "nice" properties: for instance, the multiplication is *cancellative*:
$$fg = fh \implies g = h, \quad f, g, h \in R,\ f \neq 0,$$
which follows from the fact that a polynomial ring is an *integral domain*, i.e., a ring with no *zero divisors*: for $f, g \in R$,
$$fg = 0 \implies f = 0 \text{ or } g = 0.$$

Sometimes a polynomial ring $R = k[x_1, \ldots, x_n]$ is referred to as a polynomial *algebra* (over k) when one needs to emphasize that R is a vector space over the *field of coefficients* k equipped with a bilinear product; note that bilinearity here follows from the distributivity of multiplication in the definition of a ring.

Note: A *field* is a ring where each nonzero element has a multiplicative inverse.
In this text we mostly use fields such as $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ as coefficient fields in polynomial rings. However, one other field closely related to a polynomial ring $R = k[x_1, \ldots, x_n]$ is the *field of rational functions*, denoted by $k(x_1, \ldots, x_n)$, the elements of which are of the form
$$\frac{f}{g}, \text{ where } f, g \in R; \quad \left(\frac{f}{g} = \frac{f'}{g'} \iff fg' = f'g\right).$$
Every nonzero element f/g has $(f/g)^{-1} = g/f$ as its multiplicative inverse.

3.1.1. Ideals. An *ideal* of R is a nonempty k-subspace $I \subseteq R$ closed under multiplication by elements of R:

$$gI = \{\, gf \,|\, f \in I \,\} \subseteq I, \quad g \in R.$$

Two *trivial ideals* of I are the zero ideal $\{0\}$ (denoted by 0) and the whole ring R.

One way to construct an ideal is to *generate* one using a finite set of polynomials. For $f_1, \ldots, f_r \in R$, we define

$$\langle f_1, \ldots, f_r \rangle = \{\, g_1 f_1 + \cdots + g_r f_r \,|\, g_i \in R \,\} \subseteq R,$$

the set of all linear combinations of *generators* f_i with polynomial coefficients g_i. The fact that the set $I = \langle f_1, \ldots, f_r \rangle$ is an ideal follows straightforwardly from the definition.

The set $I = \langle f \rangle = \{\, gf \,|\, g \in R \,\}$ for an element $f \in R$ is called a *principal ideal* and f is called a *principal generator* of I. Note that $R = \langle 1 \rangle$.

Exercise 3.1.1. *A ring, each ideal of which is principal, is called a* principal ideal domain *(PID). Show that the ring of univariate polynomials is a PID.*

We can construct an ideal using an arbitrary (possibly infinite) set of generators $G \subseteq R$:

$$\langle G \rangle = \bigcup_{F \subseteq G, |F| < \infty} \langle F \rangle .$$

However, every ideal $I \subseteq R$ is *finitely generated*, i.e., $I = \langle f_1, \cdots, f_r \rangle$ for some finite number r of polynomials $f_i \in R$ (see Theorem 3.2.10). This is yet another "nice" property of R: a ring with such property is called *Noetherian*.

Exercise 3.1.2. *A ring is said to satisfy the* ascending chain condition *(ACC) if every chain of ideals*

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$$

stabilizes, i.e., there is i_0 such that $I_i = I_{i_0}$ for all $i > i_0$.
For an arbitrary ring, show that this condition is equivalent to the condition of all ideals being finitely generated.

Example 3.1.3. *Consider an ideal $I = \langle x + y, x^2 \rangle \subseteq k[x, y]$. However, we can pick another set of generators of I; for instance, $I = \langle x + y, y^2 \rangle$.*

The polynomials in the second set of generators belong to I as

$$y^2 = \boxed{x^2} + (y - x)\boxed{(x + y)}.$$

This shows the containment $\langle y^2, x + y \rangle \subseteq I$. Since, in a similar way, reverse containment can be shown, the ideals are equal.

Exercise 3.1.4. *Determine whether the following subsets of R are ideals:*

(1) *k, the field of coefficients;*
(2) *a subring $k[x_1, \dots, x_m] \subset R = k[x_1, \dots, x_n]$, where $0 < m < n$;*
(3) *polynomials with no constant term;*
(4) *R_d, homogeneous polynomials of degree d, i.e. polynomials with all terms of degree d.*
(5) *$R_{\leq d}$, polynomials of degree at most d;*
(6) *homogeneous polynomials (of any degree).*

3.1.2. Sum, product, and intersection of ideals. The sum of two ideals I and J (as k-subspaces),

$$I + J = \{ f + g \,|\, f \in I, g \in J \},$$

is an ideal. So is the intersection

$$I \cap J = \{ f \,|\, f \in I, f \in J \}.$$

Exercise 3.1.5. *Prove that $I + J$ is the smallest ideal containing I and J. Show that, if $I = \langle f_1, \dots, f_r \rangle$ and $J = \langle g_1, \dots, g_s \rangle$, then $I + J = \langle f_1, \dots, f_r, g_1, \dots, g_s \rangle$.*

Exercise 3.1.6. *Show that the ideal generated by the products of the elements in I and J,*

$$IJ = \langle\, fg \,|\, f \in I, g \in J \,\rangle,$$

is contained in $I \cap J$. (Exercise 3.1.7 shows that $IJ \neq I \cap J$ in general.)

Exercise 3.1.7. *Consider the univariate polynomial ring $R = k[x]$.*

(1) *How would one find a principal generator of $\langle f \rangle \cap \langle g \rangle$?*
(2) *How would one find a principal generator of $\langle f \rangle \langle g \rangle$?*
(3) *Give an example of f and g where the ideals above (the intersection and the product) are not the same.*

3.1.3. Ring maps and quotient rings. Let R and S be rings, a map $R \to S$ is called a *ring map* if it respects both additive and multiplicative structure of the rings.

Example 3.1.8. *The following ring maps involving polynomial rings are frequently used:*

- *specialization of a variable*

$$(\cdot)|_{x_i = a_i} : k[x_1, \dots, x_n] \to k[x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n], \quad a_i \in k,$$
$$f = f(x_1, \dots, x_n) \mapsto f|_{x_i = a} = f(x_1, \dots, x_{i-1}, a_i, x_{i+1}, \dots, x_n);$$

- *evaluation a point $a = (a_1, \dots, a_n) \in k^n$,*

$$e_a : k[x_1, \dots, x_n] \to k,$$
$$f(x_1, \dots, x_n) \mapsto f(a_1, \dots, a_n);$$

- *variable substitution:*

$$k[x_1, \dots, x_n] \to k[y_1, \dots, y_m],$$
$$f(x_1, \dots, x_n) \mapsto f(g_1(y_1, \dots, y_m), \dots, g_n(y_1, \dots, y_m)),$$

where $g_1, \dots, g_n$ are polynomials in the ring $k[y_1, \dots, y_m]$.

Every polynomial ring map can be defined as the last map in Example 3.1.8, since every ring map is determined by its action on the *ring generators* of the domain, which in case of a polynomial ring are the variables.

A map $\phi : R \to S$ is called an *isomorphism*, if there is a map $\psi : S \to R$ (called the *inverse map* of ϕ) such

$$\psi\phi = \mathrm{id}_R \text{ and } \phi\psi = \mathrm{id}_S,$$

where $\mathrm{id}_R : R \to R$ denotes the identity map on R.

Exercise 3.1.9. *Let $R = k[x_1, \dots, x_n]$. A matrix $A \in k^{(n+1)\times n}$ defines a linear substitution*

$$\begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix} = A \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ 1 \end{bmatrix} \in R^n$$

that can be used to make endomorphism (the source and target of the map coincide) $\phi_A : R \to R$ using the recipe of the last map in Example 3.1.8. If the ring map ϕ_A is an automorphism (endomorphism that is an isomorphism), it is commonly referred to as a linear change of coordinates.

(1) *Find a condition on A for ϕ_A to be an automorphism (endomorphism that is an isomorphism).*
(2) *If ϕ_A is an automorphism, find B such that ϕ_B is its inverse.*

Exercise 3.1.10. *Prove that the kernel of a polynomial ring map, i.e. the set of elements that map to zero, is an ideal.*

Given an ideal $I \subseteq R$ we introduce the *quotient ring* R/I. The elements of R/I are equivalence classes $[f] = \{\, g \in R \mid f - g \in I \,\} \subseteq R$ where $f \in R$. Two elements $f, g \in R$ are equivalent *modulo I* if $[f] = [g]$; that, in turn, holds iff $f - g \in I$.

The ring structure of R/I is induced by that of the ring R:

- $[f] + [g] = [f + g]$;
- $[f][g] = [fg]$;
- $[0]$ is the additive and $[1]$ is the multiplicative identities.

This addition operation is well defined: if $f' \in [f], g' \in [g]$ are alternative representatives, then $[f' + g'] = [f + g]$, since $f' + g' - (f + g) = (f' - f) + (g' - g) \in I$.

Exercise 3.1.11. *Show that the product in a quotient ring is well defined.*

There is a natural surjective ring map

$$\begin{aligned} \phi : R &\to R/I \\ f &\mapsto [f] \end{aligned}$$

Proposition 3.1.12. *Let I be an ideal in an arbitrary ring R. There is a one-to-one correspondence between the ideals of R/I and the ideals of R that contain I. Sums, intersections, and products of ideals are preserved under this correspondence.*

PROOF. We claim that the ring map ϕ above establishes a one-to-one correspondence.

Take an ideal $J \subseteq R$, then $\phi(J)$ is an ideal of J. In fact, this is true for any map ϕ. This follows from the definition of an ideal and the fact that ϕ respects ring addition and multiplication. Similarly, if $\overline{J}$ is an ideal of R/I then $\phi^{-1}(\overline{J})$ is an ideal of R; it contains the preimage of zero $\phi^{-1}([0]) = I$. □

Exercise 3.1.13. *Let $R = k[x_1, \dots, x_n]$ and $I = \langle x_{m+1}, \dots, x_n \rangle$. Show that the rings R/I and $S = k[x_1, \dots, x_m]$ are isomorphic through a natural ring map $\psi : R/I \to S$,*

$$\psi([f]) = f(x_1, \dots, x_m, 0, ..., 0) \in S, \quad f \in R.$$

Exercise 3.1.14. *Consider the ideal $I = \langle x^2 + 1 \rangle \subset \mathbb{Q}[x]$. The quotient ring $\mathbb{Q}[x]/I$ is called (the field of)* Gaussian rational numbers.
Given an element $[f] \in \mathbb{Q}[x]/I$, represented by $f \in \mathbb{Q}[x]$, show how to construct its inverse. (This proves that Gaussian rational numbers, indeed, form a field.)

3.2. Gröbner bases

It has been pointed out (e.g., in Example 3.1.3) that the same nonzero ideal can be generated by different sets of generators. In this section, we develop a theory and algorithms to convert any generating sets into a *Gröbner basis*, a generating set with helpful special properties.

3.2.1. Monomial orders. A *monomial order* is a recipe for comparing two monomials in a polynomial ring $R = k[x_1, \dots, x_n]$ with the following properties:

(1) It is a *total order*: for every pair of distinct monomials x^α and x^β, $\alpha, \beta \in \mathbb{N}^n$,

$$\text{either } x^\alpha > x^\beta \text{ or } x^\alpha < x^\beta.$$

(2) It is a *multiplicative order*:

$$x^\alpha > x^\beta \implies x^{\alpha+\gamma} = x^\alpha x^\gamma > x^\beta x^\gamma = x^{\beta+\gamma}, \quad \alpha, \beta, \gamma \in \mathbb{N}^n.$$

(3) It is a *well-order*: every nonempty set (of monomials) has a minimal element. Together with being a total order, this implies that

$$x^0 = 1 < x^\alpha, \quad \alpha \in \mathbb{N}^n - \{0\}.$$

Exercise 3.2.1. *Show that there is only one monomial order for monomials of a univariate polynomial ring.*

Example 3.2.2. *A* lexicographic order *on $k[a, b, c, \dots, z]$ compares monomials as words in a dictionary:*

$$a^3b^2c = aaabbc > aabbbcccc = a^2b^3c^4$$

as "aaabbc" comes before "aabbbcccc" in the dictionary.

This can be used with any alphabet: for $k[x_1, \dots, x_n]$, we have

$$x^\alpha >_{\text{lex}} x^\beta \iff \alpha_1 > \beta_1 \text{ or } (\alpha_1 = \beta_1 \text{ and } x^{(0,\alpha_2,\dots,\alpha_n)} >_{\text{lex}} x^{(0,\beta_2,\dots,\beta_n)}).$$

One important class of monomial orders is *graded monomial orders*, the ones that refine the (non-total) order by degree.

Example 3.2.3. *The* graded lexicographic order *compares the degrees of monomials first and "breaks the tie", if necessary, using the lexicographic order:*

$$x^\alpha >_{\text{glex}} x^\beta \iff |\alpha| > |\beta| \text{ or } (|\alpha| = |\beta| \text{ and } x^\alpha >_{\text{lex}} x^\beta).$$

Note: The default monomial order used by many computer algebra systems is *graded reverse lexicographic order*.

Exercise 3.2.4. *For a polynomial* $f = x^3y + 2x^2y^2 + y^3 + x + y^2 + y + 1$ *find* $\text{LM}(F)$*, where*

(1) $>=>_{\text{lex}}$, $x > y$;
(2) $>=>_{\text{lex}}$, $y > x$;
(3) $>=>_{\text{glex}}$, $x > y$;
(4) $>=>_{\text{glex}}$, $y > x$.

Another useful class of monomial orders are *block orders* that compare monomials according to a fixed partition of the sets of variables into blocks.

Let $>_1$ be an order on the on monomials in $x_1, \dots, x_m$ and $>_2$ be and order on monomials in $x_{m+1}, \dots, x_n$. The *2-block order* $>_{2,1}$ on monomials in $x_1, \dots, x_n$ is

$$x^\alpha >_{1,2} x^\beta \iff x_{m+1}^{\alpha_{m+1}} \cdots x_n^{\alpha_n} >_2 x_{m+1}^{\beta_{m+1}} \cdots x_n^{\beta_n} \text{ or}$$
$$(x_{m+1}^{\alpha_{m+1}} \cdots x_n^{\alpha_n} = x_{m+1}^{\beta_{m+1}} \cdots x_n^{\beta_n} \text{ and } x_1^{\alpha_1} \cdots x_m^{\alpha_m} >_1 x_1^{\beta_1} \cdots x_m^{\beta_m}).$$

Note that $>_{\text{lex}}$ is a 2-block order with respect to the blocks $\{x_1, \dots, x_m\}$ and $\{x_{m+1}, \dots, x_n\}$.

3.2.2. Normal form algorithm. In §1.1.4 we have introduced NF_f the normal form function that maps a polynomial $g \in k[x]$ to its remainder after division by the polynomial $f \in k[x]$. We would like to define the *normal form* $\text{NF}_F : R \to R$, where $R = k[x_1, \dots, x_n]$, with respect to a system of polynomials $F \in R^r$.

Algorithm 3.2.1 $h = \text{NF}(g, F)$

Require: $g \in R$;
$F \in R^r$, $r > 0$;
Ensure: $h \in R$, such that

$$g = h + \sum_{i=1}^{r} q_i f_i, \quad q_i \in R, \ \deg q_i + \deg f_i \le \deg g \tag{3.2.1}$$

and either $h = 0$ or $\text{LM}(h)$ is not divisible by $\text{LM}(f)$ for all $f \in F$.

$h \leftarrow g$
while $h \neq 0$ and $\text{LM}(h)$ is divisible by $\text{LM}(f)$ for some $f \in F$ **do**
$f \leftarrow$ first polynomial in the set F such that $\text{LM}(f) \,|\text{LM}(h)$

$$h \leftarrow h - \frac{\text{LT}(h)}{\text{LT}(f)} f$$

end while

The leading monomials and leading terms in Algorithm 3.2.1 are taken with respect to a fixed monomial order $>$. If this needs to be emphasized, we write $\mathrm{NF}_F^{(>)}$; normal forms for the same input, but different monomial orders are not the same, in general.

PROOF OF TERMINATION AND CORRECTNESS OF ALGORITHM 3.2.1. Let h_i be the contents of h at the i-th iteration. Then

$$\mathrm{LM}(h_1) > \mathrm{LM}(h_2) > \mathrm{LM}(h_3) > \cdots$$

Since a monomial order is a well-order, the descending sequence of monomials terminates, so does the algorithm. The condition (3.2.1) holds for all $h = h_i$ by construction. When the algorithm terminates h is either 0 or $\mathrm{LM}(h)$ is not divisible by $\mathrm{LM}(f)$ for all $f \in F$. □

Exercise 3.2.5. *Let $f_1, \dots, f_r \in I$, where $I \subseteq R$ is an ideal.*
Show that $\mathrm{NF}_{(f_1,\dots,f_r)}(g) \in I$ iff $g \in I$.

Note: As its univariate analoque, Algorithm 3.2.1 can be modified to compute not only the "remainder", but also the "quotients", i.e., polynomial coefficients $q_i \in R$ in (3.2.1).

Note that, in general, the normal form also depends on the order of polynomials in the system.

Example 3.2.6. *Consider two polynomials in $k[x, y, z]$,*

$$f_1 = x - y,$$
$$f_2 = x - z^2.$$

Fix the monomial order $>=>_{\mathrm{lex}}$, $x > y > z$.
Then $\mathrm{NF}_{(f_1,f_2)}(x) = y$ and $\mathrm{NF}_{(f_2,f_1)}(x) = z^2$.

Exercise 3.2.7. *For $f_1 = x^3 + y^2$, $f_2 = xy + 1$, and*

$$g = x^3y + 2x^2y^2 + xy^3 + x + y^2 + y + 1,$$

polynomials in $k[x, y]$ with the lexicographic order such that $x > y$, find

(1) $\mathrm{NF}_{(f_1,f_2)}(g)$
(2) $\mathrm{NF}_{(f_2,f_1)}(g)$

3.2.3. Initial ideal, Dickson's Lemma, Noetherianity. For a polynomial ideal $I \subset R$, the ideal generated by the leading monomials of all polynomials of I is called the *initial ideal* and denoted

$$\mathrm{in}(I) = \langle\, \mathrm{LM}(f) \mid f \in I \,\rangle .$$

Again, if we need to emphasize the (usually fixed) monomial order $>$ that is used, we would write $\mathrm{in}_>(I)$.

Exercise 3.2.8. *For the ideal $I = \langle x - y, x - z^2 \rangle \subset k[x, y, z]$ find*

(1) *the initial ideal $\mathrm{in}_{>_{\mathrm{lex}}}(I)$ with respect to the lexicographic ordering;*
(2) *the initial ideal $\mathrm{in}_{>_{\mathrm{glex}}}(I)$ with respect to the graded lexicographic ordering.*

We need the following lemma to show that every ideal I of a polynomial ring R can be finitely generated; this is one of the ways to say that R is *Noetherian*. (We refered to this fact in §3.1.1 without a proof.)

