Chapter 4

Numerical Algebraic Geometry

Outline:

Chapter 2 presented fundamental symbolic algorithms, including the classical resultant and general methods based on Gröbner bases. Those algorithms operate on the algebraic side of the algebraic-geometric dictionary underlying algebraic geometry. This chapter develops the fundamental algorithms of numerical algebraic geometry, which uses tools from numerical analysis to represent and study algebraic varieties on a computer. As we will see, this field primarily operates on the geometric side of our dictionary. Since numerical algebraic geometry involves numerical computation, we will work over the complex numbers.

4.1 Core Numerical Algorithms

Numerical algebraic geometry rests on two core numerical algorithms, which go back at least to Newton and Euler. Newton's method refines approximations to solutions to systems of polynomial equations. A careful study of its convergence leads to methods for certifying numerical output, which we describe in Section 4.5. The other core algorithm comes from Euler's method for computing a solution to an initial value problem. This is a first-order iterative method, and more sophisticated higher-order methods are used in practice for they have better convergence. These two algorithms are used together for path-tracking in numerical homotopy continuation, which we develop in subsequent sections as a tool for solving systems of polynomial equations and for manipulating varieties on a computer. While these algorithms are standard in introductory numerical analysis, they are less familiar to algebraists. Our approach is intended to be friendly to algebraists.

By the Fundamental Theorem of Algebra, a univariate polynomial $f(x)$ of degree n has n complex zeroes. When the degree of f is at most four, there are algorithmic formulas for these zeroes that involve arithmetic operations and extracting roots. Zeroes of linear polynomials go back to the earliest mathematical writing, such as the Rhind Papyrus, and the Babylonians had a method for the zeroes of quadratic polynomials that is the precursor of the familiar quadratic formula, which was made explicit by Bramagupta c. 600 CE. The 16th century Italians del Ferro, Tartaglia, Cardano, and Ferrari colorfully gave formulas for the zeroes of cubic and quartic polynomials. It was only in 1823 that Niels Hendrik Abel proved there is no universal such formula for the zeroes of polynomials of degree five and higher.

It was later shown that the zeroes of a general polynomial cannot be expressed in terms of the coefficients using only arithmetic operations on its coefficients and extracting roots. For example, the zeroes of this sextic

$$f \;:=\; x^6 \;+\; 2x^5 \;+\; 3x^4 \;+\; 5x^3 \;+\; 7x^2 \;+\; 11x \;-\; 13 \tag{4.1}$$

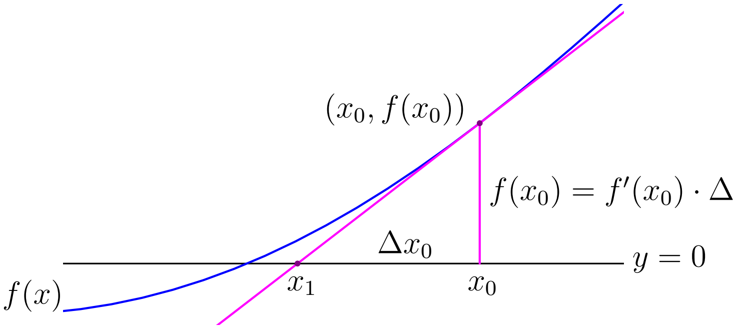
admit no such expression. This is not to say there is no formula for the zeroes of polynomials of degree five or more. For example, there are hypergeometric power series for those zeroes that depend upon the coefficients.

Numerical methods offer another path to the roots of univariate polynomials. While numerical linear algebra may be combined with the eigenvalue approaches to solving of Section 2.5, we will discuss algorithms based on Newton's method.

Newton's method uses the tangent-line approximation to the graph of a differentiable function $f(x)$ to refine an approximation x_0 to a zero of f. The tangent line to the graph of $f(x)$ at the point $(x_0, f(x_0))$ has equation

$$y \;=\; f'(x_0)(x-x_0) \;+\; f(x_0)\,.$$

If $f'(x_0) \neq 0$, we solve this for $y=0$ to get the formula $x_1 := x_0 - (f'(x_0))^{-1}f(x_0)$ for the refinement of x_0. This may also be read from the graph.



Using operator notation Df for the derivative of f, we obtain the expression

$$N_f(x) \;:=\; x \;-\; (Df(x))^{-1}f(x) \tag{4.2}$$

for the passage from x_0 to x_1 above. This *Newton iteration* (4.2) is the basis for Newton's method to compute a zero of a function $f(x)$:

- Start with an initial value x_0, and
- While $Df(x_i) \neq 0$, compute the sequence $\{x_i \mid i \in \mathbb{N}\}$ using the recurrence $x_{i+1} = N_f(x_i)$.

For $f(x) = x^2 - 2$ with $x_0 = 1$, if we compute the first seven terms of the sequence,

$$\begin{aligned}
x_1 &= 1.5\\
x_2 &= 1.416\overline{6}\\
x_3 &= 1.4142156862745098039\overline{2156862745098039}\\
x_4 &= 1.41421356237468991062629557889013491011655962211574404458490501920000\\
x_5 &= 1.414213562373095048801689623502530243614981925776197428498289498623\,1\\
x_6 &= 1.41421356237309504880168872420969807856967187537723400156101313311132\\
x_7 &= 1.4142135623730950488016887242096980785696718753769480731766797379907\,,
\end{aligned}$$

then the 58 displayed digits of x_7 are also the first 58 digits of $\sqrt{2}$.

This example suggests that Newton's method may converge rapidly to a solution. It is not always so well-behaved. For example, if $f(x) = x^3 - 2x + 2$, then $N_f(0) = 1$ and $N_f(1) = 0$, and so the sequence $\{x_i\}$ of Newton iterates with $x_0 = 0$ is periodic with period 2, and does not converge to a root of f.

In fact Newton's method is about as badly behaved as it can be, even for polynomials. The *basin of attraction* for a zero x^* of f is the set of all complex numbers x_0 such that Newton's method starting at x_0 converges to x^*. In general, the boundary of a basin of attraction is a fractal Julia set. Figure 4.1 shows basins of attraction for two univariate cubic polynomials. On the left are those for the polynomial $f(x) = x^3 - 1$, in the region $|\Re(x)|, |\Im(x)| \leq 1.3$. This polynomial vanishes at the cubic roots of unity, and each is a fixed point of a Newton iteration. There are three basins, one for each root of f, and their union is dense in the complex plane. Each basin is drawn in a different color.

On the right of Figure 4.1 are basins for the polynomial $f(x) = x^3 - 2x + 2$. The roots of f give three fixed points of a Newton iteration, and we noted there is an orbit of period 2. The roots and the orbit each have a basin of attraction and each basin is drawn in a different color. The basin of attraction for the orbit of period 2 is in red.

Despite this complicated behaviour, Newton's method is a foundation for numerical algebraic geometry, and it may be used to certify the results of numerical computation. To understand why this is so, we investigate its convergence.

Suppose that f is twice continuously differentiable in a neighborhood of a zero ζ of f and that $Df(\zeta)^{-1}$ exists. Differentiating the expression (4.2) for $N_f(x)$ gives

$$\begin{aligned}
DN_f(x) &= 1 - Df(x)^{-1}Df(x) + Df(x)^{-1}D^{(2)}f(x)Df(x)^{-1}f(x)\\
&= Df(x)^{-1}D^{(2)}f(x)Df(x)^{-1}f(x)\,.
\end{aligned}$$

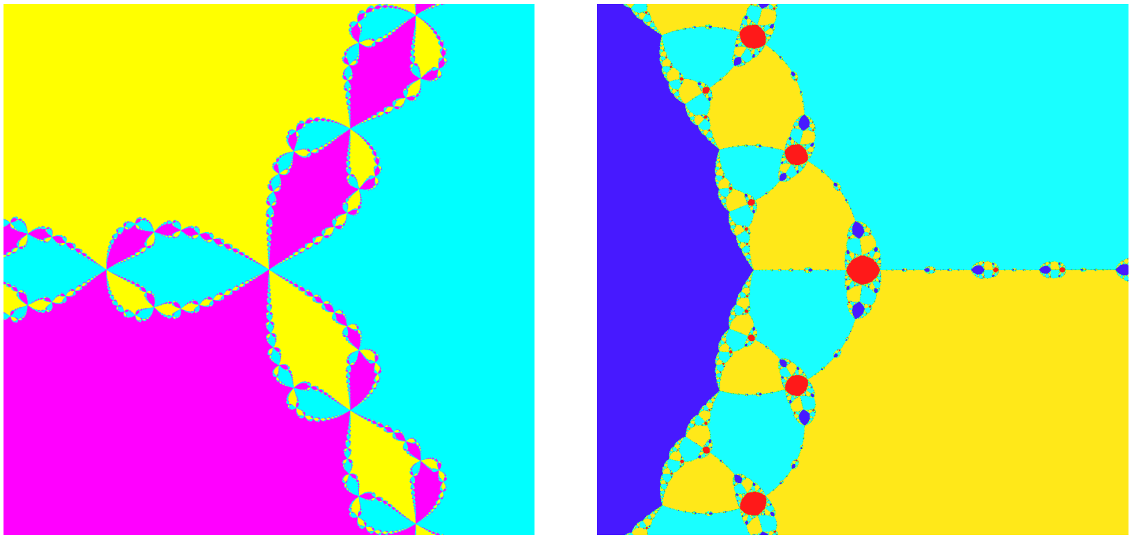


Figure 4.1: Basins of attraction for two cubics

As $f(\zeta) = 0$, we have $DN_f(\zeta) = 0$ and $N_f(\zeta) = \zeta$. The first order Taylor expansion of $N_f(x)$ about the point $x = \zeta$ with remainder is then

$$N_f(x) \ = \ \zeta \ + \ \frac{1}{2}D^{(2)}N_f(a)(x-\zeta)^2\,,$$

where a is some point with $|a-\zeta| < |x-\zeta|$, and we used that $N_f(\zeta) = \zeta$. If we set $c := \max\{\frac{1}{2}|D^{(2)}N_f(a)| \mid |a-\zeta| < r\}$, for some $r > 0$, then we have

$$|N_f(x) - \zeta| \ \leq \ c|x-\zeta|^2\,,$$

whenever $|x-\zeta| \leq r$. Suppose now that $|x-\zeta| < \min\{\frac{1}{2c}, r\}$. Then,

$$\begin{aligned} |N_f(x) - \zeta| \ &< \ c\cdot\left(\frac{1}{2c}\right)^2 \ = \ \frac{1}{4c}\,, \qquad \text{and} \\ |N_f^2(x) - \zeta| \ &< \ c\cdot\left(\frac{1}{4c}\right)^2 \ = \ \frac{1}{16c}\,, \end{aligned}$$

and in general, if $N_f^i(x)$ is the ith iteration of N_f applied to z, we have

$$|N_f^i(x) - \zeta| \ < \ 2^{1-2^i}\frac{1}{2c}\,.$$

This implies that the number of digits of ζ that have been computed in x_{i+1} (the number of *significant digits* in x_{i+1}) will be approximately twice the number of significant digits in x_i. We saw this when computing $\sqrt{2}$ using Newton's method as $x_1, x_2, \ldots, x_6$ had 1, 3, 6, 12, 24, and 48 correct decimal digits of $\sqrt{2}$.

This has a straightforward extension to the problem of approximating zeroes to a square system of multivariate polynomials,

$$F \,:\, f_1 \;=\; f_2 \;=\; \cdots \;=\; f_n \;=\; 0\,, \tag{4.3}$$

where each f_i is a polynomial in n variables. Here *square* means that the number of equations equals the number of variables and (4.3) defines a zero-dimensional variety. It is useful to consider F to be a polynomial map,

$$F \;=\; (f_1,\dots,f_n) \,:\, \mathbb{C}^n \;\longrightarrow\; \mathbb{C}^n\,,$$

and write the solutions $\mathcal{V}(F)$ as $F^{-1}(0)$. Given a point $x\in\mathbb{C}^n$ where the Jacobian matrix $DF(x)$ of partial derivatives of $f_1,\dots,f_n$ with respect to the components of x is invertible, the *Newton iteration* of F applied to x is

$$NF(x) \;:=\; x - DF(x)^{-1}F(x)\,.$$

The geometry of this map is the same as for univariate polynomials: $NF(x)$ is the unique zero of the linear approximation of F at x. This is the solution $\zeta\in\mathbb{C}^n$ to

$$0 \;=\; F(x) + DF(x)(\zeta - x)\,.$$

The elementary analysis of iterating Newton steps is also the same. As before, Newton steps define a chaotic dynamical system on $\mathbb{C}^n$, but if x is sufficiently close to a zero ζ of F, then Newton iterations starting at x converge rapidly to ζ.

We quantify this. A sequence $\{x_i \mid i\in\mathbb{N}\}\subset\mathbb{C}^n$ *converges quadratically* to a point $\zeta\in\mathbb{C}^n$ if for all $i\in\mathbb{N}$,

$$\|x_i-\zeta\| \;\leq\; 2^{1-2^i}\|x_0-\zeta\|\,. \tag{4.4}$$

For example, the sequence of Newton iterations for x^2-2 beginning with $x_0=1$ that we computed converges quadratically to $\sqrt{2}$. A point $x\in\mathbb{C}^n$ is an *approximate zero* of F with *associated zero* $\zeta\in\mathbb{C}^n$ if $F(\zeta)=0$ and if the sequence of Newton iterates defined by $x_0:=x$ and $x_{i+1}:=NF(x_i)$ for $i\geq 0$ converges quadratically to ζ.

Knowing an approximate zero x to a system F is a well-behaved relaxation of the problem of knowing its associated zero ζ. Indeed, the sequence of Newton iterations starting with x reveals as many digits of ζ as desired, in a controlled manner.

At this point, one should ask for methods to determine if x is an approximate zero to F. A heuristic that is used in practice is to treat the length of a Newton step

$$\beta(F,x) \;:=\; \|x-NF(x)\| \;=\; \|DF(x)^{-1}F(x)\|\,, \tag{4.5}$$

as a proxy for the distance $\|x-\zeta\|$ to the zero ζ of F. If we are in the basin of quadratic convergence to ζ, then we expect that

$$\tfrac{1}{2}\|x-\zeta\| \;\lesssim\; \beta(F,x) \;\lesssim\; \|x-\zeta\|\,.$$

For the heuristic, compute two Newton iterations, $NF(x)$ and $NF^2(x)$. If we have that

$$\beta(F, NF(x)) \;\leq\; \beta(F,x)^2\,, \tag{4.6}$$

with $\beta(F,x)$ below some threshold $\beta \ll 1$, then we replace x by $NF^2(x)$, and declare it to be an approximate zero of F with some certitude. The condition (4.6) implies that the number of digits common to $NF(x)$ and $NF^2(x)$ is at least twice the number of digits common to x and $NF(x)$. See Exercise 5 for potential limitations of this heuristic.