Lemma 3.2.9 (Dickson's Lemma)**.** *Every* monomial ideal *(i.e., ideal generated by monomials) is finitely generated.*

THEOREM 3.2.10. *A polynomial ring R is Noetherian.*

PROOF. Let $I \subseteq R$ be a nonzero ideal of R, then, by Dickson's Lemma, its initial ideal is finitely generated:

$$\mathrm{in}(I) = \langle m_1, \dots, m_r \rangle, \quad r > 0.$$

Pick $f_i \in I$ such that $\mathrm{LM}(f_i) = m_i$ and let

$$J = \langle f_1, \dots, f_r \rangle, \quad J \subseteq I.$$

Take $g \in I$ and compute $h = \mathrm{NF}_{(f_1,\dots,f_r)}(g)$. On one hand, by Exercise 3.2.5, $h \in I$. On the other, if $h \neq 0$, then $\mathrm{LM}(h) \notin \mathrm{in}(I)$ as it is not divisible by monomials m_i, which leads to a contradiction. Therefore, $h = 0$ and $g \in J$; we conclude that $J = I$. □

PROOF OF DICKSON'S LEMMA. Let G be a (possibly infinite) set monomials generating the ideal $J = \langle G \rangle$. Without a loss of generality we may assume G consists of minimal elements with respect to divisibility: if two monomials $x^\alpha, x^\beta \in G$ are such that x^α divides x^β, then the latter can be excluded from G.

First, we can see a monomial ideal $J \subseteq k[x_1, \dots, x_n]$ generated as follows

$$J = \left\langle J_0 \cup x_1 J_1 \cup x_1^2 J_2 \cup \cdots \right\rangle,$$

where $J_i \subseteq k[x_2, \dots, x_n]$ are monomial ideals (in a ring with one fewer variable) such that

$$\left\{ x_1^i x^{\beta_2 \cdots \beta_n} \,\middle|\, x^{\beta_2 \cdots \beta_n} \in \mathrm{in}(J_i) \right\} = \left\{ x^{\alpha_1 \alpha_2 \cdots \alpha_n} \in \mathrm{in}(J) \,|\, \alpha_1 = i \right\}.$$

Using induction on the number of variables in a polynomial ring, we may assume that $k[x_2, \dots, x_n]$ is Noetherian. The base of induction is the case $R = k$, a polynomial ring with no variables, which has only trivial ideals.

Observe that $J_1 \subseteq J_2 \subseteq \cdots$ is an ascending chain of ideals. By Noetherianity it stabilizes; we also may pick finite generating sets of monomials G_i for J_i.

Now the infinite union above becomes finite: for some $s > 0$,

$$\begin{aligned} J &= \left\langle J_0 \cup x_1 J_1 \cup x_1^2 J_2 \cup \cdots \cup x_1^s J_s \right\rangle \\ &= \left\langle J_0 \cup x_1 G_1 \cup x_1^2 G_2 \cup \cdots \cup x_1^s G_s \right\rangle, \end{aligned}$$

which shows that J is generated by a finite number of monomials. □

3.2.4. Gröbner bases and their properties. Fix a polynomial ring R and a monomial order.

A set $G \subseteq R$ is a *Gröbner basis* of an ideal $I \subseteq R$ if

- $I = \langle G \rangle$, and
- $\mathrm{in}(I) = \langle \mathrm{in}(G) \rangle$, where $\mathrm{in}(G) = \{ \mathrm{in}(g) \,|\, g \in G \}$.

Example 3.2.11. *The set $G = \{x - y, x - z^2\} \subseteq k[x, y, z]$ is*

- *not a Gröbner basis of $I = \langle G \rangle$ with respect to $>_{\mathrm{lex}(x,y,z)}$, since $\mathrm{in}(I) \ni y = \mathrm{in}(y - z^2)$, however $\mathrm{in}(G) = \langle x \rangle \not\ni y$;*

- *a Gröbner basis of* $I = \langle G \rangle$ *with respect to* $>_{\mathrm{lex}(z,y,x)}$*: one can show that* $\mathrm{in}_{\mathrm{lex}(z,y,x)}(I) = \langle y, z^2 \rangle$.

Proposition 3.2.12. *Let* G *be a Gröbner basis of an ideal* I *and consider a polynomial* $f \in R$.

(1) $\mathrm{NF}_G(f) = 0 \iff f \in I$.

Proof. Let $h = \mathrm{NF}_G(f)$; note that $h \in I \iff f \in I$, by Exercise 3.2.5. However, either $h = 0$ or $\mathrm{LM}(h) \notin \mathrm{in}(I)$, since the leading monomials of elements in G generate $\mathrm{in}(I)$.The conclusion is that $h \in I \iff h = 0$. □

Given a fixed monomial order, define the *normal form* $\mathrm{NF}_I(f)$ of $f \in R$ with respect to an ideal I to be the output of Algorithm 3.2.2.

Algorithm 3.2.2 $h = \mathrm{NF}(f, I)$

Require: $f \in R = k[x_1, \ldots, x_n]$ with a fixed monomial order;
$I \subseteq R$, an ideal (given by a finite set of generators);
Ensure: $h \in R$, such that $h \equiv f \pmod{I}$ and all monomials of h are not in $\mathrm{in}(I)$.

```
G ← a Gröbner basis of I
h ← 0
t ← f                          - This is the "tail" that we reduce.
while t ≠ 0 and LM(t) is divisible by LM(g) for some g ∈ G do
  t ← NF_G(t)
  if h ≠ 0 then
    h ← h + LT(t)
    t ← t − LT(t)
  end if
end while
```

Corollary 3.2.13 (of Proposition 3.2.12)**.** *A polynomial* $f \in R$ *belongs to an ideal* $I \subseteq R$ *iff* $\mathrm{NF}_I(f) = 0$.

Proposition 3.2.14. *There is a unique* $h \in R$*, such that* $h \equiv f \pmod{I}$ *and all monomials of* h *are not in* $\mathrm{in}(I)$.

Proof. Suppose two distinct $h', h \in R$ satisfy the hypotheses. On one hand, $h - h' = (h - f) - (h' - f) \in I$; on the other, monomials of $h - h'$ do not belong to $\mathrm{in}(I)$, hence, $h - h' = \mathrm{NF}_I(h - h')$. We conclude that $h - h' = 0$ by Corollary 3.2.13. □

Corollary 3.2.15. *For any polynomial* $f \in R$ *and any ideal* $I \subseteq R$*, the normal form* $\mathrm{NF}_I(f)$ *does not depend*

- *neither on the choice of the Gröbner basis* G *in Algorithm 3.2.2*
- *nor on the order of reductions in Algorithm 3.2.1.*

A Gröbner basis G of an ideal I is called *reduced* if

- $\mathrm{LC}(g) = 1$ for all $g \in G$ (g is monic),
- $\mathrm{LM}(g)$, $g \in G$, are distinct,
- $\mathrm{NF}_I(g - \mathrm{LM}(g)) = g - \mathrm{LM}(g)$ (no other monomials in $\mathrm{in}(I)$).

Exercise 3.2.16. *Show that (provided a fixed monomial order) the reduced Gröbner basis is unique for any ideal.*

Exercise 3.2.17. *Fix the monomial order* $>_{\text{glex}}$. *Knowing that*

$$G = \left\{2x^2 - 2y^2,\ y^3 - 2xy - y^2 + 2x,\ xy^2 + y^3 - 5xy - y^2 + 4x\right\}$$

is a Gröbner basis of the ideal $I = \langle G\rangle$, *find the reduced Gröbner basis of* I.

3.2.5. Buchberger's algorithm. Now we are ready to provide the missing piece of Algorithm 3.2.2 is a subroutine that would compute a Gröbner basis for an ideal generated by a finite set of polynomials.

For two nonzero polynomials $f, g \in R$. Define the *s-polynomial* of f and g

$$S_{f,g} = \frac{\mathrm{LT}(g)}{\gcd(\mathrm{LM}(f), \mathrm{LM}(g))} f - \frac{\mathrm{LT}(f)}{\gcd(\mathrm{LM}(f), \mathrm{LM}(g))} g \in R.$$

THEOREM 3.2.18 (Buchberger's criterion). *Let* $G \subseteq R$ *be a finite set of polynomials, then* G *is a Gröbner basis of the ideal* $I = \langle G\rangle$ *(with respect to a fixed monomial order) iff* $\mathrm{NF}_G(S_{f,g}) = 0$ *for all* $f, g \in G$.

PROOF. If G is a Gröbner basis, then $S_{f,g} \in I$ implies $\mathrm{NF}_G(S_{f,g}) = 0$ by Proposition 3.2.12. To prove the statement in the other direction, we will show that, when every s-polynomial reduces to zero, every element $f \in I$ also reduces to zero with respect to G. This is sufficient, since it implies $\mathrm{in}(I) = \langle \mathrm{in}(G)\rangle$.

Let $G = \{g_1, \ldots, g_r\}$. If $f = \sum_{i=1}^{r} h_i g_i$ for $h_i \in R$, we shall call the sequence $h = (h_1, \ldots, h_r)$ a *representation* of $f \in I$. Define the *leading monomial* λ of a representation to be

$$\lambda = \lambda(h) = \max_i \mathrm{LM}(h_i g_i)$$

and the *multiplicity* μ of the representation to be the number of times the equality $\mathrm{LM}(h_i g_i) = \lambda(h_1, \ldots, h_r)$ holds for $i = 1, \ldots, r$.

Let $f = \mathrm{NF}_G(f)$ be a (reduced) polynomial in I and suppose it is nonzero. Suppose $(h_1, \ldots, h_r)$ is a representation of f with the smallest possible leading monomial λ and multiplicity μ.

If $\mu = 1$, then $\mathrm{LM}(f) = \mathrm{LM}(h_i g_i)$ for some i, which contradicts our assumption (that f is reduced).

For $\mu > 1$, take $1 \le i < j \le r$ such that $\mathrm{LM}(h_i g_i) = \mathrm{LM}(h_j g_j)$. This means that for the monomial $m = \lambda / \mathrm{lcm}(\mathrm{LM}(g_i), \mathrm{LM}(g_j))$ and some $c \in k$,

$$\mathrm{LT}(h_i)\, g_i = c\, m\, \mathrm{lcm}(\mathrm{LM}(g_i), \mathrm{LM}(g_j)).$$

Since $\mathrm{NF}_G(S_{g_i,g_j}) = 0$, there are $\hat{h}_i$ such that

$$S_{g_i,g_j} = \sum_{i=1}^{r} \hat{h}_i g_i \quad \text{and} \quad \mathrm{LM}\left(\hat{h}_i g_i\right) < \mathrm{lcm}(\mathrm{LM}(g_i), \mathrm{LM}(g_j)).$$

Consider a representation h' of f obtained by adding a representation of 0 corresponding to the above:

$$\begin{aligned} h'_l &= h_l + c\, m\, \hat{h}_l, && \text{if } l \notin \{i, j\}, \\ h'_i &= h_i - c\, m \left(\tfrac{\mathrm{LT}(g_j)}{m'} - \hat{h}_i\right), \\ h'_j &= h_j + c\, m \left(\tfrac{\mathrm{LT}(g_i)}{m'} + \hat{h}_j\right), \end{aligned}$$

where $m' = \gcd(\mathrm{LM}(g_i), \mathrm{LM}(g_j))$. One can check that this representation has either $\lambda(h') < \lambda(h)$ (this happens if $\mu(h) = 2$) or $\lambda(h') = \lambda(h)$ but $\mu(h') < \mu(h)$. This contradicts the minimality of the representation h. Hence, $\mathrm{NF}_G(f) = 0$ for every $f \in I$. □

The criterion translates into *Buchberger's algorithm* for finding a Gröbner basis (Algorithm 3.2.3).

Algorithm 3.2.3 $G = \textsc{Buchberger}(I)$

Require: $I = \langle F \rangle \subseteq R$, an ideal given by a finite set of generators F;
Ensure: $G \subseteq R$, a Gröbner basis of I (with respect to a fixed monomial order).

```
G ← F
S ← G × G                    - The queue of s-pairs.
while S ≠ ∅ do
  Pick (f1, f2) ∈ S.
  S ← S − {(f1, f2)}
  g ← NF_G(S_{f1,f2})
  if g ≠ 0 then
    S ← S ∪ ({g} × G)
    G ← G ∪ {g}
  end if
end while
```

Proof of termination and correctness of Algorithm 3.2.3. Let G_i be an intermediate set of generators at step i of the algorithm. The sequence

$$G_1 \subseteq G_2 \subseteq \cdots$$

has a property that either $G_{i+1} = G_i$ or $\mathrm{LM}(G_i) \subsetneq \mathrm{LM}(G_{i+1})$, which which mirrors in the sequence

$$\langle \mathrm{LM}(G_1) \rangle \subseteq \langle \mathrm{LM}(G_2) \rangle \subseteq \cdots$$

Since the latter sequence has to stabilize due to Noetherianity of the polynomial ring, the former one stabilizes too. This means that no new elements are appended to the set $G = G_{final}$ after some step and the algorithm runs through the remaining s-pairs reducing each of them to zero and stops.

The s-polynomials of s-pairs that resulted in a new element $g \in G$ reduce to zero, since $g \in G_{final}$. Therefore, every s-pair considered during the run reduces to zero and the algorithm goes through all pairs $G_{final} \times G_{final}$ by construction. □

3.3. Basic computations in polynomial rings

Here we discuss basic computations in polynomial rings that Gröbner bases enable.

Proposition 3.2.12 already provides us with a way to test if a polynomial belongs to an ideal: the so-called *ideal membership* test.

3.3.1. Computations in a quotient ring. Given an ideal $I \subseteq R$ consider the quotient ring R/I. Proposition 3.2.14 and Corollary 3.2.15 give a way to pick a canonical representative for $[f] \in R/I$: take the normal form of the representative $f \in R$:

$$[\mathrm{NF}_I(R)] = [f].$$

Note that representation with normal forms gives a one-to-one correspondence between polynomials involving only *standard monomials* (i.e., monomials outside $\mathrm{in}(I)$) and R/I.

Example 3.3.1. *The set*

$$G = \left\{ \boxed{x^2} - y^2, \boxed{y^3} - 2xy - y^2 + 2x, \boxed{xy^2} - 3xy + 2x \right\}$$

is a Gröbner basis of $I = \langle G \rangle$ with respect to $>_{\mathrm{glex}}$. The $S = \{1, x, y, xy, y^2\}$ is the set of standard monomials.

Therefore, as a k-space, R/I is finite-dimensional. (This is equivalent to saying that ideal I and the system of polynomials G are 0-dimensional *in the ring-theoretic sense.)*

We used this fact in Section 1.2.1 to construct the multiplication map

$$M_f : R/I \to R/I, \quad [g] \mapsto [fg]$$

and applied it to solving the polynomial system G via eigenvalues of operators M_f where f is set equal to one of the variables.

3.3.2. Elimination. Another fundamental problem is that of *elimination*: given and ideal $I \subset k[x, y] = k[x_1, \ldots, x_n, y_1, \ldots, y_m]$ find $J = I \cap k[x]$ (an ideal of $k[x]$), i.e., eliminate y_i.

Fix a block order $>_{2,1}$ (see §3.2.1) constructed from some monomial orders $>_1$ on $k[x]$ and $>_2$ on $k[y]$. We say that such order *eliminates* the variables y_i and sometimes write $y_i \gg x_j$ for all i, j.

One can show that if G is a Gröbner basis of I with respect to $>_{2,1}$, then $G \cap k[x]$ is not only a generating set, but also a Gröbner basis of J with respect to $>_1$.

Example 3.3.2. *Fix the elimination order with $y \gg x$ on $R = k[x, y]$ and consider the ideal I of Example 3.3.1. The set*

$$G = \left\{ x^4 - 2x^3 - x^2 + 2x,\ 3yx - x^3 - 2x,\ y^2 - x^2 \right\}$$

is a Gröbner basis of I with respect to this order.

Therefore, $J = I \cap k[x] = \langle x^4 - 2x^3 - x^2 + 2x \rangle$. Now solving the univariate equation and substituting the values of x in the other equations gives a solving method that was also discussed in Chapter 1.

An ideal $I \subset R = k[x_1, \ldots, x_n]$ is said to be in *shape position* if there exist univariate polynomials $f_1, \ldots, f_n \in k[x_n]$ such that

$$I = \langle x_1 - f_1, \ldots, x_{n-1} - f_{n-1}, f_n \rangle . \tag{3.3.1}$$

Exercise 3.3.3. *Let I be an ideal of R. Prove that for any monomial order that eliminates $x_1, \ldots, x_{n-1}$, the reduced Gröbner basis for I is of the form in Equation (3.3.1) iff I is in the shape position.*

Note that having I in shape position and knowing f_i, $i = 1, \ldots, n$, above reduces the problem of finding the variety $\mathbb{V}(I)$ to finding roots of the univariate polynomial f_n. Once values of the n-th coordinate are obtained, f_i, $i = 1, \ldots, n-1$, determine the rest of the coordinates.

This technique can be applied to a 0-dimensional ideal that is not in shape position.

Exercise 3.3.4. *Let* $I = \langle x_1^2 - 1, x_2^2 - 4 \rangle \subset \mathbb{R}[x_1, x_2]$.

Show that I *is not in shape position.*
Find a linear change of coordinates (an invertible map $\phi : R \to R$*) of the form*

$$\phi(x_1) = x_1 + ax_2$$
$$\phi(x_2) = ax_1 + x_2$$

such that $\phi(I)$ *is in shape position.*
Describe the set of all possible $a \in \mathbb{R}$ *above that result in* $\phi(I)$ *in shape position.*

Exercise 3.3.5. *Consider an ideal* $I \subset \mathbb{C}[x_1, \ldots, x_n]$ *such that* $\mathbb{V}(I)$ *is a collection of points. Show that for a generic choice of* $(a_1, \ldots, a_n) \in \mathbb{C}^n$,

$$I + \langle x_{n+1} - (a_1 x_1 + \cdots + a_n x_n) \rangle \subset k[x_1, \ldots, x_{n+1}]$$

is in shape position.

CHAPTER 4

Algebra-geometry correspondence

4.1. Ideal-variety correspondence

The correspondence between algebra and geometry about to be discussed is the core of the area called *algebraic geometry*, which uses geometric intuition on one hand and algebraic formalism on the other. Computations in polynomial rings is what drives the effective methods in algebraic geometry.