We can do much better than this heuristic. Smale studied the convergence of Newton's method and developed what is now called *α-theory* after a computable constant $\alpha = \alpha(F,x)$ such that if α is sufficiently small, then x is an approximate zero of F. We present this theory in Section 4.5. For now, we define $\alpha(F,x)$ and state Smale's theorem.

The constant $\alpha(F,x)$ is the product of two numbers. The first is the length $\beta(F,x)$ (4.5) of a Newton step at x. For the second, recall the Taylor expansion of F at x, Relate this to /eqrefEq:Taylor.

$$F(w) \;=\; F(x) + DF(x)(w-x) + D^2F(x)(w-x)^2 + \cdots + D^NF(x)(w-x)^N\,,$$

where the polynomial map F has degree N. Let is describe the meaning of the terms in this Taylor expansion. For $v \in \mathbb{C}^n$, v^k is the symmetric tensor indexed by all exponents $a \in \mathbb{N}^n$ of degree k, where

$$(v^k)_a \;=\; \tfrac{1}{a_1!}\tfrac{1}{a_2!}\cdots\tfrac{1}{a_n!}v_1^{a_1}v_2^{a_2}\cdots v_n^{a_n} \;=:\; \tfrac{1}{a!}v^a\,.$$

Let $S^k\mathbb{C}^n$ be this vector space of symmetric tensors. Writing $x = (x_1,\dots,x_n) \in \mathbb{C}^n$, the ith component of $D^kF(x)$ is the vector of partial derivatives of F_i of order k,

$$(D^kF_i(x))_a \;=\; \left(\tfrac{\partial}{\partial x_1}\right)^{a_1}\left(\tfrac{\partial}{\partial x_2}\right)^{a_2}\cdots\left(\tfrac{\partial}{\partial x_n}\right)^{a_n}F_i(x)\,,$$

for $a \in \mathbb{N}^n$ of degree k. Then the ith component of $D^kF(x)(w-x)^k$ is the sum

$$\sum_{|a|=k} D^aF_i(x)\tfrac{1}{a!}(w-x)^a\,.$$

Thus $D^kF(x)$ is a linear map from the space $S^k\mathbb{C}^n$ of symmetric tensors to $\mathbb{C}^n$. The same is true for the composition $DF(x)^{-1}\circ D^kF(x)$. If we use the standard norm for vectors $z \in \mathbb{C}^n$ and $v \in S^k\mathbb{C}^n$,

$$\|z\| \;:=\; \left(\sum_{i=1}^n |z_i|^2\right)^{1/2} \qquad\text{and}\qquad \|v\| \;:=\; \left(\sum_{|a|=k}|v_a|^2\right)^{1/2},$$

then the operator norm of this composition is

$$\left\|DF(x)^{-1}\circ D^kF(x)\right\| \;:=\; \max_{\|v\|=1}\left\|DF(x)^{-1}\circ D^kF(x)(v^k)\right\|\,.$$

With these definitions, set

$$\gamma(F,x) \ := \ \max_{k\geq 2} \tfrac{1}{k!} \left\| DF(x)^{-1} \circ D^k F(x) \right\|^{\frac{1}{k-1}} ,$$

and then define

$$\alpha(F,x) \ := \ \beta(F,x) \cdot \gamma(F,x) \,.$$

Theorem 4.1.1. *If $\alpha(F,x) < \frac{1}{4}(13-3\sqrt{17}) \simeq 0.15767\ldots$, then x is an approximate zero of F. The distance from x to its associated zero is at most $2\beta(F,x)$.*

We prove Theorem 4.1.1 and some extensions in Section 4.5. Note the similarity between the formula for $\gamma(F,x)$ and the root test/formula for the radius of convergence of a power series. The shift in the radical from $\frac{1}{k}$ to $\frac{1}{k-1}$ is because this is applied to the expansion of DF.

We apply this to the quadratic polynomial $f(x) = x^2 - 2$. When $x \in \mathbb{C}$ is non-zero, $f'(x) \neq 0$, and we have

$$\beta(f,x) \ = \ \left| \frac{f(x)}{f'(x)} \right| \qquad \text{and} \qquad \gamma(f,x) \ = \ \frac{1}{2} \left| \frac{f''(x)}{f'(x)} \right| \ .$$

Then $\alpha(f,x) = \frac{1}{2}|f(x)f''(x)/(f'(x))^2| = |\frac{x^2-2}{4x^2}|$.

Observe that $\alpha(f,1) = \frac{1}{4} > \frac{1}{4}(13-3\sqrt{17})$. Thus, while Newton iterations starting at $x_0 = 1$ converge quadratically to $\sqrt{2}$, this quadratic convergence is not detected with Smale's α-theory. Note that $\alpha(f,3/2) = \frac{1}{36} < \frac{1}{4}(13-3\sqrt{17})$, so that α-theory certifies the quadratic convergence of Newton iterations starting with $x_0 = 3/2$.

Consider the regions of convergence of Newton's method for the zeroes of x^2-2 in the complex plane. First, when $\Re(x) > 0$, Newton iterations beginning with x converge to $\sqrt{2}$ and when $\Re(x) < 0$, iterations beginning with x converge to $-\sqrt{2}$. In Figure 4.2, a point $x \in \mathbb{C}$ is yellow if Newton iterations converge quadratically, and magenta otherwise. The zeroes $\pm\sqrt{(2)}$ are as indicated and the convex regions enclosing them indicate the points whose quadratic convergence is certified using α-theory ($\alpha(f,z) < \frac{1}{4}(13-3\sqrt{17})$).

On the positive real line, the interval of quadratic convergence is $(\sqrt{2}/2, 3\sqrt{2})$, while the interval where $\alpha(f,x) < \frac{1}{4}(13-3\sqrt{17})$ is $(1.10746, 2.32710)$, which is much smaller.

Euler's method was developed to solve the initial value problem for a first-order ordinary linear differential equation,

$$y' \ = \ f(x,y)\,, \qquad y(x_0) = y_0\,,$$

where the function $f(x,y)$ is continuous near (x_0,y_0) in $\mathbb{R}^2$. Given a *stepsize* $h > 0$, Euler's method approximates the value of $y(x)$ at $x_1 = x_0 + h$ using the linear approximation given by the differential equation, $y_1 := y_0 + hf(x_0,y_0)$. Euler's method recursively computes a sequence $\{(x_i,y_i) \mid i = 0,\ldots,N\}$ of points where

$$x_{i+1} \ := \ x_i + h \quad \text{and} \quad y_{i+1} \ := \ y_i + hf(x_i,y_i)\,.$$

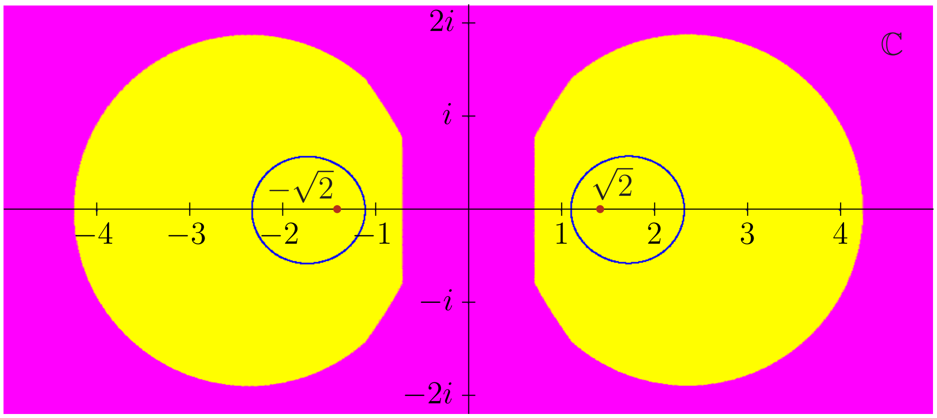
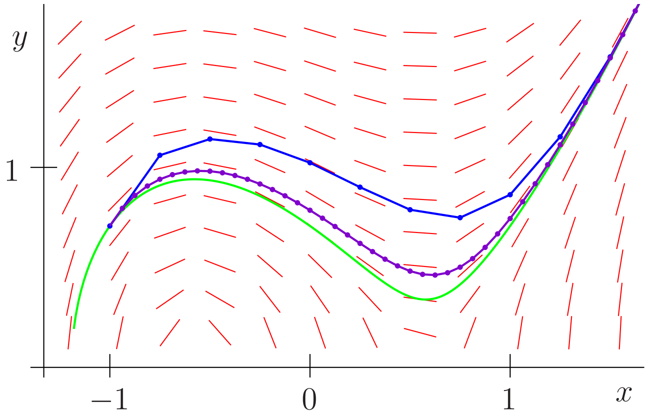


Figure 4.2: Basins of quadratic convergence for $x^2 - 2$.

Let us consider the initial value problem,

$$y' = \frac{3x^2 - 1}{2y}\,, \qquad y(-1) = \frac{1}{\sqrt{2}}\,. \tag{4.7}$$

The solution to this initial value problem is $y = \sqrt{x^3 - 3 + \frac{1}{2}}$. The picture below shows the slope field and the solution curve, and two approximations using Euler's method starting at $(-1, \frac{1}{\sqrt{2}})$ with respective stepsizes $h = \frac{1}{4}$ and $h = \frac{1}{16}$.



Like Newton's method, Euler's method extends to solving the intial value problem for a system of first order linear differential equations, so that y is a vector.

Let us investigate the accuracy of Euler's method, assuming that y (and hence $f(x, y)$) has sufficiently many derivatives. The second order Taylor expansion of $y(x)$ at $x = x_0$, together with the differential equation $y'(x) = f(y, x)$ gives

$$y(x + h) \;=\; y_0 + hf(x_0, y_0) + \tfrac{h^2}{2}y''(x_0) + \tfrac{h^3}{6}y'''(x^*)\,,$$

where x^* lies between x_0 and $x_0 + h$. We estimate

$$|y(x+h) - (y_0 + hf(x_0, y_0))| \;\leq\; h^2\left(\tfrac{1}{2}|y''(x_0)| + \tfrac{|h|}{6}|y'''(x^*)|\right).$$

When y''' is bounded near (x_0, y_0), the error in a single step of Euler's method is at most a constant mutiple of h^2.

To compute $y(x)$ for a fixed x using Euler's method, choose a stepsize h and then perform $\lceil |x - x_0|/h \rceil$ iterations. Each iteration introduces an error at most a constant multiple of h^2 so the difference between $y(x)$ and the computed value will be at most a constant multiple of h. That the global error is at most a constant multiple of the stepsize h, marks Euler's method as a *first-order* iterative method.

The primary value of Euler's method is to illustrate the idea of an iterative solver for initial value problems, as first order accuracy is typically insufficient. A simple higher-order method is the *midpoint rule*, in which the successive values of y are computed using the more complicated formula

$$x_{i+1} = \; x_i + h\,, \quad \text{and} \quad y_{i+1} = \; y_i + hf\left(x_i + \tfrac{1}{2}h\,,\, y_i + \tfrac{1}{2}hf(x_i, y_i)\right)\,,$$

which is second-order.

The classical *Runge-Kutta* method, also called RK4, is a common fourth-order method. For this,

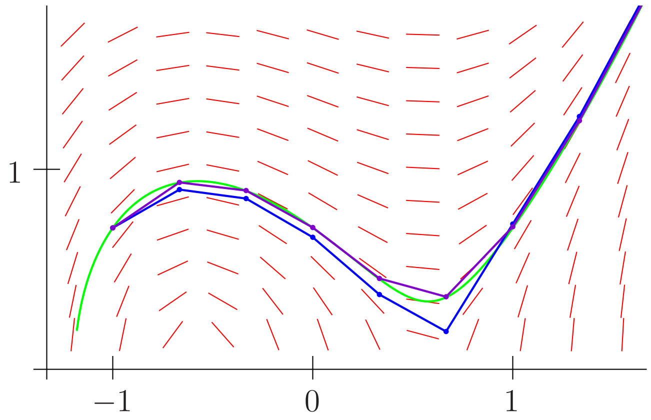
$$x_{i+1} = \; x_i + h\,, \quad \text{and} \quad y_{i+1} = \; y_i + \tfrac{1}{6}h(z_1 + 2z_2 + 2z_3 + z_4)\,,$$

where

$$\begin{aligned} z_1 &= f(x_i, y_i)\,, & z_2 &= f(x_i + \tfrac{1}{2}h\,,\, y_i + \tfrac{1}{2}hz_1)\,, \\ z_3 &= f(x_i + \tfrac{1}{2}h\,,\, y_i + \tfrac{1}{2}hz_2)\,, \text{ and} & z_4 &= f(x_i + h\,,\, y_i + hz_3)\,. \end{aligned}$$

In Exercise 9 you are asked to relate this to Simpson's rule.

We display the two piecewise linear curves obtained from the midpoint and Runge-Kutta methods with stepsize $\frac{1}{3}$ for the initial value problem (4.7) with solution $y = \sqrt{x^3 - 3 + \frac{1}{2}}$ on the same slope field as we used to illustrate Euler's method.



There is a vast literature on iterative methods for solving ordinary differential equations.

Exercises

1. Consider a depressed cubic equation, one of the form $x^3 + bx = c$. Show that

$$x \;=\; \sqrt[3]{\tfrac{c}{2} + \sqrt{\tfrac{c^2}{4} + \tfrac{b^3}{27}}} \;+\; \sqrt[3]{\tfrac{c}{2} - \sqrt{\tfrac{c^2}{4} + \tfrac{b^3}{27}}}$$

   is a solution. What are the other two solutions? How about the solutions to a general cubic equation $\alpha x^3 + \beta x^2 + \gamma x + \delta = 0$?

2. Prove the assertion about (4.1) using Galois theory. Specifically show that the polynomial f has Galois group the full symmetric group S_6 by factoring f modulo sufficiently many primes, and use lifts of Frobenius elements.

3. Consider the following iterative algorithm used by the Babylonians to compute $\sqrt{x}$ for $x > 0$. Observe that if $x_i > 0$ and $x_i \neq \sqrt{x}$, then the interval with endpoints x_i and x/x_i contains $\sqrt{x}$ in its interior. Set $x_{i+1} = \frac{1}{2}(x_i + \frac{x}{x_i})$ and repeat. Compare this method of computing square roots to Newton's Method.

4. Let $f(x) = x^3 - 2x + 2$ and compute some iterates of Newton's method beginning with the following values of x_0,

$$x_0 \;\in\; \{0.1\,,\; 0.9\,,\; \tfrac{1}{2} + \frac{1}{10}\sqrt{-1}\,,\; 1 - \frac{1}{10}\sqrt{-1}\,,\; -1\}\,.$$

5. Newton's method for $f(x) = x^2 - 2$ converges quadratically for x_0 in the interval $[\sqrt{2}/2, 3\sqrt{2}]$. Prove that Newton iterations beginning with these endpoints converge quadratically. Investigate the failure of quadratic convergence for $x > 3\sqrt{2}$: at which step of Newton's method does quadratic convergence fail (the condition (4.4) does not hold) for each of the following starting points for Newton's method.

$$5.2\,,\; 4.53\,,\; 4.36\,,\; 4.298\,,\; 4.256\,,\; 4.25\,,\; 4.246\,,\; 4.245\,,\; 4.2427\,.$$

6. Prove that the midpoint rule is a second-order method.

7. Give some examples of Euler's method. If you cannot find something more interesting, start with the exponential function. Have them study its global convergence and the dependence on stepsize.