In this section we will consider the polynomial ring $R = k[x] = k[x_1, \dots, x_n]$ over an *algebraically closed* field k, i.e., a field such that every univariate polynomial with coefficients in k has its roots in k. From the list of popular fields that we considered $(\mathbb{Q}, \mathbb{R}, \mathbb{C})$ only the field of complex numbers is algebraically closed; this follows from the *fundamental theorem of algebra*. Note that, e.g., polynomial $x^2 + 1$ has coefficients in $\mathbb{Q} \subset \mathbb{R}$, but its roots are not real. Therefore, neither $\mathbb{Q}$ nor $\mathbb{R}$ are not algebraically closed.

In the rest of the text we assume $k = \mathbb{C}$ and will describe the classical (complex) algebraic geometry, an area where the main object of study is a *variety* (as we defined above) sometimes referred to by the full name of *complex affine variety*.

We have already introduces the concept of a *variety* $\mathbb{V}(F)$ given by a system of polynomials $F \subset R$ (see §1.2). Here we would like to regard the operation $\mathbb{V}(\dots)$ as a map

$$\begin{array}{rccl} \mathbb{V}: & \{\text{ideals}\} & \to & \{\text{varieties}\} \\ & I & \mapsto & \mathbb{V}(I) = \{\, x \in k^n \mid f(x) = 0, \text{ for all } f \in I \,\} \end{array}$$

Exercise 1.2.2 implies that $\mathbb{V}(I) = \mathbb{V}(F)$ for any generating set F of the ideal I.

In the opposite direction we have a map

$$\begin{array}{rccl} \mathbb{I}: & \{\text{varieties}\} & \to & \{\text{ideals}\} \\ & V & \mapsto & \mathbb{I}(V) = \{\, f \in R \mid f(x) = 0, \text{ for all } x \in V \,\} \end{array}$$

We refer to $\mathbb{I}(V)$ as *the ideal* of V. (Fuller names are *vanishing ideal* and *defining ideal*.)

Exercise 4.1.1. *Show that for an arbitrary set $V \subset k^n$ (not necessarily a variety) $I(V)$ is an ideal.*

Exercise 4.1.2. *Show that both $\mathbb{V}$ and $\mathbb{I}$ are inclusion-reversing, i.e,*

- $I \subseteq J \implies \mathbb{V}(J) \subseteq \mathbb{V}(I)$,
- $V \subseteq W \implies \mathbb{I}(W) \subseteq \mathbb{I}(V)$.

A variety V is called a *hypersurface* if $\mathbb{I}(V)$ is a proper principal ideal, i.e., it is defined by one nonconstant polynomial.

Example 4.1.3. $\mathbb{V}\big(x^2+y^2+z^2-1\big)$ *is a hypersurface in* k^3 *with coordinates* x, y, z *and* $\mathbb{V}\big(x^2+y^2-1\big)$ *is a hypersurface in* k^2 *with coordinates* x, y.

If I is an ideal and $f \in R$, the variety

$$\mathbb{V}(I + \langle f \rangle) = \{\, x \in \mathbb{V}(I) \mid f(x) = 0 \,\}$$

is simply the intersection of the variety $\mathbb{V}(I)$ with the hypersurface $\mathbb{V}(f)$.

Exercise 4.1.4. *The variety* $V = \mathbb{V}\big(y - x^2, z - x^3\big) \subset k^3$ *is called the* twisted cubic. *Find the ideal of the projection of* V *onto*

(1) *the* xy*-plane;*
(2) *the* xz*-plane;*
(3) *the* yz*-plane. (Hint: Look at implicitization procedure in §6.2.1.)*

4.1.1. Hilbert's Nullstellensatz. "Nullstellensatz" translates as "theorem about zeros of functions" from German. We shall state most results for $k = \mathbb{C}$ but they hold more generally for any field k that is *algebraically closed*. (Note: e.g. $\mathbb{R}$ is *not* algebraically closed.)

For I, an ideal of $R_n = \mathbb{C}[x_1, \ldots, x_n]$, and $a \in \mathbb{C}$ define

$$I_{x_n=a} = \{f(x_1, \ldots, x_{n-1}, a) \mid f \in I\} \subset R_{n-1} = \mathbb{C}[x_1, \ldots, x_{n-1}].$$

Note that this is an ideal of R_{n-1} that can be obtained as $(I + \langle x_n - a \rangle) \cap R_{n-1}$.

Exercise 4.1.5. *If* $I \subset R_n$ *is a proper ideal (i.e.* $I \neq R_n$*), then there is* $a \in \mathbb{C}$ *such that* $I_{x_n=a} \neq R_{n-1}$.

THEOREM 4.1.6 (Weak Nullstellensatz). *If* $I \subseteq R_n$ *is an ideal with* $\mathbb{V}(I) = \emptyset$*, then* $I = R_n$.

PROOF. Using Theorem 4.1.5 and induction, we can see that for a proper ideal I there exist $a_1, \ldots, a_n$ such that $I_{x_1=a_1,\ldots,x_n=a_n} \subsetneq R_0 = \mathbb{C}$. However, the only proper ideal of any field is 0. This implies that the point $(a_1, \ldots, a_n) \in \mathbb{V}(I)$ and results in a contradiction. Thus, the ideal I is not proper. □

THEOREM 4.1.7 (Hilbert's Nullstellensatz). *Let* $I \subseteq R$ *be an ideal and* $f \in R$ *be a polynomial vanishing at every point of the variety* $\mathbb{V}(I)$.

Then there exists $m > 0$ *such that* $f^m \in I$.

PROOF. See the proof here[1]. □

Hilbert's Nullstellensatz is stronger than the weak Nullstellensatz: indeed, any function vanishes at every point of an empty set, hence, some power of $f = 1$ belongs to I if $\mathbb{V}(I) = \emptyset$.

Exercise 4.1.8. *Let* $V \subseteq k^n$ *be a variety and* $I \subseteq R$ *be an ideal. Show that*

(1) $\mathbb{V}(\mathbb{I}(V)) = V$;
(2) $I \subseteq \mathbb{I}(\mathbb{V}(I))$ *(In general,* $I \neq \mathbb{I}(\mathbb{V}(I))$*; e.g., see Exercise 4.1.9.)*

[1] David A. Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms*. Fourth. Undergraduate Texts in Mathematics. An introduction to computational algebraic geometry and commutative algebra. Springer, Cham, 2015, pp. xvi+646, Theorem 6, p.183.

Exercise 4.1.9. *Draw the points of the real varieties* $\mathbb{V}\big(x^2+y^2-1\big)$, $\mathbb{V}(x-1)$, *and* $V = \mathbb{V}(I) \subset \mathbb{R}^2$, *where* $I = \langle x^2+y^2-1, x-1\rangle \subset \mathbb{R}[x,y]$.
Show that $\mathbb{I}(V) \neq I$.

4.1.2. Radical ideals. The *radical* of an ideal $I \subseteq R$ is

$$\sqrt{I} = \{\, f \in R \,|\, f^m \in I \text{ for some } m \,\}.$$

An ideal $I \subseteq R$ is a *radical ideal* if $\sqrt{I} = I$.

Proposition 4.1.10. *If* I *is an ideal, then the set* $\sqrt{I}$ *is an ideal.*

PROOF. Take $f, g \in \sqrt{I}$. Then there exist a, b such that $f^a, g^b \in I$.
Each term in the binomial expansion

$$(f+g)^{a+b} = \sum_{i=0}^{a+b} \binom{a+b}{i} f^i g^{a+b-i}$$

has a factor of either f^a or g^b. Therefore, $(f+g)^{a+b} \in I$ and $f+g \in \sqrt{I}$.
Also, for all $h \in R$, the multiple $hf \in \sqrt{I}$, since $(hf)^a \in \sqrt{I}$. □

Two immediate corollaries follow.

Corollary 4.1.11. *If* I *is a proper ideal, then* $\sqrt{I}$ *is.*

PROOF. The element $1 \notin \sqrt{I}$, since all powers of 1 are not in I. □

Corollary 4.1.12. *Nontrivial ideals that are* maximal *(with respect to inclusion) are radical.*

PROOF. Let $\mathfrak{m} \subsetneq R$ be a maximal ideal. Then $\mathfrak{m} \subseteq \sqrt{\mathfrak{m}} \subsetneq R$, which forces $\mathfrak{m} = \sqrt{\mathfrak{m}}$. □

Exercise 4.1.13. *Show that for every ideal* I,
(1) $\mathbb{V}\Big(\sqrt{I}\Big) = \mathbb{V}(I)$,
(2) $\sqrt{\sqrt{I}} = \sqrt{I}$.

Exercise 4.1.14. *For a point* $a \in k^n$, *show that the ideal*

$$\mathfrak{m}_a = \mathbb{I}(\{a\}) = \langle x_1 - a_1, \ldots, x_n - a_n \rangle$$

is maximal.
Prove that every maximal ideal (in the polynomial ring $R = \mathbb{C}[x]$*) has this form.*

Exercise 4.1.15. *Show that for every* $V \subseteq k^n$ *the ideal* $\mathbb{I}(V)$ *is radical.*

In fact, restricted to radical ideals, the maps $\mathbb{I}$ and $\mathbb{V}$ establish a one-to-one correspondence

$$\{\text{varieties}\} \leftrightarrow \{\text{radical ideals}\}.$$

This is the main conceptual consequence of Hilbert's Nullstellensatz (Theorem 4.1.7).

Exercise 4.1.16. *Show that if the ideal I is radical and $f \notin I$ then $\mathbb{V}(I + \langle f \rangle)$, the intersection of $V = \mathbb{V}(I)$ with the hypersurface $\mathbb{V}(f)$, is strictly smaller than V.*

Note: One can design a membership test to determine if a polynomial f belongs to the radical of an ideal $I = \langle g_1, \dots, g_r \rangle \subset R = k[x_1, \dots, x_n]$.
Let $J = \langle g_1, \dots, g_r, tf - 1 \rangle \subset k[t, x_1, \dots, x_n]$. Note that the extra variable t is designed to play a role of f^{-1} (mod J). If f vanishes on $\mathbb{V}(I)$ then $\mathbb{V}(J) = \emptyset$. By Nullstellensatz we have

$$1 = c(1 - tf) + \sum h_i g_i,$$

for some $c, h_1, \dots, h_r \in R$.
To complete the so-called *Rabinowitsch trick*. show that a power of f is in the radical $\sqrt{I}$ note that the equality above still holds when t is substituted by f^{-1} and denominators are cleared.

4.1.3. Irreducible varieties and prime ideals. A variety V is *irreducible* if it can not be decomposed as $V = V_1 \cup V_2$ where $V_1, V_2 \subsetneq V$ are strictly smaller varieties.

An ideal I is *prime* if for every pair $f, g \in R$,

$$fg \in I \implies f \in I \text{ or } g \in I.$$

Proposition 4.1.17. *Prime ideals and irreducible varieties are in one-to-one correspondence.*

Proof. Let V be an irreducible variety and consider $I = \mathbb{I}(V)$, which is radical by Exercise 4.1.15. Suppose $f, g \in R$ are such that $fg \in I$, but $f, g \notin I$. Then both $V_1 = \mathbb{V}(I + \langle f \rangle) \subsetneq \mathbb{V}(I)$ and $V_2 = \mathbb{V}(I + \langle g \rangle) \subsetneq \mathbb{V}(I)$ by Exercise 4.1.16. On the other hand, a point $x \in V$ belongs either to V_1 or to V_2, since $fg \in I$. Therefore, $V = V_1 \cup V_2$ contradicts the irreducibility of V. We conclude that I is prime.

Let I be prime. If $V = \mathbb{V}(I)$ is reducible, then $V = V_1 \cup V_2$ for $V_1, V_2 \subsetneq V$. We can find $f, g \in R$ such that $f \in \mathbb{I}(V_1) \setminus \mathbb{I}(V_2)$ and $g \in \mathbb{I}(V_2) \setminus \mathbb{I}(V_1)$. Now $fg \in I$ (as fg vanishes on V), but $f, g \notin I$, which can not happen, since I is prime. Therefore, V is irreducible. □

Exercise 4.1.18. *Show that*

(1) *every maximal ideal is prime;*
(2) *ideal $\langle x_1, \dots, x_m \rangle$ is prime for any m;*
(3) *the hypersurface $\mathbb{V}(f)$ is irreducible iff $f = g^r$ for $r \geq 1$ for an irreducible polynomial g.*

Define an *r-plane* in k^n to be a variety $\mathbb{V}(\ell_1, \cdots, \ell_{n-r})$ where $f_i \in R_{\leq 1}$ are linearly independent linear functions. One can use Exercise 4.1.16 to show that an r-plane is irreducible. An $(n-1)$-plane is called a *hyperplane*.

4.2. Zariski topology, irreducible decomposition, and dimension

We shall introduce a *topology* on the space k^n that is weaker than the usual topology induced by the distance metric.

4.2.1. Varieties as Zariski closed sets. The *closed sets* of *Zariski topology* are varieties $V \subseteq k^n$. The *open sets* are their complements $U = k^n \setminus V$, where V is closed. The geometric intuition dictated by the definition of a variety makes the axioms of a topolgy hold. One can use basic operations discussed in §4.3 to rigorously check that this indeed is a topology:

(1) The trivial subsets $\emptyset$ and k^n are closed sets.
(2) The intersection of a collection of closed sets is a closed set.
(3) The union of a finite number of closed sets is a closed set.

Given any subset $S \subseteq k^n$, we define the Zariski closure of S

$$\overline{S} = \mathbb{V}(\mathbb{I}(S)).$$

Exercise 4.2.1. *Show that the unit ball $B = \{ x \in \mathbb{C} \,|\, |x| \leq 1 \} \subseteq \mathbb{C}$ is neither a closed nor an open set in Zariski topology.*
What is the Zariski closure of B?

4.2.2. Irreducible decomposition of a variety. Every Zariski closed set can be decomposed into a finite union of irreducible components.

Proposition 4.2.2. *Let V be a variety. Then there exist irreducible varieties $V_i \subseteq V$, $i = 1, \ldots, r$, (called* irreducible components *of V) such that*

- *$V = V_1 \cup \cdots \cup V_r$ and*
- *V_i is not contained in V_j for $i \neq j$.*

Moreover, such a decomposition (called (irredundant) irreducible decomposition*) is unique.*

Proof. Consider a sequence of decompositions

$$V = V_1^{(i)} \cup \cdots \cup V_{r_i}^{(i)}$$

starting with $V = V_1^{(1)}$ where $V_1^{(1)} = V$. Given the i-th decomposition, if some component is reducible, then it can be replaced with a union of two strictly smaller varieties producing a finer decomposition at step $i + 1$.

Suppose this process does not terminate, i.e., each decomposition contains at least one reducible component. Then we can construct an infinite descending chain of varieties

$$V_{j_1}^{(1)} \supsetneq V_{j_2}^{(2)} \supsetneq V_{j_3}^{(3)} \supsetneq \cdots$$

which translates into the (strictly) ascending chain of their ideals

$$\mathbb{I}\Big(V_{j_1}^{(1)}\Big) \subsetneq \mathbb{I}\Big(V_{j_2}^{(2)}\Big) \subsetneq \mathbb{I}\Big(V_{j_3}^{(3)}\Big) \subsetneq \cdots$$

However, since the polynomial ring R is Noetherian, this can not happen. Therefore, an irreducible decomposition exists.

Suppose there are two irredundant irreducible decompositions: $V = V_1 \cup \cdots \cup V_r$ and $V = W_1 \cup \cdots \cup V_s$. Since a component V_i is irreducible,

$$V_i = V_i \cap V = (V_i \cap W_1) \cup \cdots \cup (V_i \cap W_s)$$

there exists j such that $V_i \cap W_j = V_i$. Therefore, either (1) $V_i = W_j$ or (2) $V_i \subsetneq W_j$. However, if (2) is the case, using the same argument, we may conclude that $W_j \subset V_{i'}$ for some i', which implies that $V_i \subsetneq V_{i'}$ with $i' \neq i$ and contradicts irredundancy.

Since (1) is the only option, we conclude that the irreducible decomposition is unique: every component in one decomposition has to match a component in the other. □

4.2.3. Dimension. First, we define the *dimension* of an irreducible variety $V \neq \emptyset$ to be the maximal length d of a chain

$$\emptyset \neq V_0 \subsetneq V_1 \subsetneq \cdots \subsetneq V_d = V$$

where V_i are irreducible varieties. Note that the dimension is finite due to Noetherianity.

The *dimension* of an arbitrary variety V is defined as the maximal dimension of its irreducible components.

Exercise 4.2.3. *Show that a finite set of points is 0-dimensional.*

Note: The geometric notion of dimension defined above corresponds to the algebraic notion of *Krull dimension*: the dimension of an ideal $I \subseteq R$ is the maximal length d of the chain

$$R \neq P_0 \supsetneq P_1 \supsetneq \cdots \supsetneq P_d \supset I$$

where P_i are prime ideals.
For a variety $\dim V = \dim \mathbb{I}(V)$; for an ideal $\dim I = \dim \mathbb{V}(I)$.

Define the *tangent plane* at a point $a \in V \subset k^n$ to be

$$T_a(V) = \ker\left(\frac{\partial F}{\partial x}(a)\right) + a \tag{4.2.1}$$

where $F = (f_1, \ldots, f_r)$ is a system of generators of $I = \mathbb{I}(V)$ and

$$\frac{\partial F}{\partial x} = \frac{\partial(f_1, \ldots, f_r)}{\partial(x_1, \ldots, x_n)} \in R^{r \times n}$$

be the jacobian of the system F. Note that $T_a(V)$ is a c-plane where $c = \operatorname{corank}\left(\frac{\partial F}{\partial x}(a)\right)$.

Exercise 4.2.4. *Prove that for every ideal $I \subset R$ the right hand side of (4.2.1) does not depend on the generating system F of I.*

We can now give an analytically intuitive definition of the *local dimension* of a variety V at a point $a \in V$:

$$\dim_a V = \dim T_a(V) = \operatorname{corank}\left(\frac{\partial F}{\partial x}(a)\right).$$

Exercise 4.2.5. *Show that for an irreducible variety V its local dimension is constant for points $a \in V \setminus W$ for some proper Zariski closed subset W of V. (Hint: In a stronger topology,* corank *is an upper semicontinuous function.)*

Exercise 4.2.6. *Show that for a positive-dimensional variety V and a generic hyperplane H going through a point $a \in V$*

$$\dim_a(V \cap H) = \dim_a(V) - 1.$$

In fact, the dimension of an irreducible variety equals its local dimension at a generic point (reference?).