8. Compare Euler, midpoint, and Runge-Kutta on some examples.

9. Have the students prove the equivalence of Runge-Kutta with Simpson's rule.

4.2 Numerical Homotopy Continuation

The core numerical algorithms introduced in Section 4.1—Newton's method to refine an approximation to a solution of a system of polynomial equations and iterative methods to solve an initial value problem—are the building blocks of higher-level predictor-corrector methods that track smooth implicitly defined paths. Numerical homotopy continuation uses path tracking to compute the zeroes of a system of polynomials, given the zeroes of a different, but related system. The Bézout Homotopy Algorithm, which is based on numerical homotopy continuation, computes the isolated zeroes of any system of multivariate polynomials and is optimal for generic systems. Algorithms based on numerical homotopy continuation have the added virtue of being inherently parallelizable.

To motivate and illustrate these ideas, consider a toy problem. Suppose we want to compute the (four) solutions to the system of equations

$$z(y-1)-1 \;=\; y^2+2z^2-9 \;=\; 0\,. \tag{4.8}$$

Consider instead the system

$$z(y-1) \;=\; y^2+2z^2-9 \;=\; 0\,, \tag{4.9}$$

whose solutions are found by inspection to be

$$(\pm 3, 0) \qquad \text{and} \qquad (1, \pm 2)\,. \tag{4.10}$$

Figure 4.3 shows the plane curves defined by the polynomials in these systems. The first system (4.8) seeks the intersection of the hyperbola with the ellipse, while the second, simpler, system (4.9) replaces the hyperbola by the two lines.

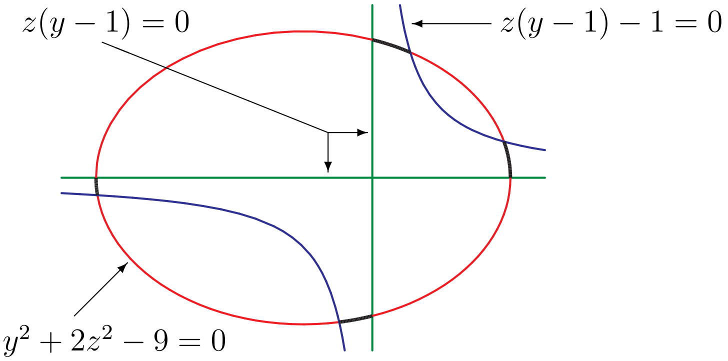


Figure 4.3: The intersection of a hyperbola with an ellipse.

These systems are connected by the one-parameter (in t) family of systems

$$z(y-1)-(1-t) \;=\; y^2+2z^2-9 \;=\; 0\,. \tag{4.11}$$

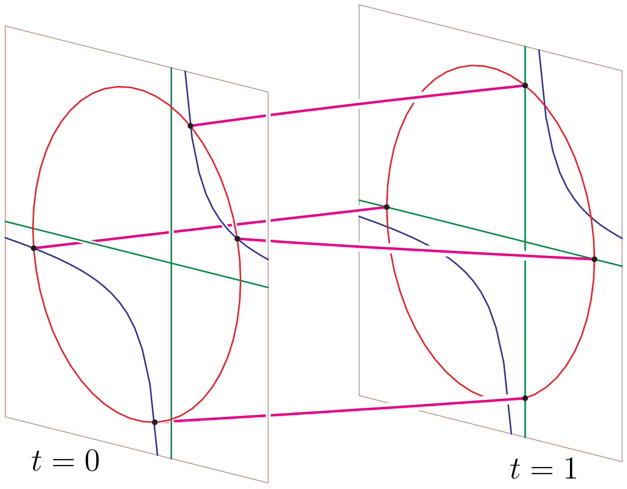


Figure 4.4: Paths connecting solutions.

This defines a space curve C in $\mathbb{C}^2_{yz} \times \mathbb{C}_t$. Restricting C to $t \in [0, 1]$ gives four paths that connect the known solutions to (4.9) at $t = 1$ to the unknown solutions to (4.8) at $t = 0$. These paths are shown in Figure 4.4. To find the unknown solutions at $t = 0$, we need to track these four paths, starting from the known solutions at $t = 1$.

Let $H(y, z; t)$ be the system of two polynomials in (4.11) that define the space curve C, and let $x(t) := (y(t), z(t))$ for $t \in [0, 1]$ be the projection of one of the paths in Figure 4.4, which is a parametrization of one of the thickened arcs in Figure 4.3.

Then $H(y(t), z(t); t) \equiv 0$ for $t \in [0, 1]$. Differentiating this gives

$$\frac{\partial H}{\partial y} \cdot \frac{dx}{dt} \;+\; \frac{\partial H}{\partial z} \cdot \frac{dy}{dt} \;+\; \frac{\partial H}{\partial t} \;=\; 0\,.$$

Solving, shows that $x(t)$ satisfies the *Davidenko differential equation*

$$x' \;=\; -(D_x H)^{-1} \frac{\partial H}{\partial t}\,. \tag{4.12}$$

Here, $D_x H$ is the Jacobian matrix of H with respect to its first two (y, z) variables. Thus each of the four paths in Figure 4.4 is a solution of an initial value problem for this differential equation, one for each of the four points (4.10). Using an iterative method to solve these initial value problems computes approximations to the solutions of (4.8).

This is dramatically improved when combined with Newton's method. Fix a sequence of points $1 = t_0 > t_1 > \cdots > t_m = 0$ in $[0, 1]$, let z_0 be one of the four solutions (4.10) to the system (4.9) and $x(t)$ the corresponding path. Having computed x_i for some $i < m$, apply one step of any iterative method (Euler, midpoint, RK4, ...) with stepsize $t_{i+1} - t_i$ to get a prediction x^*_{i+1} for $x(t_{i+1})$. Next, perform Newton iterations for $H(x; t_{i+1})$ starting at x^*_{i+1}, until some stopping criterion is reached, obtaining x_{i+1}. If each x_i lies in the basin of attraction of $x(t_i)$ for Newton iterations, then x_m converges to $x(0)$ under

Newton iterations. Better is when the x_i are approximate solutions, for then x_m is an approximate solution to $H(x;0)=0$ with associated zero $x(0)$.

This discussion about the system (4.11) is in fact quite general. Let $H(x;t)$ be a system of n polynomials in $n+1$ variables ($x \in \mathbb{C}^n$ and $t \in \mathbb{C}_t$). Then every component of the affine variety $\mathcal{V}(H)$ has dimension at least 1. Let C be the union of all components of dimension 1 whose projection to $\mathbb{C}_t$ is dense.Then A point $(x;t) \in \mathcal{V}(H)$ is *nondegenerate* if the Jacobian matrix D_xH with respect to its x-variables is invertible at $(x;t)$. A nondegenerate point is isolated from other points of $\mathcal{V}(H)$ in its fiber $\mathbb{C}^n \times \{t\}$ and the nondegenerate points are exactly the points where the Davidenko differential equation (4.12) is defined. You are asked to prove the following lemma in Exercise 2.