The *dimension reduction principle* (following from Exercise 4.2.6) says that intersecting a variety V, $\dim V = m$, with a generic r-plane L of $\operatorname{codim} L = n - r \leq m$, reduces the dimension by $n - r$:

$$\dim(V \cap L) = \dim V - \operatorname{codim} L = m - n + r.$$

Note that if $\operatorname{codim} L > m$ then a generic r-plane L misses V, i.e., $V \cap L = \emptyset$.

Remark 4.2.7. *A variety $V \subseteq k^n$ and a generic r-plane, where $r = \operatorname{codim} V = n - \dim V$, meet at finitely many points.*

In fact, this generic intersection is 0-dimensional iff $r = \operatorname{codim} V$. This observation gives a way to determine $\dim V$.

Exercise 4.2.8. *Show that*

(1) *$\mathbb{V}(\langle x_1, \ldots, x_m \rangle) \subseteq k^n$ has dimension $n - m$ (codimension m);*
(2) *the dimension of an r-plane is r;*
(3) *a hypersurface $\mathbb{V}(f)$ in k^n has dimension $n - 1$ (codimension 1).*

A polynomial system $F = (f_1, \ldots, f_c) \subseteq R$ is a *regular sequence* if

$$\dim \mathbb{V}(f_1, \ldots, f_m) = n - m, \quad m = 1, \ldots, c.$$

Note: In algebraic terms, F is a regular sequence if f_m is not a zero-divisor in $R/\langle f_1, \ldots, f_{m-1} \rangle$, $m = 1, \ldots, c$.

A system $F = (f_1, \ldots, f_c) \subseteq R$ is a *local regular sequence* with respect to an irreducible variety V if

$$\dim (\text{irreducible component of } \mathbb{V}(f_1, \ldots, f_m) \text{ that contains } V) = n - m, \quad m = 1, \ldots, c.$$

Note: Our definition of a local regular sequence is equivalent to the definition of *set-theoretic local regular sequence* (at a generic point of an irreducible variety V) given in *commutative algebra*.

Every irreducible variety V is a *local complete intersection*: there is a local regular sequence F (with respect to V), such that V is an irreducible component of $\mathbb{V}(F)$. The following algorithm gives a way to construct such F given polynomials that define V.

Algorithm 4.2.1 $F = \textsc{LocalCompleteIntersection}(G, V)$

Require: $G = (g_1, \ldots, g_r) \subset R = \mathbb{C}[x_1, \ldots, x_n]$, a system of polynomials;
V, an irreducible component of $\mathbb{V}(G)$.
Ensure: F is a local regular sequence with respect to V.

```
F ← ∅
W ← ℂ^n
for i = 1 to r do
  if there is an irreducible component W' of 𝕍(F ∪ {g_i}) such that
                                  V ⊆ W' ⊊ W
  then
    F ← F ∪ {g_i}
    W ← W'
  end if
end for
```

Proof of correctness of Algorithm 4.2.1. At every step of the algorithm when a polynomial gets appended to the system F the new irreducible component W' containing V is strictly smaller than the old (irreducible variety) W. Hence, the dimension of W' is smaller than that of W; in fact,

$$\dim W' = \dim W - 1,$$

since we add only one more polynomial.

Let $c = |F|$ when algorithm terminates and

$$W_{i_1} \supsetneq W_{i_2} \supsetneq \cdots \supsetneq W_{i_c}$$

be the sequence of irreducible varieties produced (the values that W takes at the end of the loop at steps $1 \leq i_1 < i_2 < \cdots < i_c \leq r$). Showing that $W_{i_c} = V$ would conclude this proof.

Suppose $W_c \neq V$. Then there exists $g_i \in G$ such that $W_i' = W_{i-1}$ (the case when g_i does not get appended to F) and $W_{i_c} \cap \mathbb{V}(g_i) \neq V$. But then $W_i' \subseteq W_{i-1} \cap \mathbb{V}(g_i) \neq W_{i-1}$, which produces a contradiction. □

Note that it is implied that the number c of polynomials that the algorithm outputs equals $\operatorname{codim} V$.

4.3. Basic operations with ideals

Here we shall discuss what operations on varieties correspond to basic operations on ideals

4.3.1. Intersection of varieties: sum of ideals. Let $V, W \subseteq k^n$ be varieties; $V = \mathbb{V}(I)$ and $W = \mathbb{V}(J)$ for some ideals $I, J \in R$. Then $V \cap W = \mathbb{V}(I + J)$, since $V \cap W$ is precisely the set of points on which both polynomials of I and J vanish.

Note that even if I and J are radical ideals, $I + J$ is not radical in general.

Example 4.3.1. *The ideals $I = \langle y \rangle$ and $J = \langle y - x^2 \rangle$ in $k[x, y]$ are radical, since they have irreducible principle generators. However,*

$$I + J = \langle y, y - x^2 \rangle = \langle x^2, y \rangle$$

is not radical as $\sqrt{I + J} = \mathfrak{m}_0 = \langle x, y \rangle$.

4.3.2. Union of varieties: intersection or multiplication of ideals. Both intersection and multiplication of ideals correspond to taking union of the corresponding varieties.

Exercise 4.3.2. *Let $V, W \subseteq k^n$ be varieties; $V = \mathbb{V}(I)$ and $W = \mathbb{V}(J)$ for some ideals $I, J \in R$.*

1. *Show that $V \cup W = \mathbb{V}(IJ) = \mathbb{V}(I \cap J)$;*
2. *Show that if I and J are radical, $I \cap J$ is.*
3. *Find an example of radical I and J such that IJ is not.*

Let $I = \langle f_1, \ldots, f_r \rangle$ and $J = \langle g_1, \ldots, g_s \rangle$. While finding a set of generators for IJ is straightforward, namely

$$IJ = \langle\, f_i g_j \,|\, i = 1, \ldots, r; j = 1, \ldots, s \,\rangle,$$

how would one construct generators of $I \cap J$?

Proposition 4.3.3. *Given $I = \langle f_1, \dots, f_r \rangle$ and $J = \langle g_1, \dots, g_s \rangle$, ideals in $R = k[x] = k[x_1, \dots, x_n]$, define*

$$K = tI + (1-t)J = \langle tf_1, \dots, tf_r, (1-t)g_1, \dots, (1-t)g_s \rangle \subset k[x,t].$$

Then $K \cap k[x] = I \cap J$.

PROOF. The inclusion $I \cap J \subseteq K \cap k[x]$ is straightforward: if $f \in I \cap J \subseteq k[x]$ then $f = tf + (1-t)f$ is in K (and still in $k[x]$).

Suppose $f \in K \cap k[x]$, i.e., $f = g + h$ where $g \in tI$ and $h \in (1-t)J$. Then $g|_{t=0} = 0$, therefore,

$$f = f|_{t=0} = g|_{t=0} + h|_{t=0} = h|_{t=0} \in J.$$

Similarly,

$$f = f|_{t=1} = g|_{t=1} + h|_{t=1} = g|_{t=1} \in I.$$

We conclude that $f \in I \cap J$. □

Note that Proposition 4.3.3 provides an algorithmic way to compute generators of $I \cap J$ via elimination.

4.3.3. Difference of varieties: colon ideal. The difference of two varieties V and W is not a variety in general: for instance take $V = \mathbb{V}(x)$ and $W = \mathbb{V}(y)$ as varieties in k^2. Then $V \setminus W = V \setminus \{(0,0)\}$ is not Zariski closed: $\overline{V \setminus W} = V$ as every polynomial that vanishes on $V \setminus W$ must vanish on $\{(0,0)\}$ as well.

A good questions is: how to find $\overline{V \setminus W}$ in general?

The answer is not hard to give if one can construct an irreducible decomposition

$$V = V_1 \cup \dots \cup V_r.$$

In this case,

$$\overline{V \setminus W} = \bigcup_{\{\, i \,|\, V_i \not\subseteq W \,\}} V_i.$$

In algebraic language (without resorting to decomposition algorithms), the construction of *colon ideal* (also called *quotient ideal*) provides an answer. Let $I, J \subseteq R$ be ideals, define

$$I : J = \{\, f \in R \,|\, fJ \subseteq I \,\}.$$

Exercise 4.3.4. *Show that $I : J$ is an ideal.*

Proposition 4.3.5. *If $I, J \subseteq R$ and I is a radical ideal, then*

$$\overline{\mathbb{V}(I) \setminus \mathbb{V}(J)} = \mathbb{V}(I : J).$$

Moreover, $\mathbb{I}\Big(\overline{\mathbb{V}(I) \setminus \mathbb{V}(J)}\Big) = I : J$.

PROOF. Let $V = \mathbb{V}(I) = V_1 \cup \dots \cup V_r$ be the irreducible decomposition, $W = \mathbb{V}(J)$, and set

$$V' = \bigcup_{\{\, i \,|\, V_i \not\subseteq W \,\}} V_i.$$

Taking $f \in I : J$ we can show that f vanishes on all $V_i \not\subseteq W$: pick a polynomial $g \in J$ such that $g \notin \mathbb{I}(V_i)$ then $fg \in fJ \subseteq I$. Since g does not vanish on V_i, the polynomial f has to.

On the other hand, take $f \in \mathbb{I}(V')$ then $fg \in I$ for every $g \in J$ and, hence, $f \in I : J$. □

Exercise 4.3.6. *If $I = \langle f_1, \dots, f_r \rangle$ and $J = \langle g \rangle$ one can compute the generators h_i of the intersection,*

$$I \cap J = \langle h_1, \dots, h_s \rangle .$$

Show that $I : J = \langle h_1/g, \dots, h_s/g \rangle$.

Exercise 4.3.7. *Show that $I : (J + K) = (I : J) \cap (I : K)$ for any ideals I, J, K.*

The exercises above give a way of constructing generators of $I : J$ algorithmically for any $I = \langle f_1, \dots, f_r \rangle$ and $J = \langle g_1, \dots, g_s \rangle$:

(1) compute generators of $K_i = I : \langle g_i \rangle$ using the conslusion of Exercise 4.3.6,
(2) compute generators of $K_1 \cap \cdots \cap K_s$.

The output of the second step is generators of $I : J$ by Exercise 4.3.7.

One can find algorithmically $\overline{\mathbb{V}(I) \setminus \mathbb{V}(J)}$ without the assumption of radicality of I that Proposition 4.3.5 makes. To that end, one needs to compute the *saturation* of I with respect to J,

$$I : J^\infty = \{ f \in R \mid fJ^m \subseteq I \text{ for some } m \} .$$

The saturation ideal $I : J^\infty$ is the ideal at which the chain

$$I : J \subseteq I : J^2 \subseteq I : J^3 \subseteq \cdots$$

stabilizes.

4.3.4. Projection of variety: intersection with a subring. Define $\pi_m : k^n \to k^m$ to be the projection map that sends $(x_1, \dots, x_n)$ to $(x_1, \dots, x_m)$.

Proposition 4.3.8. *Let $V = \mathbb{V}(I) \subseteq k^n$ be a variety given by some ideal $I \subseteq k[x_1, \dots, x_n]$. Then $\overline{\pi_m(V)} = \mathbb{V}(I_m)$, where $I_m = I \cap k[x_1, \dots, x_m]$.*

Proof. A point $(a_1, \dots, a_m) \in \pi_m(V)$ clearly satisfies all polynomials in $I \cap k[x_1, \dots, x_m]$. Hence, $\overline{\pi_m(V)} \subseteq \mathbb{V}(I)$

Let $f \in \mathbb{I}(\pi_m(V))$ then $f \in \mathbb{I}(V) = \sqrt{I}$. Since $\mathbb{I}(\pi_m(V)) \subseteq \sqrt{I}$, the inclusion reversing property of ideal-variety correspondence (Exercise 4.1.2) implies that $\mathbb{V}(I) \subseteq \overline{\pi_m(V)}$. □

Note that this proposition makes the closure of the projection of a variety computable via elimination.

Projections of varieties (more generally images of varieties under algebraic maps) are not Zariski closed.

Example 4.3.9. *Consider $V = \mathbb{V}(xy - 1) \subset k^2$. The projection $\pi_x(V)$ to the coordinate x is the set $k \setminus \{0\}$, which is not Zariski closed.*

Note: In fact, projections of varieties (more generally, images of varieties under algebraic maps) are *constructible sets*, i.e., sets of the form

$$V = (V_1 \setminus W_1) \cup \cdots \cup (V_r \setminus W_r),$$

where V_i and W_i are Zariski closed.

4.4. Multiplicity

Describe how to compute the multiplicity of the origin. Mention "multiplicity structure".

CHAPTER 5

Numerical algebraic geometry

The classical way of viewing varieties is through the lens of commutative algebra, namely, the ideal-variety correspondence studied in Chapter 4. While this approach has its indisputable advantages, its main deficiency is the reliance on Gröbner bases techniques for computation: the worst case complexity of Buchberger-type algorithms is doubly exponential in the number of variables and Gröbner bases are not suited well for approximate computation.

What this chapter introduces is a relatively novel approach of the *numerical algebraic geometry*, which uses numerical polynomial homotopy continuation of Chapter 2 as its main computational engine. Unlike Gröbner bases algorithms, algorithms for homotopy continuation are numerical in their essence (they may use approximate computations) and also allow for straightforward parallelization.

5.1. Witness sets

Recall two properties of an irreducible variety $V \subset \mathbb{C}^n$ of dimension m:

- There is a local regular sequence $F \in R^{n-m}$, $R = \mathbb{C}[x_1, \ldots, x_n]$, with respect to V. (See Algorithm 4.2.1.) In particular, V is an irreducible component of $\mathbb{V}(F)$.
- Every generic $(n-m)$-plane L intersects V in finitely many points. (See Remark 4.2.7.)

The latter is commonly referred to as *Bertini's theorem*, which classically makes a stronger claim: all points of the intersection are regular and their number equals the *degree* of the variety. One can take this as a geometric definition of the degree.

5.1.1. Definition. Let $F \subseteq R$ be a system of $n-m$ polynomials and $V_1, \ldots, V_r$ be irreducible components of $\mathbb{V}(F)$ of dimension m. Note that F is a local regular sequence with respect to every component V_i. The union $V = V_1 \cup \cdots \cup V_r$ is an *equidimensional variety*, i.e., a variety whose irreducible components have the same dimension.

A *witness set* w representing an equidimensional variety V of dimension m is a triple $w = [F, L, W]$ where

(1) $F \subseteq R = \mathbb{C}[x_1, \ldots, x_n]$ is a polynomial system that is a local regular sequence with respect to every irreducible component of V;
(2) L is a generic $(n-m)$-plane, a so-called *slicing plane*;
(3) $W = V \cap L \subseteq \mathbb{V}(F) \cap L$, a finite set of points referred to as *witness points*.

We denote the variety represented by a witness set w by $\mathbb{V}(w)$. A witness set w is an *irreducible witness set* if $\mathbb{V}(w)$ is irreducible.

Example 5.1.1. *The variety*

$$\mathbb{V}\begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \mathbb{V}\begin{pmatrix} (x^2+y^2+z^2-4)(y-x^2)(y+x^2) \\ (x^2+y^2+z^2-4)(y-1)z \\ (x^2+y^2+z^2-4)(z-1)z \end{pmatrix} \subseteq \mathbb{C}^3$$

breaks into the following irreducible components:

dim=2: *the sphere* $V_\circ = \mathbb{V}\big(x^2+y^2+z^2-4\big)$ *(corresponding to the common factor present in all defining polynomials);*

dim=1: *two parabolas* $V_\cup = \mathbb{V}\big(y-x^2, z\big)$ *and* $V_\cap = \mathbb{V}\big(y+x^2, z\big)$ *in* (x,y) *coordinate plane (corresponding to the case when* $x^2+y^2+z^2-4 \neq 0$*, but* $z=0$*);*

dim=0: *four points* $V_1, \ldots, V_4$ *that are irreducible components of*

$$\mathbb{V}\big((y-x^2)(y+x^2), y-1, z-1\big)$$

(corresponding to the case $x^2+y^2+z^2-4 \neq 0$ *and* $z \neq 0$*), namely,*

$$V_1 = \{(1,1,1)\}\,,\ V_2 = \{(-1,1,1)\}\,,\ V_3 = \{(\boldsymbol{i},1,1)\}\,,\ \text{and } V_4 = \{(-\boldsymbol{i},1,1)\}\,.$$

To represent $V_\circ$ *with a witness set we need a generic line (1-plane): let us take* $L_\circ = \mathbb{V}(x-z-1, y-z-2)$*. The two points of the intersection* $V_\circ \cap L_\circ$ *are*

$$(z+1, z+2, z),\ \text{where } z = -1 \pm \sqrt{\frac{2}{3}}.$$

Now $\mathbb{V}(w_\circ) = V_\circ$ *for a witness set*

$$w_\circ = \left[(f_1)\,, L_\circ, \left\{ (z+1, z+2, z) \,\middle|\, z = -1 \pm \sqrt{\frac{2}{3}} \right\} \right].$$

Note that there is a lot of freedom in constructing a witness set $w_\circ$*: for intance, we can*

- *replace* f_1 *with* f_2 *or* f_3*;*
- *choose any other generic line instead of* $L_\circ$ *above.*

We can represent the equidimensional variety $V_\cup \cup V_\cap$ *with one witness set: take the 2-plane* $L_{\cup\cap} = \mathbb{V}(y+z-4)$ *and construct*

$$w_{\cup\cap} = \left[(f_1, f_2)\,, L_{\cup\cap}, \{(2,4,0), (-2,4,0), (2\boldsymbol{i},4,0), (-2\boldsymbol{i},4,0)\} \right].$$

Again, there is a lot of freedom of choice here: for example, f_2 *can be replaced with* f_3 *since* (f_1, f_3) *is also a regular sequence with respect to both* $V_\cup$ *and* $V_\cap$*.*

Now, to represent one of the irreducible components – say, $V_\cup$*, for instance – we need to make only a small change in* $w_{\cup\cap}$*: namely, keep only the witness points that belong to* $V_\cup$*, i.e.,*

$$w_\cup = \left[(f_1, f_2)\,, L_{\cup\cap}, \{(2,4,0), (-2,4,0)\} \right].$$

To represent the equidimensional variety $V_1 \cup V_2 \cup V_3 \cup V_4$ *we need a* 3*-plane: the only* 3*-plane is the whole space* $\mathbb{C}^3$ *making a witness set*

$$w_{1234} = \left[(f_1, f_2, f_3)\,, \mathbb{C}^3, \{(1,1,1), (-1,1,1), (\boldsymbol{i},1,1), (-\boldsymbol{i},1,1)\} \right].$$

Exercise 5.1.2. *For the varieties* $V_\circ$ *and* $V_\cup$ *of Example 5.1.1 construct witness sets corresponding to alternative choices of slicing planes*

$$L_\circ = \mathbb{V}(x+y+z-3, x-y+2z-5) \text{ and}$$
$$L_\cup = \mathbb{V}(x-y+2z-5).$$

Exercise 5.1.3. *Show that for a hypersurface $V = \mathbb{V}(f)$ given by a squarefree polynomial f the number of witness points in a witness set equals $d = \deg f$. (Hint: Eliminate all variables but one.)*

5.1.2. Numerical construction. If $F = (f_1, \ldots, f_c)$ is a polynomial system that defines an equidimensional variety $V = \mathbb{V}(F) \subseteq \mathbb{C}^n$ of codimension c, then we already have the first ingredient for a witness set: F is a locally regular sequence with respect to the irreducible components of V. For the second ingredient take a generic c-plane $L = \mathbb{V}(\ell_1, \cdots, \ell_m)$, where $m = n - c = \dim V$ and ℓ_i are linear independent linear functions. The third (missing) ingredient is $V \cap L$.