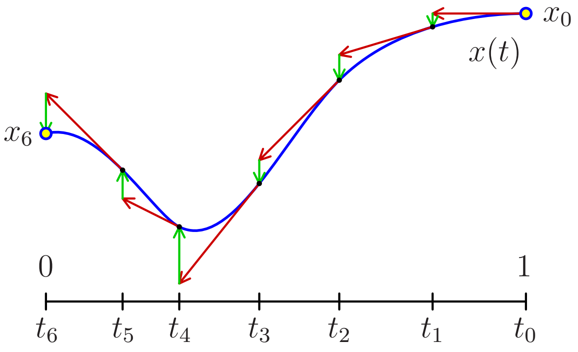
Lemma 4.2.1. *The curve C contains every point $(x;t)$ of $\mathcal{V}(H)$ that is isolated from other points of $\mathcal{V}(H)$ in its fiber. If C is nonempty, then there is a positive integer d and a nonempty Zariski-open subset U of $\mathbb{C}_t$ consisting of all points u such that $H(x;u)=0$ has d nondegenerate solutions. Above every point b of the complement $B := \mathbb{C}_t \smallsetminus U$ there is some point where either the curve C meets another component of $\mathcal{V}(H)$ or the map $C \to \mathbb{C}_t$ is ramified or C has a branch tending to infinity near C.*

We call $H(x;t)$ a *homotopy* if C is nonempty and $1 \in U$. Suppose further that

$$\text{the interval } [0,1] \text{ of } \mathbb{R} \text{ is a subset of } U. \tag{4.13}$$

Then the restriction $C|_{[0,1]}$ of C to $t \in [0,1]$ is a collection of d smooth paths. The *start system* is $H(x;1)=0$ and the *target system* is $H(x;0)=0$, and both have d nondegenerate solutions. Each path in $C|_{[0,1]}$ connects one nondegenerate solution to the start system to one nondegenerate solution to the target system.

Given a nondegenerate solution x_0 to the start system, let $x(t)$ be the path in $C|_{[0,1]}$ with $x_0 = x(1)$. It satisfies the Davidenko differential equation (4.12). Choosing a sequence of points $1 = t_0 > t_1 > \cdots > t_m = 0$ in $[0,1]$, a *predictor-corrector method* constructs a sequence $x_1, \ldots, x_m$ of approximations to the points $x(t_i)$ along the path alternately applying an iterative method to x_i to get a prediction x^*_{i+1} to $x(t_{i+1})$, which is refined using Newton iterations to get the next point x_{i+1}. Here is a schematic of the predictor-corrector method using Euler predictions to trace a smooth path $x(t)$.



A predictor-corrector method only computes refinable approximations to a sequence of points on an implicitly defined path $x(t)$ for $t \in [0,1]$. We will largely ignore this distinction and refer to the computed points as lying on the path with $x(1)$ the starting point and $x(0)$ the output, and use the term *path tracking* for this process.

The algorithm of *numerical homotopy continuation* begins with a homotopy $H(x;t)$ satisfying (4.13) and the set of nondegenerate solutions to the start system $H(x;1)=0$. By assumption (4.13), each solution x is the starting point $x(1)$ of a smooth path $x(t)$ defined implicitly by $H(x;t)=0$ for $t\in[0,1]$. For each solution x to the start system, the algorithm tracks the path $x(t)$ from $t=1$ to $t=0$ to obtain $x(0)$, a solution to the target system.

Theorem 4.2.2. *Numerical homotopy continuation under assumption* (4.13) *computes all nondegenerate solutions to the start system* $H(x;0)=0$.

Proof. For each solution x to the start system, the path $x(t)$ is smooth, so that path tracking starting from $x=x(1)$ will compute its endpoint $x(0)$. By assumption (4.13), each nondegenerate solution of the target system is connected to a nondegenerate solution of the start system along one of these paths. This completes the proof. □

The Bézout homotopy is one of the simplest homotopies. Suppose that $F=(f_1,\dots,f_n)$ is a system of n polynomials in n variables with $\deg f_i = d_i$. By Bézout's Theorem ????, $\mathcal{V}(F)$ has at most $d := d_1 d_2 \cdots d_n$ isolated solutions, and if F is generic, it has exactly d solutions, all nondegenerate. Given the degrees $d_1,\dots,d_n$, for each $i=1,\dots,n$, set

$$g_i(x) \;:=\; x_i^{d_i} - 1\,. \tag{4.14}$$

Then the system $G=(g_1,\dots,g_n)$ has the d solutions,

$$\{(\zeta_1,\dots,\zeta_n) \;\mid\; \zeta_i = e^{\frac{2\pi k}{d_i}\sqrt{-1}} \text{ for } k=0,\dots,d_i \text{ and } i=1,\dots,n\}\,.$$

The *Bézout homotopy* is the convex combination of F and G,

$$H(x;t) \;:=\; tG \;+\; (1-t)F\,. \tag{4.15}$$

For any $t\in\mathbb{C}_t$, $H(x;t)=0$ has at most d isolated solutions. As there are d nondegenerate solutions when $t=1$, the curve C of Lemma 4.2.1 is nonempty and $1\in U$. Thus (4.15) is a homotopy with start system G having known solutions and target system F.

Proposition 4.2.3. *If the system F is general, then numerical homotopy continuation using the Bézout homotopy will compute all d solutions to $F=0$.*

This follows from Theorem 4.2.2, once we see that F general implies that assumption (4.13) holds. While this follows from the discussion below, our goal is to modify the homotopy (4.15) and our path-tracking algorithm to prove the stronger result that the modified Bézout homotopy computes all isolated solutions to the system F.

Let us examine condition (4.13) for the Bézout homotopy. At the endpoints $t = 0, 1$, (4.13) is ensured if the target system is general. To help understand what may happen for $t \in (0, 1)$, consider the Bézout homotopy in one variable

$$H(x;t) \ := \ t(x^2-1) + (1-t)(x^2+x+1) \ = \ x^2 + (1-t)x + 1 - 2t\,. \tag{4.16}$$

The zeroes of $H(x;t) = 0$ as a function of t are found using the quadratic formula

$$x(t) \ = \ \frac{t-1}{2} \pm \frac{\sqrt{t^2+6t-3}}{2}\,.$$

The system $H(x; 2\sqrt{3}-3)$ has a single root $\sqrt{3}-1$ of multiplicity 2. At that point, $\frac{\partial H}{\partial x} = 0$ and so assumption (4.13) fails as $2\sqrt{3} - 3 \approx 0.464$ is a branch point in $[0, 1]$.

The Bézout homotopy is a *straight-line homotopy*, which is a convex combination

$$H(x;t) \ := \ tG + (1-t)F \tag{4.17}$$

of two polynomial systems that forms a homotopy (defines a curve C with $t = 1$ not a branch point). When both F and G are real as in (4.16), the branch locus likely contains real points that meet the interval $[0, 1]$ even when 0 is not a branch point. A simple modification gives smooth paths above $[0, 1]$. Let γ be any nonzero complex number, and set

$$H_\gamma(x;t) \ := \ \gamma t G \ + \ (1-t)F\,. \tag{4.18}$$

The modification (4.18) is called the 'γ-trick'.

Theorem 4.2.4. *Let F, G be as above. For any nonzero $\gamma \in \mathbb{C}$, $H_\gamma(x;t)$ is a homotopy with start system G and target system F. When 0 is not a branch point of* (4.17)*, there is a finite set Θ of arguments such that if* $\arg(\gamma) \notin \Theta$, *then $H_\gamma(x;t)$ satisfies* (4.13).