To compute the point of $V \cap L$, solve the square polynomial system $(f_1, \ldots, f_c, \ell_1, \cdots, \ell_m)$. We may apply Theorem 2.1.5; in fact, we can fix the linear equations defining the slicing plane in the homotopy:

$$H_t = \begin{pmatrix} (1-t)G + \gamma t F \\ \ell_1 \\ \vdots \\ \ell_m \end{pmatrix}, \quad t \in [0,1],$$

where $G = (g_1, \ldots, g_c)$ are the first c polynomials in the total-degree start system and $\gamma \in \mathbb{C}$ is generic.

This is the driving idea behind respresenting varieties with witness sets, since for their construction we can rely exclusively on numerical polynomial homotopy continuation algorithms. Since we intend to use witness sets as a practical data structure, instead of the witness points their numerical approximations are stored.

We will come back to the question of creating witness sets for components of $\mathbb{V}(F)$ for general F when we discuss decomposition of a general variety $\mathbb{V}(F)$.

5.1.3. Equivalence of witness sets. Two witness sets $[F, L, W]$ and $[F, L', W']$ are *equivalent* if $|W| = |W'|$ and for the homotopy

$$H_t^{[F, L \to L', \gamma]} = \begin{pmatrix} F \\ (1-t)\ell_1 + \gamma t \ell_1' \\ \vdots \\ (1-t)\ell_m + \gamma t \ell_m' \end{pmatrix}, \quad t \in [0,1], \tag{5.1.1}$$

with $\gamma \in \mathbb{C}$ generic there are $d = |W| = |W'|$ homotopy paths that

- do not intersect,
- start with witness points W at $t = 0$, and
- end with the witness points W' at $t = 1$.

Equivalent witness sets represent the same (equidimensional) variety.

Note: We make no assumption that the starting points W in the homotopy $H_t^{[F, L \to L', \gamma]}$ in (5.1.1) are regular.

However, due to genericity of the slicing planes L and L', the points on the homotopy paths starting at W all have "similar singularity structure". This enables the deflation technique of §2.2.1 to regularize all points on the paths uniformly making it possible to track $H_t^{[F, L \to L', \gamma]}$ with a numerical algorithm.

5.1.4. Sampling and the membership test. Given a witness set

$$w = [F, L, W]$$

representing a positive-dimensional variety, one can sample infinitely many points on $\mathbb{V}(w)$ repeatedly picking a random slicing plane L' in (5.1.1) and tracking the resulting homotopy to get new points $\mathbb{V}(w) \cap L'$.

On the other hand, given a point $p \in \mathbb{C}^n$ one can perform a *membership test*, i.e., determine whether $p \in \mathbb{V}(w)$. To test membership,

- pick a random slicing plane L' that contains the point p;
- track the homotopy $H_t^{[F, L \to L', \gamma]}$ in (5.1.1) for a random $\gamma \in \mathbb{C}$ to get points $W' = \mathbb{V}(w) \cap L'$;
- $p \in \mathbb{V}(w)$ iff $p \in W'$.

Note that $[F, L', W']$ may not form a witness set; nevertheless, W' is a finite set with $|W'| \leq |W| = \deg \mathbb{V}(w)$ for a generic choice L' passing through p.

Exercise 5.1.4. *For the varieties $V_\circ$ and $V_\cup$ in Example 5.1.1*

- *determine their degrees;*
- *show that the slicing line $L_\circ = \mathbb{V}(y, z-2)$ does not produce a witness set for $V_\circ$;*
- *give an example of an exceptional slicing plane $L_\cup$, one that can not be a part of a witness set representing $V_\cup$.*

Remark 5.1.5. *The membership test can be used to perform determine containment (with probability 1): for a witness set $\widetilde{w} = \left[\widetilde{F}, \widetilde{L}, \widetilde{W}\right]$ the variety $\mathbb{V}(\widetilde{w}) \subseteq \mathbb{V}(w)$ iff $\widetilde{W} \subseteq \mathbb{V}(w)$.*

5.2. Numerical irreducible decomposition

In this section we discuss the following problem:

> Given a witness set
>
> $$w = [F, L, W]$$
>
> representing an equidimensional variety, how does one obtain the partition of the witness points
>
> $$W = W_1 \sqcup \cdots \sqcup W_r$$
>
> such that $V_i = \mathbb{V}([F, L, W_i])$, $i = 1, \ldots, r$, are the irreducible components of $V = \mathbb{V}(w)$?

For instance, for the witness set $w_{\cup\cap}$ in Example 5.1.1 we would like to have an algorithmic way to produce $w_\cup$ and $w_\cap$ using only the data of $w_{\cup\cap}$.

5.2.1. Irreducible witness sets. Let us assume that we have the required partition

$$W = W_1 \sqcup \cdots \sqcup W_r.$$

Pick a generic slicing plane L' and consider homotopy paths starting at the points of W_i traced out by $H_t^{[F, L \to L', \gamma]}$ in (5.1.1). Due to genericity of L,L', and γ, all points on the paths are generic: in particular, they do not belong to intersections $V_i \cap V_j$, $i \neq j$, of the irreducible components. We conclude that the set of endpoints W_i' is contained in the same component V_i and, therefore, gives another (equivalent to $[F, L, W_i]$) witness set $[F, L, W_i']$ for V_i.

Now consider a map $\phi_{L', \gamma, \gamma'} : W \to W$ which maps a witness point $p \in W$ to the result of tracking of two homotopy paths:

- the path of $H_t^{[F,L\to L',\gamma]}$ starting at p and ending at some point $q \in W'$,
- then the path of $H_t^{[F,L'\to L,\gamma']}$ starting at q and ending at some point $\phi(p) \in W$.

Assuming that all ingredients (L, L', γ, γ') are generic, the map $\phi_{L',\gamma,\gamma'}$

- is a permutation of W,
- restricts to a permutation of each W_i (according to our discussion above).

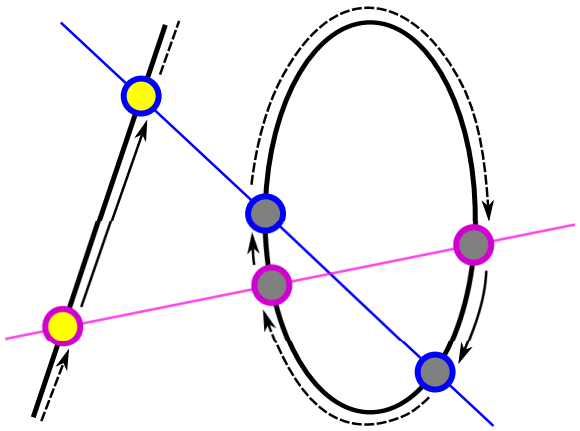


FIGURE 1. A cubic is a union of an ellipse and a line: homotopies $H_t^{[F,L\to L',\gamma]}$ (solid arrows) and $H_t^{[F,L'\to L,\gamma']}$ (dashed arrows) establish that two witness points belong to the same component (ellipse).

Figure 1 depicts a scenario when two points in a witness set of a reducible cubic are permuted, wherefore, showing that they belong to the same irreducible component.

Proposition 5.2.1. *Using the setting of this section, restrict to W_i, the witness points corresponding to an irreducible component V_i of V.*

Then the group generated by permutations $\phi_{L',\gamma,\gamma'}|_{W_i} : W_i \to W_i$, where L', γ, γ' are generic, acts transitively on W_i.

In other words, for every pair of points points $p, q \in W_i$, there exists $L_i', \gamma_i, \gamma_i'$, $i = 1, \dots, r$, such that

$$\left(\phi_{L_r',\gamma_r,\gamma_r'} \circ \cdots \circ \phi_{L_1',\gamma_1,\gamma_1'}\right)(p) = q,$$

i.e., there is a finite number of "moves", whose composition maps p to q.

PROOF. add a reference □

5.2.2. Monodromy breakup algorithm. The modern numerical algebraic geometry software often uses the following probabilistic algorithm, Algorithm 5.2.1, for *numerical irreducible decomposition* of an equidimensional variety.

The only missing ingredient in this algorithm is a stopping criterion. We postpone the proof of correctness until the end of the next section.

Algorithm 5.2.1 $P = \text{MonodromyBreakup}(F, L, W)$

Require: $[F, L, W]$, a witeness set representing an equidimensional variety of dimension m:
- $F = (f_1, \dots, f_{n-m})$, a polynomial system in $R = \mathbb{C}[x_1, \dots, x_n]$;
- $L = \mathbb{V}(\ell_1, \dots, \ell_m)$, an $(n-m)$-plane given by a system of linearly independent linear functions;
- $W \subseteq \mathbb{C}^n$, a finite number of points.

Ensure: $P = \{W_1, \dots, W_r\}$ is a partition of the set of witness points W according to the irreducible decomposition of $\mathbb{V}([F, L, W])$.

$P \leftarrow \{\, \{p\} \,|\, p \in W \,\}$ – initialize P with the partition into singletons
while a stopping criterion is not satisfied **do**
Pick randomly
- linear functions ℓ_i' to get an $(n-m)$-plane $L' = \mathbb{V}(\ell_1', \dots, \ell_m')$;
- constants $\gamma, \gamma' \in \mathbb{C}$.

$\phi \leftarrow \phi_{L', \gamma, \gamma'}$
In the partition P merge parts $A \subseteq W$ and $B \subseteq W$ if there exists a pair of points $p \in A$ and $q \in B$ such that $\phi(p) = \phi(q)$.
end while

5.2.3. Linear trace test. Here we describe an idea of a stopping criterion for Algorithm 5.2.1 in the case of a curve in a plane (1-equidimensional variety in $\mathbb{C}^2$), however the approach generalizes to varieties of arbitrary dimension and codimension.

A curve V in a plane has $\dim V = \operatorname{codim} V = 1$ and can be represented by a witness set $w = [\{f\}, \mathbb{V}(\ell), W]$, i.e., with

(1) one polynomial $f \in \mathbb{C}[x, y]$,
(2) one linear function $\ell \in \mathbb{C}[x, y]$, and
(3) a finite set of witness points $W = \mathbb{V}(f) \cap \mathbb{V}(\ell) \subseteq \mathbb{C}^2$.

Moreover, after a linear change of coordinates, we may assume that

- the monomial x^d, where $d = \deg f$, occurs in f with coefficient 1 and
- the slicing line $L = \mathbb{V}(\ell)$ is the x-axis, i.e., $\ell = y$.

Consider the family of lines parallel to L defined by

$$\ell_t = y - t, \quad t \in \mathbb{C}$$

and the witness sets obtained by deforming W into $W_t \subseteq \mathbb{V}(f) \cap \mathbb{V}(\ell_t)$ as the parameter t changes. There may be finitely many values of t that go not give a generic line $\mathbb{V}(\ell_t)$, i.e., the intersection $V \cap \mathbb{V}(\ell_t)$ would be not typical: recall that typically $|V \cap \mathbb{V}(\ell_t)| = |W| = d$, where $d = \deg V = \deg f$.

Let $W_t = \{(x_1(t), t), \dots, (x_d(t), t)\}$, then $x_i(t)$ are the roots of the univariate polynomial

$$\begin{aligned} f(x, t) &= x^n + a_{n-1}(t)x^{n-1} + \cdots + a_1(t)x + a_0(t) \\ &= (x - x_1(t)) \cdots (x - x_d(t)) \\ &= x^d - \boxed{(x_1(t) + \cdots + x_d(t))} x^{d-1} + \cdots + (-1)^d (x_1 \cdots x_n) \end{aligned}$$

The *trace* $\operatorname{tr}(f)$ of a univariate monic polynomial f is defined to be the sum of its roots. In our case the $\lambda(t) = \operatorname{tr}(f(x, t)) = x_1(t) + \cdots + x_d(t) = -a_{n-1}$ is a linear function in t: indeed, we substituted

the parameter t for y in the bivariate polynomial f of degree d, therefore, only terms with monomials x^{d-1} and tx^{d-1} contribute to the term $a_{n-1}(t)x^{n-1}$.

The idea of the *linear trace test* is to check the linearity of the trace numerically. In practice, it is enough to compute the trace $\lambda(t)$ for three distinct generic values of parameter t to conclude whether $\lambda(t)$ is linear.

Proposition 5.2.2. *Consider a witness set $w = [F, L, W]$, $L = \mathbb{V}(\ell_1, \ldots, \ell_m)$, representing an m-equidimensional variety in $\mathbb{C}^n$ and a subset of witness points $W' \subseteq W$. Let $W'_t \subseteq W_t \subseteq \mathbb{V}(F) \cap L_t$, where*

$$L_t = \mathbb{V}(\ell_1 - t, \ldots, \ell_m - t), \quad t \in \mathbb{C},$$

be the points obtained by tracking homotopy $H^{[F, L \to L_t, \gamma]}$, for a fixed generic $\gamma \in \mathbb{C}$. (Note that W_t is defined for a dense open subset of $\mathbb{C}$.)

Then, provided the choice of ℓ_i is generic, $W' = W \cap V$ for some subvariety of $V \subseteq \mathbb{V}(w)$ iff

$$\lambda(t) = \sum_{p \in W'_t} p \ \in \ \mathbb{C}^n$$

is a linear function of t.

Proof. For a subvariety $V \subseteq \mathbb{V}(w)$ and a generic choice of ℓ_i, the intersection $L_t \cap V$ gives a curve in $\mathbb{C}^{n+1}$ with coordinates $x_1, \ldots, x_n, t$. The projection of this curve to $\mathbb{C}^2$ with coordinates x_i, t, for $i \in [n]$, is either a point or, again, an irreducible curve.

This observation reduces the argument to the above discussion of a planar curve. □

Using this proposition we can design a stopping criterion for Algorithm 5.2.1: the algorithm should stop when all parts W' in the partition P pass the linear trace test, i.e., $\lambda(t)$ in Proposition 5.2.2 is linear.

Proof of correctness of Algorithm 5.2.1. (with a linear trace test as a stopping criterion)

Parts in the partition P are merged by the algorithm if points in distinct parts are discovered to be in the same irreducible component in accordance with Proposition 5.2.1. Since the algorithm starts with partitioning W into single-element sets, during the run of the algorithm it is always true that every part in the partition is contained in exactly one irreducible component.

The algorithm stops when each part in P is a "complete" set of witness points corresponding to a subvariety according to Proposition 5.2.2.

Putting these two observations together, we coclude that the parts of P represent an irreducible components when the algorithm terminates. □

Exercise 5.2.3. *For the ellipsoid $V = \mathbb{V}\big(x^2 + 2y^2 + 3z^2 - 5\big)$*

(1) *find the points p_0 and q_0 of the intersection $V \cap L$ for*
$$L = \mathbb{V}(y - x - 1, z - x)\,;$$
(2) *find the points p_t and q_t of the intersection $V \cap L_t$ for*
$$L_t = \mathbb{V}(y - x - 1 - t, z - x - t)\,,$$
where $t \in \mathbb{C}$ is a parameter;
(3) *show that $p_t + q_t$ is a linear function in t, while p_t and q_t are not.*

5.3. Numerical variety

A *numerical variety* $\mathbf{w}$ is defined a finite collection of witness sets $\{w_1, \ldots, w_r\}$.

A numerical variety is viewed as a (non-unique) representation of the variety $\mathbb{V}(\mathbf{w}) = \mathbb{V}(w_1) \cup \cdots \cup \mathbb{V}(w_r)$. Note that we **do not** require

- witness sets w_i to be irreducible, although one can produce a numerical variety with irreducible witness sets using numerical irreducible decomposition algorithms;
- the collection $\mathbf{w}$ to be *irredundant*, i.e,

$$\mathbb{V}(w_i) \not\subseteq \mathbb{V}(w_j), \quad i \neq j,$$

  although one can easily use the containment test (see Remark 5.1.5) to discard the redundant parts.

In this section we explain the basic operations for numerical varieties and describe an algorithm for constructing a numerical variety representing $\mathbb{V}(F)$ where F is an arbitrary polynomial system.

5.3.1. Union and difference. While both taking a union and subtracting two varieties is a nontrivial operation using the ideal-variety correspondence, these two operations are simple in the numerical setting. Let $V = \mathbb{V}(\mathbf{w})$ and $\widetilde{V} = \mathbb{V}(\widetilde{\mathbf{w}})$.

Then the union is

$$V \cup \widetilde{V} = \mathbb{V}(\mathbf{w} \cup \widetilde{\mathbf{w}})$$

and the Zariski closure of the difference is

$$\overline{V \setminus \widetilde{V}} = \mathbb{V}\left\{ [w \in \mathbf{w} \,|\, \mathbb{V}(w) \not\subseteq \mathbb{V}(\widetilde{w}) \text{ for all } \widetilde{w} \in \widetilde{\mathbf{w}} \right\},$$

assuming the witness sets of $\mathbf{w}$ are irreducible.

5.3.2. Intersection with a hypersurface. On the other hand, the intersection is a simple operation when varieties are represented by ideals, but intersecting numerical varieties is a nontrivial task. We start by showing how to intersect a numerical variety with a hypersurface.

Consider a numerical variety $\mathbf{w}$ and a polynomial $g \in R = \mathbb{C}[x_1, \ldots, x_n]$. Our goal is to construct a numerical variety $\mathbf{w}'$ such that

$$\mathbb{V}(\mathbf{w}') = \mathbb{V}(\mathbf{w}) \cap \mathbb{V}(g).$$

For a witness set $w = [F, L, W] \in \mathbf{w}$ the following scenarios and corresponding actions are possible:

(1) $\mathbb{V}(w) \cap \mathbb{V}(g) = \mathbb{V}(w)$, i.e., $\mathbb{V}(w) \subseteq \mathbb{V}(g)$ or, equivalently, g vanishes on all witness points W.
   Action: append w to $\mathbf{w}'$.