Proof. For the first statement, substitute $t = 1, 0$ into the formula for $H_\gamma(x;t)$. To understand the modification (4.18) and prove the second statement, note that

$$\frac{\gamma t}{\gamma t + (1-t)} A + \Big(1 - \frac{\gamma t}{\gamma t + (1-t)}\Big) B \ = \ \frac{1}{\gamma t + (1-t)}\big(\gamma t A + (1-t)B\big)\,,$$

for indeterminates A, B, t. Consequently, if we define $\tau_\gamma(t) := \gamma t/(\gamma t + (1-t))$ and if $\gamma t + (1-t) \neq 0$ for $t \in [0, 1]$, then for every $t \in [0, 1]$,

$$\gamma t F + (1-t)G \qquad \text{and} \qquad \tau_\gamma(t)F + (1-\tau_\gamma(t))G$$

have the same solutions. That is, if $\gamma t + (1-t) \neq 0$ for $t \in [0, 1]$, then the homotopy $H_\gamma(x;t)$ (4.18) for $t \in [0, 1]$ is a straight-line homotopy (4.17), but over the image of $\tau_\gamma \colon [0, 1] \to \mathbb{C}$. Solving $\gamma t + (1-t) = 0$ gives $\gamma = 1 - \frac{1}{t}$, so γ cannot be a negative real number, $\arg\gamma \neq \pi$. Identifying $\mathbb{C}$ with $\mathbb{R}^2$ where $(x, y) \leftrightarrow x + y\sqrt{-1}$ and writing $\gamma = a + b\sqrt{-1}$, the path $\tau_\gamma(t)$ for $t \in [0, 1]$ lies on the circle $x^2 - x + y^2 + \frac{a}{b}y = 0$ with

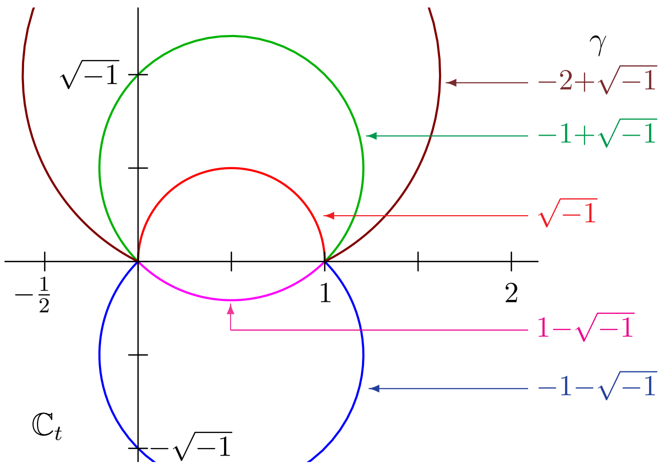


Figure 4.5: Paths τ_γ for $\gamma = -1-\sqrt{-1}, 1-\sqrt{-1}, \sqrt{-1}, -1+\sqrt{-1}, -2+\sqrt{-1}$

center $(\frac{1}{2}, -\frac{a}{2b})$ and radius $\frac{a^2+b^2}{4b^2}$. This circle contains the points 0 and 1, and the path $\tau_\gamma(t)$ for $t \in [0,1]$ traces the arc of that circle lying in the same half-plane as γ.

The paths defined by $H_\gamma(x;t) = 0$ for $t \in [0,1]$ are those in the curve C lying above the image $\tau_\gamma[0,1]$. They will be continuous and satisfy the Davidenko differential equation exactly when $\tau_\gamma[0,1]$ does not meet the branch locus B. As B is finite, τ_γ depends only upon the argument of γ, and the arcs $\tau_\gamma(0,1)$ foliate $\mathbb{C}_t \smallsetminus \mathbb{R}$, there are only finitely many arguments for γ such that $\tau_\gamma[0,1]$ meets B. This completes the proof. □

For the Bézout homotopy and any other straight-line homotopy, we use the γ-trick for a general $\gamma \in \mathbb{C} \smallsetminus \mathbb{R}$. This γ-trick is a systematic (and easy) way to choose a smooth path in $\mathbb{C}_t \smallsetminus B$ between 0 and 1, but any such path will suffice. For a general homotopy, we will assume that the tracking is done over a general smooth path $\tau \subset \mathbb{C}_t$ between 0 and 1. These assumptions imply that branch points in $\mathbb{C}_t \smallsetminus \{0\}$ may be avoided. This is commonly expressed as "the homotopy defines smooth paths, *with probability one*". That is, the set of paths between 0 and 1 in $\mathbb{C}_t$ that meet the base locus has measure zero in the collection of all paths considered.

Suppose that we have a homotopy $H(x;t)$ and assume for simplicity that the interval $(0,1] \subset \mathbb{C}_t$ does not meet the branch locus B. (In general choose a smooth path τ in $(\mathbb{C}_t \smallsetminus B) \cup \{0\}$ between 0 and 1.) Then $C|_{(0,1]}$ is a collection of d half-open smooth paths, and at each point of every path the Jacobian matrix DH_x is invertible. When $0 \notin B$, the homotopy satisfies (4.13) and by Theorem 4.2.2 numerical homotopy continuation computes all solutions to the target system, given the solutions to the start system. When $0 \in B$, by Lemma 4.2.1 there are several cases for a path $x(t)$ in $C|_{(0,1]}$ in the limit as $t \to 0$.

(1) The path $x(t)$ does not have a limit as it becomes unbounded as $t \to 0$.

(2) The path $x(t)$ has a limit $x(0)$ and D_xH is invertible at $x(0)$.

(3) The path $x(t)$ has a limit $x(0)$ that lies on another component of $\mathcal{V}(H)$ so that D_xH is not invertible at $x(0)$.

(4) The path $x(t)$ has a limit $x(0)$ that is a branch point of $C \to \mathbb{C}_t$, so that D_xH is not invertible at $x(0)$ and at least one other path also ends at $x(0)$.

Figure 4.6 is a schematic showing these four cases.

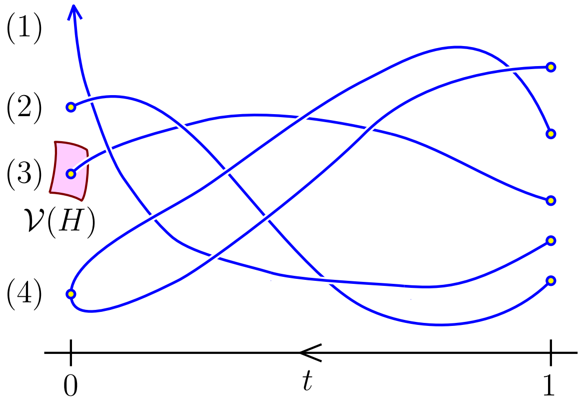


Figure 4.6: Possible behavior of homotopy paths near $t = 0$.

In Case (2), when D_xH is invertible at the endpoint $x(0)$ of the path $x(t)$, this path may be successfully tracked from $x(1)$ to $x(0)$. In all other cases, simple path-tracking will fail, as D_xH is not invertible at $x(0)$. and alternatives to simple path-tracking, called *endgames*, are needed.