(2) $\mathbb{V}(w) \cap \mathbb{V}(g) \neq \mathbb{V}(w)$, but $\mathbb{V}(w) \cap \mathbb{V}(g) = \dim \mathbb{V}(w)$. That means g vanishes on a proper subset $W' \neq \emptyset$ of witness points W.
   Action: append $[F, L, W']$ to $\mathbf{w}'$; proceed to case (3) with $w := [F, L, W \setminus W']$.

(3) $\dim(\mathbb{V}(w) \cap \mathbb{V}(g)) < \dim(\mathbb{V}(w))$, which is the case when $W \cap \mathbb{V}(g) = \emptyset$. There are two subcases, both addressed by Algorithm 5.3.1:
   (a) $\dim(\mathbb{V}(w) \cap \mathbb{V}(g)) = \dim(\mathbb{V}(w)) - 1$.
      Action: append $[F \cup \{g\}, L', W']$ to $\mathbf{w}'$, where L' is a random plane of dimension $n - \dim(\mathbb{V}(w)) + 1$ and W' is constructed by Algorithm 5.3.1.
   (b) $\mathbb{V}(w) \cap \mathbb{V}(g) = \emptyset$.
      Action: discard.

Algorithm 5.3.1 *output* = HypersurfaceIntersection$([F, L, W], g, L')$

Require: $[F, L, W]$, a witeness set representing an equidimensional variety of dimension m:
- $F = (f_1, \ldots, f_{n-m})$, a polynomial system in $R = \mathbb{C}[x_1, \ldots, x_n]$;
- $L = \mathbb{V}(\ell_1, \ldots, \ell_m)$, an $(n-m)$-plane given by a system of linearly independent linear functions;
- $W \subseteq \mathbb{C}^n$, a finite number of points;

$g \in R$ such that $W \cap \mathbb{V}(g) = \emptyset$;
$L' = \mathbb{V}\big(\ell'_1, \ldots, \ell'_{m-1}\big)$, a generic $(n-m+1)$-plane.

Ensure: *output* $= [F \cup \{g\}, L', W']$ such that

$$\mathbb{V}([F \cup \{g\}, L', W']) = \mathbb{V}([F, L, W]) \cap \mathbb{V}(g)$$

or *output* $= \emptyset$ if $\mathbb{V}([F, L, W]) \cap \mathbb{V}(g) = \emptyset$.

$d \leftarrow \deg g$
for $i = 1$ to d **do**
  Pick a random linear function $\ell^{(i)}$.
  $L^{(i)} \leftarrow \big\{\ell^{(i)}, \ell'_1, \ldots, \ell'_{m-1}\big\}$
  For a random $\gamma \in \mathbb{C}$, track the homotopy $H_t^{[F, L \to L^{(i)}, \gamma]}$
  - starting with points W
  - to get witness points $W^{(i)}$.

    - Note that $\big[F, L^{(i)}, W^{(i)}\big]$ is a witness set equivalent to $[F, L, W]$.

end for
For a random $\gamma \in \mathbb{C}$, track the homotopy

$$H_t = \begin{pmatrix} F \\ (1-t)(\ell^{(1)}\ell^{(2)}\cdots\ell^{(d)}) + \gamma t g \\ \ell'_1 \\ \vdots \\ \ell'_{m-1} \end{pmatrix}, \quad t \in [0, 1], \tag{5.3.1}$$

starting with the points $W^{(1)} \cup \cdots \cup W^{(d)}$. Let W' be the set of points obtained as the result.
if $W' = \emptyset$ **then**
  output $\leftarrow \emptyset$
else
  output $\leftarrow [F \cup \{g\}, L', W']$
end if

Note that the only part homotopy H_t in (5.3.1) that depends on t,

$$h_t = (1-t)(\ell^{(1)}\ell^{(2)}\cdots\ell^{(d)}) + \gamma t g,$$

evaluates to a constant multiple of g at $t = 1$ and

$$h_0 = \ell^{(1)}\ell^{(2)}\cdots\ell^{(d)}$$

at $t = 0$. While g is an arbitrary polynomial, h_0 factors into a product of linear functions; also, $\deg h_0 = \deg g$.

Remark 5.3.1. *Going from $t = 1$ to $t = 0$ above is commonly referred to as* degeneration*: a general polynomial* degenerates *into a polynomial of the same "kind" (in this case, of the same degree), but with special properties (in this case, factors into a product of linear functions).*

Going in the opposite direction, from $t = 0$ to $t = 1$ can be called undegeneration *or* regeneration*.*

The special form of h_0 makes the problem simpler than that given by a general polynomial g. Indeed, intersecting $V = \mathbb{V}([F, L, W])$ with $\mathbb{V}(h_0)$ breaks into d smaller problems:

$$V \cap \mathbb{V}(h_0) = \left(V \cap \mathbb{V}\left(\ell^{(1)}\right)\right) \cup \cdots \cup \left(V \cap \mathbb{V}\left(\ell^{(d)}\right)\right).$$

Not only is each $\mathbb{V}\left(\ell^{(i)}\right)$ a hyperplane, but we also have a solution for this smaller problem,

$$V \cap \mathbb{V}\left(\ell^{(i)}\right) = \mathbb{V}\left(\left[F, L^{(i)}, W^{(i)}\right]\right),$$

which is precomputed in the initial steps of the algorithm.

Proof of correctness of Algorithm 5.3.1. The algorithm works assuming the genericity of all random ingredients, which we do not prove here. add a reference □

Note: One can readily modify the hypersurface intersection procedure to work for a hypersurface represented with a witness set, in particular, in the case when the hypersurface in question is a proper subset of $\mathbb{V}(g)$ (i.e., a union of some but not all irreducible components of $\mathbb{V}(g)$).

5.3.3. Constructing a numerical variety. One major, yet basic, question is: how to pass from a representation of variety $V = \mathbb{V}(F)$ with a polynomial system F to a representation with a witness set?

Algorithm 5.3.2 provides an answer using a cascade of intersections with hypersurfaces, which uses the discussion of possible cases in §5.3.2 and Algorithm 5.3.1.

The numerical variety produced by Algorithm 5.3.2 depends not only on the random choices, but also notably on the order of polynomials in the input polynomial system.

Exercise 5.3.2. *Apply (the ideas of) Algorithm 5.3.2 to the system of three polynomials in the Example 5.1.1 where polynomials are sorted in the following order $F = (f_3, f_1, f_2)$. Making suitable choices of random ingredients, compute* **w**

(1) *at the step when $\mathbb{V}(\mathbf{w}) = \mathbb{V}(f_3)$,*
(2) *at the step when $\mathbb{V}(\mathbf{w}) = \mathbb{V}(f_3, f_1)$, and*
(3) *at the end of the algorithm,*

describing the transitions between these three points in the algorithm.

5.3.4. Intersection. The problem of intersecting two numerical varieties boils down to intersecting two equidimenaional varieties $V_1 = \mathbb{V}(w_1)$ and $V_2 = \mathbb{V}(w_1)$ in $\mathbb{C}^n$ represeented by witness sets $w_i = [F_i, L_i, W_i]$, $i = 1, 2$.

An algorithm to compute a numerical representation for $V_1 \cap V_2$ rests on two simple facts. First,

$$V_1 \cap V_2 \cong (V_1 \times V_2) \cap D, \quad D = \{\, (p, p) \,|\, p \in \mathbb{C}^n \,\} \subset \mathbb{C}^n \times \mathbb{C}^n.$$

Second, the direct product $V_1 \times V_2$ is represented by the witness set

$$w_{1\times 2} = \left[\begin{pmatrix} F_1(x) \\ F_2(y) \end{pmatrix}, L_1(x) \times L_2(y), W_1(x) \times W_2(y) \right],$$

Algorithm 5.3.2 $\mathbf{w} = \text{NumericalVariety}(F)$

Require: $F = (f_1, \dots, f_r)$, a polynomial system in $R = \mathbb{C}[x_1, \dots, x_n]$.
Ensure: $\mathbf{w}$ is a numerical variety such that $\mathbb{V}(F) = \mathbb{V}(\mathbf{w})$.

Create a witness set w representing the hypersurface $\mathbb{V}(f_1)$ using the homotopy discussed in §5.1.2.

$\mathbf{w} \leftarrow w$
for $i = 2$ to r **do**
  $g \leftarrow f_i$
  $\mathbf{w}' \leftarrow \emptyset$
  while $\mathbf{w} \neq \emptyset$ **do**
    Pick $w = [F, L, W] \in \mathbf{w}$.
    $\mathbf{w} \leftarrow \mathbf{w} \setminus \{w\}$
    if $g(p) = 0$ for all $p \in W$ **then**
      $\mathbf{w}' \leftarrow \mathbf{w}' \cup \{w\}$ - *case (1) in §5.3.2*
    else if $W' = \{\, p \in W \mid g(p) = 0 \,\} \neq \emptyset$, but $W' \neq W$ **then**
      $\mathbf{w} \leftarrow \mathbf{w} \cup \{[F, L, W'], [F, L, W \setminus W']\}$ - *case (2)*
    else if $g(p) \neq 0$ for all $p \in W$ **then**
      Pick a random $(n - \dim(\mathbb{V}(w)) + 1)$-plane L'.
      $w' \leftarrow \text{HypersurfaceIntersection}(w, g, L')$
      if $w' \neq \emptyset$ **then**
        $\mathbf{w}' \leftarrow \mathbf{w}' \cup \{w'\}$ - *case (3a)*
      end if
    end if
  end while
  $\mathbf{w} \leftarrow \mathbf{w}'$
end for

where (x, y) are coordinates on $\mathbb{C}^n \times \mathbb{C}^n$.

Now, since the diagonal $D = \mathbb{V}(x_1 - y_1, \dots, x_n - y_n)$, we can simply intersect $V_1 \cap V_2$ with n hyperplanes $x_i - y_i$, $i \in [n]$, obtaining $(V_1 \times V_2) \cap D$ as a numerical variety via Algorithm 5.3.1.

Note: To solve a special problem of intersecting a numerical variety $V_1 = \mathbb{V}(\mathbf{w})$ with a variety $V_2 = \mathbb{V}(g_1, \dots, g_r)$ it is sufficient to employ Algorithm 5.3.1 to compute intersections with hypersurfaces $\mathbb{V}(g_i)$, $i \in [r]$. This does not involve doubling the number of the variables.

The projection $\pi : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}^n$, $\pi(x, y) = x$, induces an isomorphism of $(V_1 \times V_2) \cap D \cong V_1 \cap V_2$. Thus one can, for example, sample points on the variety $V_1 \cap V_2$ by sampling points on $(V_1 \times V_2) \cap D$ and projecting.

Exercise 5.3.3. *Let* $\mathbf{w}$ *be a numerical variety,* $V = \mathbb{V}(\mathbf{w}) \subseteq \mathbb{C}^n \times \mathbb{C}^m$. *Describe an algorithm to detemine whether a point* $p \in \mathbb{C}^n$ *belongs to* $\pi(V) \subseteq \mathbb{C}^n$, *where* π *is the projection to the first* n *coordinates.*

5.3.5. Singular witness sets. Numerical intersection is not the only operation leading to representing a variety as a *projection* of an isomorphic variety. Suppose in the witness set $w = [F, L, W]$, a wintess point $p \in W$ is singular, then tracking a homotopy $H_t^{[F, L \to L', \gamma]}$ in (5.1.1)

starting at p is practically impossible: every point on the continuation path is singular. Fortunately, one can *deflate* a parametric family of systems using methods of §2.2.1.

Example 5.3.4. *Let* $F = \left(x^2\right) \in \mathbb{C}[x,y]^1$ *and* $V = \mathbb{V}(F) = \mathbb{V}(w_t)$, *where* $w_t = [F, L_t, W_t]$ *with*

$$L_t = \mathbb{V}(\ell_t), \quad \ell_t = (1-t)(x+2y) + t(3x+5y+7) = (1+2t)x + (2+3t)y + 7t,$$
$$W_t = \{p_t\}, \quad p_t = (0, -\frac{7}{2+3t}).$$

The witness point p_t *is singular for all* $t \in [0,1]$, *since the jacobian of* $F_t = \left(x^2, \ell_t\right)$,

$$J_t = \frac{\partial F_t}{\partial(x,y)} = \begin{bmatrix} 2x & 0 \\ 1+2t & 2+3t \end{bmatrix}$$

is rank deficient at p_t, *it has rank* $r = 1$. *The deflation in the form (2.2.2) with the identity matrix* B *is*

$$D_B F_t = \begin{pmatrix} F_t \\ J_t \begin{bmatrix} \lambda \\ 1 \end{bmatrix} \end{pmatrix} = \begin{pmatrix} x^2 \\ \ell_t \\ 2x\lambda \\ (1+2t)\lambda + (2+3t) \end{pmatrix} \in \mathbb{C}[x,y,\lambda]^4.$$

One may check that $\frac{\partial D_B F_t}{\partial(x,y,\lambda)}$ *has full rank at* p'_t *for* $t \in [0,1]$, *where* $p'_t \in \mathbb{C}^3$ *is the* lifting *of* p_t *(a unique point projecting to* p_t*).*

Replace $D_B F_t$ *with a square system* $G_t \in \mathbb{C}[x,y,\lambda]^3$ *regular at* p'_t *for* $t \in [0,1]$. *For example, one may use the technique described in §2.1.6; in this example another alternative is to simply drop the polynomial* x^2. *Now instead of following* p_t *for the singular homotopy* $H_t = H_t^{[F, L_0 \to L_1, 1]}$ *we may track* G_t, $t \in [0,1]$.

5.4. Trilingual dictionary

The purpose of the dictionary above is to outline the correspondences between concepts and algorithms in algebra (ideals), geometry (varieties), and numerical algebraic geometry (numerical varieties).

Before giving the dictionary let us fill in the pieces of missing notation

- We denote by $\mathbb{W}(F)$ the output of Algorithm 5.3.2 given a polynomial system F as input: $\mathbb{W}(F)$ is a numerical variety representing the variety $\mathbb{V}(F)$.
- The *numerical intersection* operation $\cap_{\mathbb{W}}$ is discussed in §5.3.4: for numerical varieties $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ it produces a numerical variety denoted $\mathbf{w} \cap_{\mathbb{W}} \widetilde{\mathbf{w}}$ such that
$$\mathbb{V}(\mathbf{w} \cap_{\mathbb{W}} \widetilde{\mathbf{w}}) = \mathbb{V}(\mathbf{w}) \cap \mathbb{V}(\widetilde{\mathbf{w}}).$$
- For a set of points $S \subset \mathbb{C}^n$, we denote by $\mathbb{I}(S)$ the radical ideal obtained by *interpolation* from the points sampled on the variety $V = \overline{S}$.

  Also, $\mathbb{I}(\mathbf{w})$ stands for the interpolating ideal for the set of points $S \subset \mathbb{V}(\mathbf{w})$ obtained by sampling each component of $\mathbb{V}(\mathbf{w})$ using the corresponding witness set of a numerical variety $\mathbf{w}$.

There are strengths and weaknesses in both classical and numerical approaches. For instance, on one hand, computing the intersection of two varieties is easy classically (with ideals) and hard numerically (via witness sets). On the other hand, computing the union of two varieties is hard classically (intersection of ideal), but trivial numerically (union of numerical varieties).

ALGEBRA	GEOMETRY	NUMERICAL ALGEBRAIC GEOMETRY
radical ideal	variety	numerical variety
$\mathbb{I}(V) = \mathbb{I}(\mathbf{w})$	V	$\mathbf{w}$ such that $V = \mathbb{V}(\mathbf{w}) = \bigcup_{w \in \mathbf{w}} \mathbb{V}(w)$
I (not necessarily radical)	$\mathbb{V}(I)$	$\mathbb{W}(F)$ for a finite generating set F of I
inclusion (via membership test)		inclusion/containment test
$I \subseteq J$	$\mathbb{V}(I) \supseteq \mathbb{V}(J)$	containment (via point membership test)
$\mathbb{I}(V) \subseteq \mathbb{I}(V')$	$V \supseteq V'$	$\forall w' \in \mathbf{w}'$, $\exists w \in \mathbf{w}$ such that $\mathbb{V}(w) \supseteq \mathbb{V}(w')$
addition of ideals	intersection of varieties	numerical intersection
$I + J$	$\mathbb{V}(I) \cap \mathbb{V}(J)$	
$\sqrt{\mathbb{I}(V) + \mathbb{I}(V')}$	$V \cap V'$	$\mathbf{w} \cap_{\mathbb{W}} \mathbf{w}' = \bigcup_{w \in \mathbf{w}, w' \in \mathbf{w}'} (w \cap_{\mathbb{W}} w')$
product or intersection of ideals	union of varieties	union of numerical varieties
IJ or $I \cap J$	$\mathbb{V}(I) \cup \mathbb{V}(J)$	
$\sqrt{\mathbb{I}(V)\mathbb{I}(V')}$ or $\mathbb{I}(V) \cap \mathbb{I}(V')$	$V \cup V'$	$\mathbf{w} \cup \mathbf{w}'$
quotient (saturation) of ideals	difference of varieties	difference of numerical varieties
$I : J^\infty$	$\overline{\mathbb{V}(I) \setminus \mathbb{V}(J)}$	
$\mathbb{I}(V) : \mathbb{I}(V')$	$\overline{V \setminus V'}$	$\{w \in \mathbf{w} \,\vert\, \mathbb{V}(w) \not\subset \mathbb{V}(w') \text{ for any } w' \in \mathbf{w}'\}$
elimination of variables	projection of varieties	interpolation from projected points
$\sqrt{I \cap k[x_{m+1}, \ldots, x_n]}$	$\overline{\pi_m(V(I))}$	$\mathbb{I}(\pi_m(S))$ where points S are sampled using $\mathbf{w}$
prime ideal	irreducible variety	irreducible witness set
maximal ideal	point of affine space	approximation of a point of affine space
ascending chain condition (for a chain of ideals)	descending chain condition (for a chain of varieties)	numerical variety is a finite collection of witness sets

CHAPTER 6

Applications

6.1. Robotics

Consider the planar robot arm consisting of three moving links of length ℓ_i, $i = 1, 2, 3$, and a base connected in series by rotational joints. This is a *planar 3R robot*; "R" stands for a rotational joint. A sensor at each joint measures one of the parameters of the *configuration* of the robot, the relative rotation angle θ_i between successive links i and $i-1$, where $i = 0$ corresponds to the base. Assuming the immovable base of the robot is at $(0, 0)$, we denote by (x_i, y_i) the position of the end and by ϕ_i the absolute angle of the i-th link for $i = 1, 2, 3$.

To be completed!!!

FIGURE 1. Planar 3R robot and its kinematic skeleton.

Given the values of the joint angles θ_i, the *forward kinematics problem* is to compute the position of the tip (x_3, y_3) and the absolute angle ϕ_3 for the last link, known as "hand".

The *inverse kinematics problem* is to determine the joint angles θ_i that will place the hand in a desired position and orientation.

The forward kinematics problem admits a straightforward solution. First, relating the absolute angles to the relative ones is easy: for instance, $\phi_3 = \theta_1 + \theta_2 + \theta_3$. Second,

$$x_3 = \ell_1 \cos\phi_1 + \ell_2 \cos\phi_2 + \ell_3 \cos\phi_3 \,;$$
$$y_3 = \ell_1 \sin\phi_1 + \ell_2 \sin\phi_2 + \ell_3 \sin\phi_3 \,.$$

Solving the inverse kinematics problem, given (x_3, y_3, ϕ_3), we readily find

$$x_2 = x_3 - \ell_3 \cos\phi_3 \,;$$
$$y_2 = y_3 - \ell_3 \sin\phi_3 \,.$$

However, finding either (ϕ_1, ϕ_2) or, alternatively, (x_1, y_1) in order to determine the rest of the parameters of the configuration requires solving systems of equations. For example, to find (ϕ_1, ϕ_2) we need to solve a trigonometric system:

$$\begin{pmatrix} \ell_1 \cos\phi_1 + \ell_2 \cos\phi_2 - x_2 \\ \ell_1 \sin\phi_1 + \ell_2 \sin\phi_2 - y_2 \end{pmatrix}. \tag{6.1.1}$$

Introducing $c_i = \cos\phi_i$ and $s_i = \sin\phi_i$ for $i = 1, 2$, we can rewrite (6.1.1) as

$$\begin{pmatrix} \ell_1 c_1 + \ell_2 c_2 - x_2 \\ \ell_1 s_1 + \ell_2 s_2 - y_2 \\ c_1^2 + s_1^2 - 1 \\ c_2^2 + s_2^2 \end{pmatrix}, \tag{6.1.2}$$

which is a square system of polynomials in $\mathbb{R}[c_1, c_2, s_1, s_2]$. The system (6.1.2) is of total degree $4 = 1 \cdot 1 \cdot 2 \cdot 2$, however one expects only two solutions: two circles with centers at (x_0, y_0) and (x_2, y_2) intersect at two points.

Seeking another formulation, we look at a system with (x_1, y_1) as unknowns. In fact, a solution (x_1, y_1) has to be an intersection point of two circles discussed above. Subtracting the first equation of the system

$$\begin{pmatrix} x_1^2 + y_1^2 - \ell_1^2 \\ (x_1 - x_2)^2 + (y_1 - y_2)^2 - \ell_2^2 \end{pmatrix}$$

from the second we get

$$\begin{pmatrix} x_1^2 + y_1^2 - \ell_1^2 \\ 2x_2x_1 + x_2^2 + 2y_2y_1 + y_2^2 + \ell_1^2 - \ell_2^2 \end{pmatrix}. \tag{6.1.3}$$

The system 6.1.3 consists of a quadratic and a linear polynomial in (x_1, y_1) and has at most 2 solutions generically.

Needless to say, for application purposes, we would like to find real solutions to the inverse kinematic problem.

Exercise 6.1.1. *Describe all possible position-orientations $(x_3, y_3, \phi_3) \in \mathbb{R}^3$ the planar 3R robot can reach. This set is called the* workspace *of the robot.*

A robotic mechanism can be presented by a map from the *joint space* J to the *operational space* C. In the case of the planar 3R robot the solution to the forward kinematics problem describes the map M_{ℓ_1,ℓ_2,ℓ_3} sending

$$(\theta_1, \theta_2, \theta_3) \in J = S^1 \times S^1 \times S^1,$$

where S^1 is a circle, to

$$(x_3, y_3, \phi_3) \in C = \mathbb{R} \times \mathbb{R} \times S^1.$$

We say that the robot M_{ℓ_1,ℓ_2,ℓ_3} is at a *kinematic singularity* $a \in J$ (of the inverse kinematics problem) if

$$\ker \frac{\partial M_{\ell_1,\ell_2,\ell_3}}{\partial j}(a) \neq 0,$$

in other words, when a is a nonisolated or singular isolated solution to $M_{\ell_1,\ell_2,\ell_3}(j) = 0$, $j \in J$.

Exercise 6.1.2. *Find kinematic singularities for the planar 3R robot $M_{1,1,2}$.*

Consider two robots M_{ℓ_1,ℓ_2,ℓ_3} and $M_{\ell_1',\ell_2',\ell_3'}$. Let us call a simultaneous configuration of these two robots a handshake if $(x_3', y_3') = (x_3, y_3)$ and $\phi_3' = \pi + \phi_3$.

Exercise 6.1.3. *Set up a polynomial system describing the "handshake".*

- *How many ways are there to shake hands for two (generic) planar 3R robots?*
- *What if we restrict handshake to be horizontal: e.g., fix $\phi_3 = 0$?*
- *How about a horizontal handshake at a fixed height?*

6.2. Automatic theorem proving

6.2.1. Implicitization. First, let us discuss the so-called *implicitization problem* which is a problem of finding equations describing the image of a polynomial map F:

$$\begin{array}{rcl} k^m & \to & k^n \\ t=(t_1,\dots,t_m) & \mapsto & x=(x_1,\dots,x_n)=F(t)=(f_1(t),\dots,f_n(t)). \end{array}$$

In other words, given explicit parametric description of a set (in terms of parameters t_i), we would like to produce a description without the parameters (in terms of x_i).

Example 6.2.1. *Let $F=(f_1,f_2):k\to k^2$ where $f_1(t)=t^2$ and $f_2(t)=t^3$.*

The graph

$$\Gamma(f)=\{\,(t,x_1,x_2)\,|\,x_1=f_1(t),\,x_2=f_2(t)\,\}$$

is defined by the ideal (of polynomial equations)

$$I=\left\langle x_1-t^2,x_2-t^3\right\rangle.$$

Eliminating variable t gives the ideal describing the image of F:

$$J=I\cap k[x_1,x_2]=\left\langle x_1^3-x_2^2\right\rangle.$$

6.2.2. Proving geometric theorems. Most of the theorems in plane geometry proved by mathematicians of ancient Greece can be reproved *automatically* from the first principles and a few algorithmic ideas such as elimination and implicitization.

Example 6.2.2. Heron's formula *for the area S of the triangle with the sides of length a,b,c says*

$$S=\sqrt{p(p-a)(p-b)(p-c)},$$

where $p=\frac{1}{2}(a+b+c)$ is the semiperimeter.

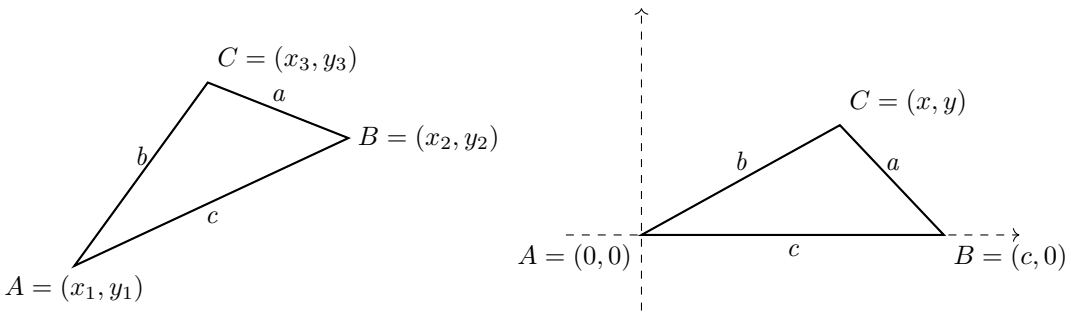


FIGURE 2. Two ways of placing a triangle with sides of lengths a,b,c.

The idea behind deriving this formula automatically consists of two steps

- *Construct a polynomial ideal that describes the relationship between*
  - *coordinates of the vertices A,B,C,*
  - *lengths of sides sides a,b,c, and*
  - *the area S.*
- *Eliminate variables that describe coordinates of A,B,C to get polynomial relations in a,b,c, and S.*

Performing the setup using either of the constructions in Figure 2 and eliminating the intermediate variables, we obtain a principal ideal

$$\left\langle a^4 - 2a^2b^2 + b^4 - 2a^2c^2 - 2b^2c^2 + c^4 + 4S^2 \right\rangle$$

proving that

$$S^2 = -\frac{1}{4}\left(a^4 - 2a^2b^2 + b^4 - 2a^2c^2 - 2b^2c^2 + c^4\right) = p(p-a)(p-b)(p-c).$$

It is impossible to pin an area of a quadrilateral with given side lengths: the quadrilateral is not *rigid*, i.e. the relative positions of vertices can change continuously (there is some "wiggle room"). Thus one can't possibly determine the lengths of diagonals, let alone the area.

One can make a quadrilateral rigid by fixing the length of a diagonal. Another way to make it rigid, is to place all vertices on a circle.

Exercise 6.2.3. *Construct an automatic proof of Heron's formula for a quadrilateral with sides of lengths a, b, c, d inscribed in a circle:*

$$S = \sqrt{(p-a)(p-b)(p-c)(p-d)},$$

where $p = \frac{1}{2}(a+b+c+d)$ is a semiperimeter.

Note that while a quadrilateral inscribed in the circle is rigid, there are two combinatorial configurations: one is convex (the four sides AB, BC, CD, and DA don't intersect) and another is not (either AB and CD intersect or BC and DA intersect). An intrinsically algebraic automatic formula derivation may discover formulas for all possible cases.

Exercise 6.2.4. *It is, in principle, possible to derive automatically formulas for an area of an n-gon inscribed in a circle. What is the largest n for which you can set up an automatic derivation (and, therefore, proof)?*

6.3. Chemical reaction networks

A *chemical reaction network* (CRN) comprises a set of reactants, a set of products and a set of reactions. For simplicity we denote reactants and products by single letters: for example, the combustion reaction for hydrogen,

$$2H_2 + O_2 \to H_2O,$$

could be written as $2A + B \to C$. Moreover, we will always assume there is an inverse reaction in our CRN (it may or may not progress "at rate 0"), so we write

$$2H_2 + O_2 \leftrightharpoons H_2O.$$

The following CRN

$$\begin{aligned} A + B &\leftrightharpoons D \\ B + C &\leftrightharpoons E \\ C &\leftrightharpoons 2A \end{aligned}$$

corresponds to the *dynamical system*

$$
\begin{array}{rcl}
\dot{c}_A &=& c_A^f - c_A - k_{A+B\to D}c_Ac_B + k_{D\to A+B}c_D + 2k_{C\to 2A}c_C - 2k_{2A\to C}c_A^2 \\
\dot{c}_B &=& c_B^f - c_B - k_{A+B\to D}c_Ac_B + k_{D\to A+B}c_D - k_{B+C\to E}c_Bc_C + k_{E\to B+C}c_E \\
\dot{c}_C &=& c_C^f - c_C - k_{B+C\to E}c_Bc_C + k_{E\to B+C}c_E + 2k_{C\to 2A}c_C - 2k_{2A\to C}c_A^2 \\
\dot{c}_D &=& c_D^f - c_D + k_{A+B\to D}c_Ac_B - k_{D\to A+B}c_D \\
\dot{c}_E &=& c_E^f - c_E + k_{B+C\to E}c_Bc_C - k_{E\to B+C}c_E.
\end{array}
\tag{6.3.1}
$$

Here all symbols besides *concentrations* $c_A, \dots, c_E$ are positive real constants: $c_A^f, \dots, c_E^f$ measure the inflow of each reactant; constants k stand for the rates of corresponding reactions.

Exercise 6.3.1. *For each CRN below set up a dynamical system similarly to the example above.*

- $2A + B \leftrightharpoons 3A$
- $A + 3B \leftrightharpoons 3A$
- $A + B \leftrightharpoons C$ *and* $A \leftrightharpoons 2B$
- $A + B \leftrightharpoons D$ *and* $B + C \leftrightharpoons E$ *and* $C \leftrightharpoons A$

An *equilibrium* of this dynamical system is a point $c \in \mathbb{R}^5$ with positive coordinates making the right hand side $F(c)$ of the system (6.3.1) vanish.

Exercise 6.3.2. *Compute the* determinant *of the jacobian $\partial F/\partial c$ of the right hand sides of the resulting dynamical systems in Exercise 6.3.1.*
For which ones of these the coefficients of all terms in $\det(\partial F/\partial c)$ *have the same sign?*

Remark 6.3.3. *Even invertibility of the jacobian on the whole real space does* not *guarantee the injectivity of the function $F(c)$; there is a counterexample*[1] *to the so-called* real Jacobian conjecture. *However, there is a proof of existence and uniqueness of an equilibrium in the case of systems $\dot{c} = F(c)$ coming from CRN that have* $\det(\partial F/\partial c)$ *with all coefficients of the same sign*[2].

Exercise 6.3.4. *Show that if all c^f equal* 1 *then all systems in Exercise 6.3.1 (regardless of the invertibility of the jacobian) have a unique equilibrium.*

6.4. Astrodynamics

Astrodynamics, also know as *orbital mechanics*, aims to solve practical problems concerning the motion of spacecraft.

Consider the *CR3BP (circular restricted three-body problem)*: two heavy bodies with masses $1 - m$ and m, $m \in (0, 0.5]$ are in a *relative equilibrium* (for a 2-body problem), i.e. the distance between them is not changing with time, and the third body has negligible mass (assumed to be zero).

Newton's gravitation governs the dynamics of the system. This means that, in some fixed coordinate frame, the heavy bodies move in circular orbits around their center of mass

[1] Sergey Pinchuk. "A counterexamle to the strong real Jacobian conjecture". In: *Mathematische Zeitschrift* 217 (1994), pp. 1–4.

[2] Gheorghe Craciun and Martin Feinberg. "Multiple equilibria in complex chemical reaction networks: I. The injectivity property". In: *SIAM Journal on Applied Mathematics* 65.5 (2005), pp. 1526–1546.

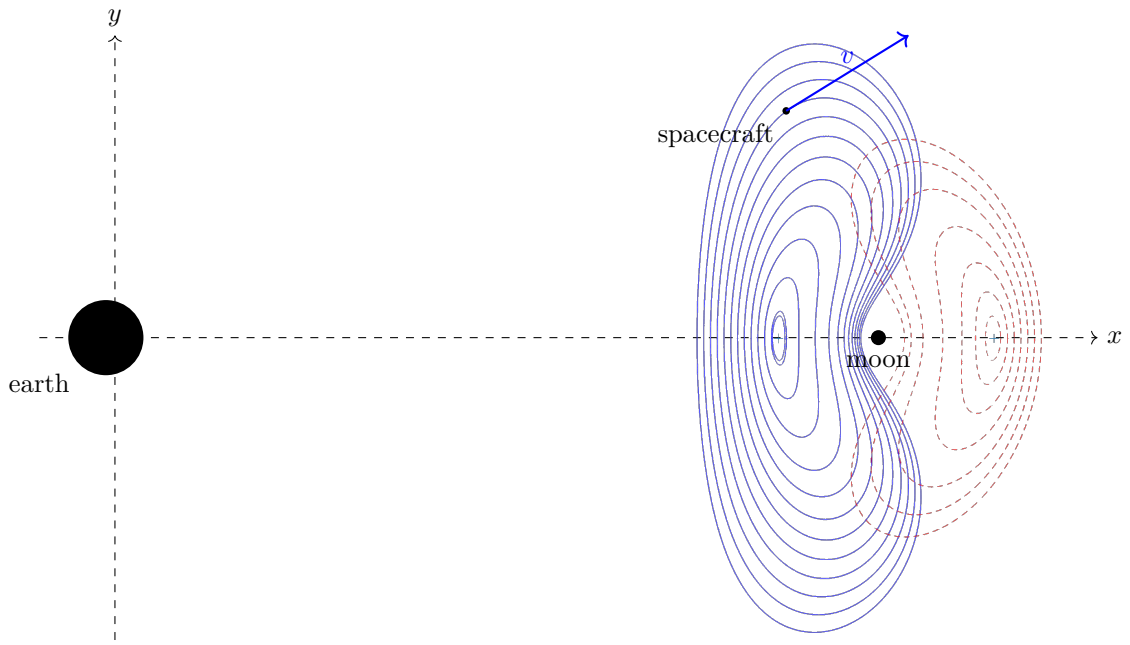


FIGURE 3. Lyapunov orbits in synodic coordinates for a planar CR3BP.

A *synodic coordinate frame* (inroduced by Euler) puts the heavy bodies in the xy-plane with their center of mass at the origin. Further assumptions are made:

- the frame rotates around z-axis so that the heavy bodies are relatively static;
- more precisely, the heavy bodies are placed on the x-axis: the first (with mass $1-m$) at $(-m, 0)$ and the second (with mass m at $(1-m, 0)$;

Finally, we will consider only the *planar* problem:

- the massless body position and velocity vector are both in the xy-plane.

The *dynamical system* describing the motion of the massless body is a system of ODEs for four functions: $x(t), y(t)$, $v_x(t)$, $v_y(t)$. The first two are coordinates in the position vector, the last two in the velocity vector.

Here is the system of ODEs:

$$\begin{aligned}
\dot{x} &= v_x \\
\dot{y} &= v_y \\
\dot{v}_x &= 2v_y + x - \frac{(1-m)(x+m)}{r_1^3} - \frac{m(x-(1-m))}{r_2^3} \\
\dot{v}_y &= -2v_x + y - \frac{(1-m)y}{r_1^3} - \frac{my}{r_2^3}
\end{aligned}$$

where r_1 and r_2 stand for distances from the first and second body:

$$r_1^2 = (x+m)^2 + y^2$$
$$r_2^2 = (x-(1-m))^2 + y^2$$

Exercise 6.4.1. *Set up a polynomial system that is satisfied by an equilibrium point (X,Y), i.e. a point such that $x = X$, $y = Y, v_x = v_y = 0$ make the right-hand side of the system of ODEs evaluates to zero.*
Compute all equilibria points (X,Y) exactly or approximately, or show that there are the variety of these is positive-dimensional.

Exercise 6.4.2. *Develop a simple numerical ODE solver. The main step of the algorithm would take $(x(t_0), y(t_0), v_x(t_0), v_y(t_0))$ and approximate $(x(t_1), y(t_1), v_x(t_1), v_y(t_1))$ for time $t_1 = t_0 + \Delta t$ for a chosen step Δt.*

Jacobi constant (a.k.a. Jacobi integral) is expression

$$C = (x^2+y^2) + \frac{2(1-m)}{r_1} + \frac{2m}{r_2} - (v_x^2 + v_y^2)$$

which happens to be the first integral of our system of ODEs, i.e. it stays constant along a trajectory solving the equations. (Can you prove this?)

Exercise 6.4.3. Lyapunov orbits *are bean-shaped closed (i.e. cyclic) orbits around the equilibria points on the x-axis in the synodic coordinates.*
Modify the numerical solver constructed in Theorem 6.4.2 by adding a correction step that adjusts a prediction at $t = t_1$ given by the old method to produce a nearby point on the orbit (use Jacobi constant).
Compare the method of Theorem 6.4.2 using Lyapunov orbits to the new method: for instance, what is the discrepancy after completing one cycle?

6.5. Signal processing

We consider a problem of estimating position and velocity vectors, r and v, in the plane $\mathbb{R}^2$ of a *transmitter* sending a signal of frequency $f > 0$ into the medium (e.g., air, water, intergalactic space), which is received by n receivers.

Two problems:

(1) Given the position r_i and v_i for the i-th receiver as well as the frequency f_i of the received signal, for $i = 1, \dots, n$; determine r, v, and f.
(2) Same problem, but the transmitter's frequency f is given.

There are n *Doppler effect* (illustrated in Figure 4) equations

$$c^2(f-f_i)^2\|r_i - r\|^2 - f^2\left((r_i - r)\cdot(v_i - v)\right)^2 = 0,$$

where c is the (constant) speed of signal (e.g. sound in air/water or light in the intergalactic space).

Frequency and speed ratios:

$$\frac{f - f_i}{f} = \frac{(v_i - v) \cdot \frac{r_i - r}{\|r_i - r\|}}{c}$$

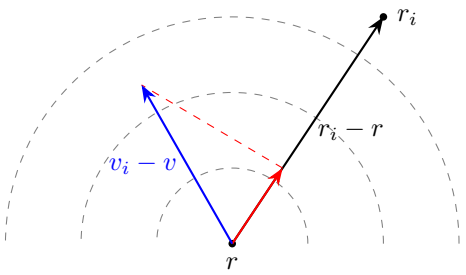


FIGURE 4. Doppler effect with transmitter (r, v) and receiver (r_i, v_i).

Exercise 6.5.1. *Find the number of receivers (i.e. of Doppler equations) necessary to form a system that is a 0-dimensional but not overdetermined (for generic input). Consider both Problems (1) and (2).*

Exercise 6.5.2. *How many complex solutions do systems for Problems (1) and (2) have generically?*
What is the smallest (respectively, largest) number of real solutions you can produce with your choice of input data?

Let us define the exceptional locus E of the problem as the set of inputs in the space of inputs for which some of the output solutions are singular. For instance, for m receivers (where m was determined in Theorem 6.5.1) for Problem (2), this locus E would be a subset of points in $\mathbb{R}^{4m}$ where the i-th receiver contributes 4 coordinates (r_i, v_i).

Exercise 6.5.3. *For both Problems (1) and (2),*

(1) *Show, experimentally, that a generic input is not in the exceptional locus E.*
(2) *Find a point on E. (This would demonstrate that E is not an empty set. Hint: having symmetries may result in singular solutions; for instance, place receivers in a configuration that has some symmetry.)*

6.6. Computer vision

In computer vision one wants to solve the problem of *3D reconstruction* from known arrays of points in $\mathbb{P}^2$ the *views* of some points in a 3D *scene*, an unknown array of points in $\mathbb{P}^3$.

If the problem involves two *calibrated* cameras, it can be rephrased via an *essential matrix* $E \in \mathbb{R}^{3\times 3}$, a matrix that satisfies

$$EE^TE - \frac{1}{2}\mathrm{tr}\,(EE^T)E = 0. \tag{6.6.1}$$

Each pair of *views* $x, y \in \mathbb{R}^3$ representing the image of the same point in the scene in cameras 1 and 2, respectively, imposes a constraint: $x^TEy = 0$.

There are several steps on the way to image reconstruction, but the main one is finding the matrix E.

Exercise 6.6.1. *Determine the number n of points such that the reconstruction problem above is* minimal*: i.e., there are finitely many solutions for E up to scaling given the views of the n* generic *points in two cameras and there are infinitely many solutions for any number smaller than n.*

Remark 6.6.2. *Given the solution set for one generic instance of the minimal reconstruction problem for n points in Exercise 6.6.1 one can construct a homotopy continuation algorithm to solve another instance. It is enough to track a homotopy induced by a line segment in the space of pairs of views connected the input for the solved instance to that of the unsolved.*

APPENDIX A

Homotopy continuation revisited

A.1. Singular solutions revisited

A.1.1. Puiseux series. A complex *Puiseux series* is a fractional power series in one variable of the form

$$x(t) = \sum_{i=i_0}^{\infty} c_i t^{\frac{i}{m}}, \quad i_o, m \in \mathbb{Z},\ m > 0,\ c_i \in \mathbb{C}\,. \tag{A.1.1}$$

A Puiseux series x is said to be *convergent* near the origin if it converges absolutely on a punctured complex disk with the center at the origin.

Note: A Puiseux series (A.1.1) can be written as $x(t) = y(t^{1/m})$ where $y(s)$ is a *Laurent series*. If $x(t)$ is convergent, $y(s)$ is a meromorphic function near the origin (with a possible pole at the origin).

Exercise A.1.1. *Show that Puiseux series (convergent Puiseux series) form a field.*

The field of Puiseux series is often denoted by $\mathbb{C}\{\{t\}\}$.

THEOREM A.1.2 (Newton-Puiseux theorem). *A polynomial equation $f = 0$ for $f \in \mathbb{C}[x,t]$ has a solution $x(t)$ in the field of convergent Puiseux series.*

PROOF. add a reference

Prove this theorem by constructing $x(t)$. Hint: set m to be the multiplicity of $x(t)$ as a root of f. □

A.1.2. Endgame. For $F, G \in \mathbb{C}[x_1, \cdots, x_n]^n$, consider homotopy

$$H_t = \gamma t G + (1-t)F, \quad t \in [0,1],$$

with a generic $\gamma \in \mathbb{C}$ and t starting from $t = 1$ and the target value $t = 0$. (Note that we swapped the ends in comparison to the homotopy considered in the previous section: now the target is $F = H_0$.)

Suppose x_0 is an isolated solution of H_0, then a homotopy path $x(t)$ with $x(0) = x_0$ can be locally, in a neighborhood of $t = t_0$, viewed as a vector in $\mathbb{C}\{\{t\}\}^n$ of the form

$$x(t) = x_0 + \sum_{i=1}^{\infty} c_i t^{\frac{i}{m}}, \quad i \in \mathbb{N},\ c_i \in \mathbb{C}^n\,, \tag{A.1.2}$$

for some integer $m > 0$. The smallest m for which it is possible to write down $x(t)$ in the form (A.1.2) is called the *winding number* of $x(t)$ around $t = 0$.

Note: In fact, as a consequence of the Newton-Puiseux Theorem, every homotopy path $x(t)$ can be viewed as a vector of convergent Puiseux series near $t = 0$. A divergent path has at least one coordinate with a pole at $t = 0$.

Using the form (A.1.2), a path $x(t)$ converging to an isolated $x_0 \in \mathbb{V}(H_0)$ can be viewed as $x(t) = y(t^{1/m})$ where the coordinates of

$$y(s) = x_0 + \sum_{i=1}^{\infty} c_i s^i, \quad i \in \mathbb{N},\ c_i \in \mathbb{C}^n$$

are holomorphic functions in the neighborhood of $t = 0$. The following elementary fact from complex analysis gives rise to a numerical procedure of approximating a singular isolated solution called the *Cauchy endgame*.

THEOREM A.1.3 (Cauchy integral formula). *Suppose U is an open subset of the complex plane $\mathbb{C}$, $f : U \to \mathbb{C}$ is a holomorphic function, and a closed disk $D_\varepsilon \subset U$ of radius $\varepsilon > 0$ centered at 0. Then*

$$f(0) = \frac{1}{2\pi i} \oint_{\partial D_\varepsilon} \frac{f(t)}{t}\, dt\,,$$

where the contour integral is taken counter-clockwise.

PROOF. add a reference □

To approximate x_0 using Cauchy endgame one would need to approximate numerically the integral

$$\frac{1}{2\pi i} \oint_{\partial D_{\varepsilon^{1/m}}} \frac{y(s)}{s}\, ds\,,$$

which one can rewrite as an integral with respect to t: the parameter t shall "wind" m times around 0 along the circle $|t| = \varepsilon$, where m is the winding number for $x(t)$.

Knowing the winding number *a priori* is not necessary. Suppose one starts at $x_\epsilon^{(0)} \in \mathbb{V}(H_\varepsilon)$, a solution on a real path computed by tracking a homotopy H_t numerically to $t = \varepsilon$. Then, as t makes one pass along the circle $|t| = \varepsilon$, the solution to $H_t = 0$ moves to $x_\varepsilon^{(1)} \in \mathbb{V}(H_\varepsilon)$. Repeating, we construct the sequence $x_\varepsilon^{(0)}, x_\varepsilon^{(1)}, x_\varepsilon^{(2)}, \ldots$ which *must* revisit $x_\varepsilon^{(0)}$, since $\mathbb{V}(H_\varepsilon)$ is 0-dimensional.

Proposition A.1.4. *If $V = \{\,(x,t) \in \mathbb{C}^n \times \mathbb{C} \,|\, H_t(x) = 0\,\}$ is an irreducible variety locally at $x_0 \in \mathbb{V}(H_0)$, then*

- *$\mu_{x_0}(H_0)$, the multiplicity of H_0 at x_0 (not defined at the moment),*
- *the number of distinct (real) homotopy paths converging to x_0, and*
- *its winding number*

are all equal.

PROOF. add a reference □

Note: If $x^{(1)}(t), \ldots, x^{(m)}(t)$ are the homotopy paths converging to x_0 as $t \to 0$, their *centroid*

$$x_c(t) = \frac{x^{(1)}(t) + \cdots + x^{(m)}(t)}{m}$$

is a numerical approximation of the Cauchy integral above.
In fact, if $m > 1$, one can show that that the centroid converges to x_0 asymptotically faster than any of the m paths.

APPENDIX B

Certification

B.1. Alpha theory

In this section we go back to the analysis of Newton's method and outline the cornerstone results of Smale's alpha theory. These can be used to show that heuristically obtained approximate solutions are certifiably correct.

B.1.1. Approximate zeros. Let $F \in \mathbb{C}[x]^n$ be a square system of polynomials. For $m \in \mathbb{N}$, let

$$N_F^m(x) \;=\; \underbrace{N_F \circ \cdots \circ N_F(x)}_{m \text{ times}}$$

be the m^{th} Newton iteration of F starting at x. Let $\|\cdot\|$ be the *hermitian norm* on $\mathbb{C}^n$:

$$\|(x_1, \ldots, x_n)\| = (|x_1|^2 + \cdots + |x_n|^2)^{1/2}.$$

A point x is a *approximate zero* of F with the *associated zero* $x^* \in \mathbb{V}(F)$ if

$$\|N_F^m(x) - x^*\| \;\leq\; \left(\frac{1}{2}\right)^{2^m-1} \|x - x^*\|, \tag{B.1.1}$$

for every $m \in \mathbb{N}$. In other words, the sequence $\{N_F^m(x) \mid m \in \mathbb{N}\}$ *converges quadratically* to x^*.

B.1.2. Smale's α-theorem. Smale's α-theory provides sufficient conditions for a given point x to be a approximate zero of F. It operates with the numbers $\alpha(F,x)$, $\beta(F,x)$, and $\gamma(F,x)$ that are defined if the Jacobian $DF(x) = \frac{\partial F}{\partial x}(x)$ is invertible:

$$\beta(F,x) = \|x - N_F(x)\| \;=\; \|DF(x)^{-1}F(x)\|$$

was used before as the absolute backward error estimator,

$$\gamma(F,x) = \sup_{m \geq 2} \left\| \frac{DF(x)^{-1}D^mF(x)}{m!} \right\|^{\frac{1}{m-1}}, \quad \text{(see Remark B.1.2)} \tag{B.1.2}$$

and

$$\alpha(F,x) = \beta(F,x)\gamma(F,x).$$

THEOREM B.1.1. *The point $x \in \mathbb{C}^n$ with*

$$\alpha(F,x) \;<\; \frac{13 - 3\sqrt{17}}{4} \;\approx\; 0.157671 \tag{B.1.3}$$

is a approximate zero of F. Moreover, $\|x - x^*\| \leq 2\beta(F,x)$ where $x^* \in \mathbb{V}(F)$ is the associated zero for x.

PROOF. add a reference □

Remark B.1.2. *In (B.1.2) for $\gamma(f,x)$, the m^{th} derivative $D^mF(x) = \frac{\partial^m F}{\partial x^m}$ has components that are symmetric tensors given by the partial derivatives of F of order m. This is a linear map from $\mathrm{Sym}^m\mathbb{C}^n$ to $\mathbb{C}^n$. The norm in (B.1.2) is the operator norm of $DF(x)^{-1}D^kF(x)$ (a map from $\mathrm{Sym}^m\mathbb{C}^n$ to $\mathbb{C}^n$), defined using the norm on $\mathrm{Sym}^m\mathbb{C}^n$ that is dual to the standard unitarily invariant norm on homogeneous polynomials*

$$\Big\| \sum_{|\sigma|=d} c_\sigma x^\sigma \Big\|_h^2 \;=\; \sum_{|\sigma|=d} \tbinom{d}{\sigma}^{-1} |c_\sigma|^2 \,, \tag{B.1.4}$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is an exponent vector of non-negative integers with $x^\sigma = x_1^{\sigma_1}\cdots x_n^{\sigma_n}$, $|\sigma| = \sigma_1 + \cdots + \sigma_n$, and $\binom{d}{\sigma} = \frac{d!}{\sigma_1!\cdots\sigma_n!}$ is the multinomial coefficient.

One can compute $\gamma(F,x)$ in finite number of steps, since the supremum is taken over a finite number of values m: the derivatives of order higher than the order of polynomials vanish. There is also a simpler procedure that bounds $\gamma(F,x)$ from above.

For a polynomial $g : \mathbb{C}^n \to \mathbb{C}$ define $\|g\| = \|g^h\|_h$, where g^h is the homogenization of g, using the norm in (B.1.4).

For $x \in \mathbb{C}^n$, define $\Delta(x)$ be the $n \times n$ diagonal matrix with

$$(\Delta(x)_{i,i})^2 \;=\; d_i \left(1 + \|x\|^2\right)^{d_i - 1} \,, \quad d_i = \deg f_i \,,$$

and, if $DF(x)$ is invertible, define

$$\mu(f,x) \;=\; \max\{1, \|F\| \cdot \|Df(x)^{-1}\Delta(x)\|\} \,, \quad \|F\|^2 = \sum_{i=1}^n \|f_i^h\|_h^2 \,,$$

where $\|\cdot\|_h$ is the norm in (B.1.4).

Let $D = \max(\deg f_i)$. If $x \in \mathbb{C}^n$ such that $Df(x)$ is invertible, then (according to add a reference)

$$\gamma(f,x) \;\le\; \frac{\mu(f,x) D^{\frac{3}{2}}}{2\sqrt{1+\|x\|^2}} \,. \tag{B.1.5}$$

Exercise B.1.3. *For a polynomial $f = x^2 - 2x + 3$ determine whether the point x passes the α-test (B.1.3) for*

(1) $x = 1$;
(2) $x = 1 + \boldsymbol{i}$;
(3) $x = 1 + \frac{3}{2}\boldsymbol{i}$.

If so, what is the associated zero of the point?

THEOREM B.1.4. *Let $x \in \mathbb{C}^n$ with $\alpha(F,x) < 0.03$ and $x^* \in \mathbb{V}(f)$ the associated zero for x. If $y \in \mathbb{C}^n$ satisfies*

$$\|x - y\| \;<\; \frac{1}{20\gamma(F,x)} \,, \tag{B.1.6}$$

then y is a approximate zero of F with associated solution x^*.

PROOF. add a reference □

Exercise B.1.5. *For a polynomial $f = x^2 - 1$ find an upper bound on $\varepsilon > 0$ such that $x = 1 + \varepsilon$ and $y = 1 - \varepsilon$ pass the* robust α-test*, i.e., satisfy the hypotheses of Theorem B.1.4.*

Exercise B.1.6. *Let $f \in \mathbb{R}[x]$ be a polynomial with* real *coefficients.*
Design a procedure that given an approximate zero x of f determines if it is associated to a real zero $x^* \in \mathbb{V}(f)$. (Hint: The conjugate $\overline{x^*} \in \mathbb{V}(f)$.)

In addition to certification of an approximate zero, certified homotopy continuation is possible in case the homotopy H_t is regular: i.e., it is possible to ensure that *path jumping* does not happen (two paths "jump" in Figure 1).

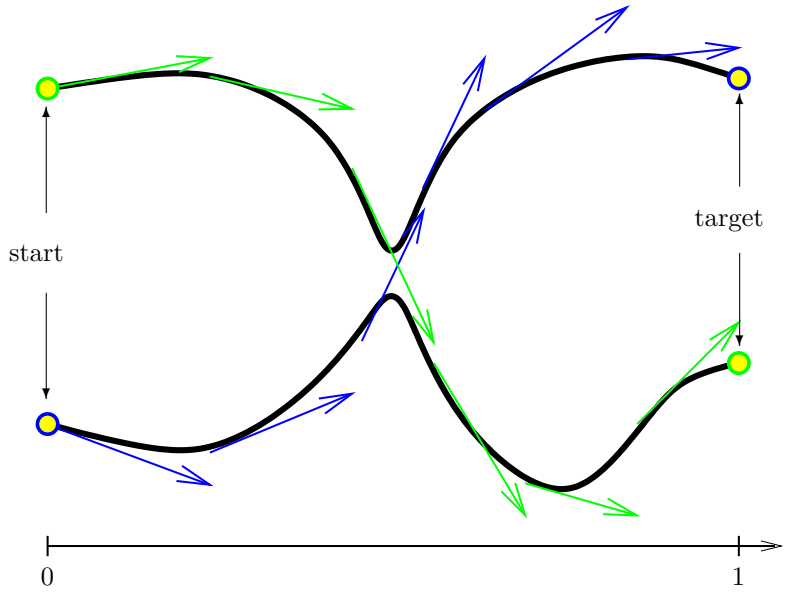


FIGURE 1. Path–crossing scenario: all target solutions may be reached, however the *order* of the solutions is incorrect.

The path-jumping may lead to obtaining an incomplete set of target solutions. In the *path-crossing* scenario of Figure 1, a more subtle piece of information is lost: the information about which path leads to which target solution.