Case (1) There are at least two endgames when $x(t)$ becomes unbounded as $t \to 0$. For one, when $\|x(t)\|$ exceeds a heuristic threshold, tracking is halted and the path is declared to diverge. The other applies, for example, when the homotopy $H(x;t)$ for $x \in \mathbb{C}^n_x$ and $t \in \mathbb{C}_t$ is the restriction of a homotopy $\widetilde{H}(z;t)$ for $z \in \mathbb{P}^n$ to an affine patch $\mathbb{C}^n_x \subset \mathbb{P}^n$. Choosing a different affine patch $\mathbb{C}^n_y$ and applying a change of coordinates expresses the homotopy and the computed points on the path $x(t)$ in the coordinates y of $\mathbb{C}^n_y$. If $\mathbb{C}^n_y$ is chosen propitiously (e.g. at random), then the resulting path $y(t)$ converges in $\mathbb{C}^n_y$ to a point $y(0)$ as $t \to 0$, and this path falls into one of cases (2), (3), or (4).

Case (3) While geometrically distinct from Case (4), this is treated in the same way as Case (4).

Case (4) The curve C is either singular at $x(0)$ or the map to $\mathbb{C}_t$ is ramified at $x(0)$, or both. Let us examine in detail its geometry near $x(0)$ before describing the Cauchy endgame, which also applies to the other cases (2) and (3).

Let $f\colon \widetilde{C} \to C$ be the normalization of C, so that $\widetilde{C}$ is smooth. The *ramification index* $r = r(x')$ of a preimage $x' \in f^{-1}(x(0))$ is the order of vanishing at x' of the (rational) function t that is the composition $\pi\colon \widetilde{C} \to C \to \mathbb{C}_t$, with the second map the projection onto the t-coordinate. If s is a nonconstant rational function on $\widetilde{C}$ that vanishes to order 1 at x', then $t = s^r g$, for some rational function g with $g(x') \neq 0$.

Suppose that $\Delta \subset \mathbb{C}_t$ is a disc centered at the origin small enough so that 0 is the only branch point in Δ. Then each component of its preimage $\pi^{-1}(\Delta)$ in $\widetilde{C}$ contains a unique point $x' \in \pi^{-1}(0)$. On the component Δ' containing x', the map $\pi\colon \Delta' \smallsetminus \{x'\} \to \Delta \smallsetminus \{0\}$ of punctured neighborhoods is an r-fold covering space, analytically isomorphic to the map $s \mapsto s^r$. Over the punctured disc $\Delta \smallsetminus \{0\}$, the two curves $\widetilde{C}$ and C agree, and the image in C of a component of $\pi^{-1}(\Delta)$ is a *branch* of C at the corresponding point above 0. Figure 4.7 shows some possibilities near a ramification point. The map $C \to \mathbb{C}_t$ is the vertical projection and only one veritcal real dimension is shown. (The self-intersections, except at $x(0)$, are artifacts of this.) In (a), C is singular at $x(0)$ with two smooth branches, each with ramification index 1. In (b), C is smooth with one branch at $x(0)$ and ramification index 2. In (c), the ramification index is 3.

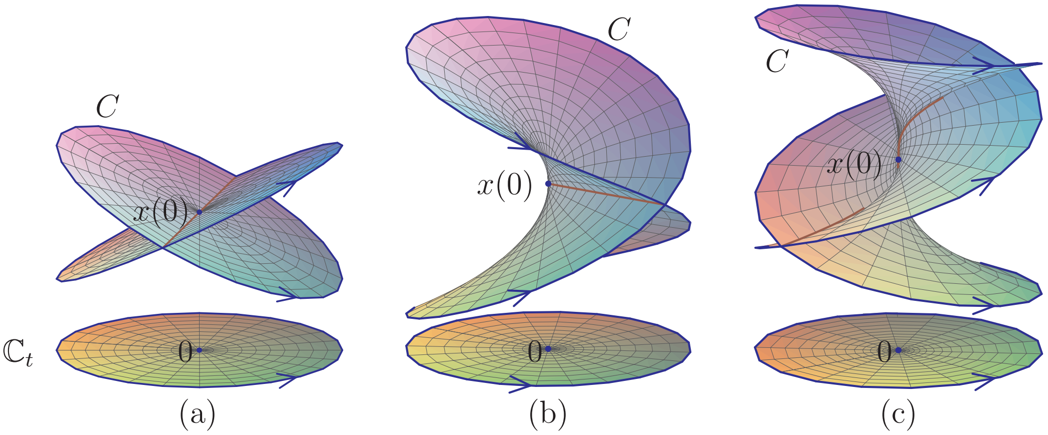


Figure 4.7: Local behaviour near a ramification point.

The ramification index is also a winding number. Given a point $x(\epsilon) \in C$ for $\epsilon > 0$ in the disc $\Delta \subset \mathbb{C}_t$, analytic continuation on C starting at $x(\epsilon)$ above the circle $\epsilon e^{2\pi\theta\sqrt{-1}}$ for $\theta \geq 0$ gives a closed path $(y(\theta), \epsilon e^{2\pi\theta\sqrt{-1}})$ in C which is parametrized by $\theta \in [0, r]$. The path $(y(\theta), \epsilon e^{2\pi\theta\sqrt{-1}})$ encircling $x(0)$ in C has image in $\mathbb{C}_t$ winding r times around 0. Figure 4.7 shows the circle and these paths. The ramification index may be computed numerically by tracking the path $(y(\theta), \epsilon e^{2\pi\theta\sqrt{-1}})$ for $\theta \geq 0$.

We may compute the endpoint $x(0)$ of the path $x(t)$ in C without tracking the path $x(t)$ to $t = 0$ using Cauchy's integral formula. Recall that if g is a function that is holomorphic in a neighborhood of closed disc D centered at the origin, then

$$g(0) \;=\; \frac{1}{2\pi\sqrt{-1}} \oint_{\partial D} \frac{g(z)}{z} dz \;.$$

This holds also when g is vector-valued. In our case, let D be the unit disc and consider the map $D \to \Delta$ where $z \mapsto \epsilon z^r =: t$. This lifts to a map $g\colon D \to C$ with $g(0) = x(0)$ and

$g(e^{2\pi\alpha\sqrt{-1}}) = (y(r\alpha), e^{2\pi\alpha\sqrt{-1}})$ to which we may apply the Cauchy integral formula. After a change of coordinates, we obtain

$$x(0) \;=\; \Bigl(\frac{1}{\sqrt{-1}}\int_0^1 \frac{y(r\alpha)}{e^{2\pi\alpha\sqrt{-1}}}\,d\alpha\ ,\ 0\Bigr)\ . \tag{4.19}$$

The *Cauchy endgame* is a numerical algorithm using this.

Algorithm 4.2.5 (Cauchy Endgame)**.**
INPUT: A homotopy $H(x;t)$, curve C as in Lemma 4.2.1, a path $x(t)$ on C with limit $x(0)$ as $t \to 0$, and a point $x(\epsilon)$ on the path with $\epsilon > 0$.
OUTPUT: A numerical approximation to $x(0)$ and the ramification index.

1. Starting at $x(\epsilon) = (y(0), \epsilon)$, track the path $y(\theta)$ above the circle $\epsilon e^{2\pi\theta\sqrt{-1}}$ from $\theta = 0$ until the first integer $r > 0$ with $y(r) = y(0)$.

2. Using the computed intermediate values of θ and $y(\theta)$, estimate $x(0)$, using numerical integration for the integral (4.19).

3. Track $x(t)$ from $x(\epsilon)$ to $x(\epsilon/2)$, replace ϵ by $\epsilon/2$, and repeat steps (1) and (2), obtaining another estimate for $x(0)$.

   If successive estimates agree up to a tolerance, or do so after it repeating (3), then exit and return the computed value of $x(0)$, along with r.

Remark 4.2.6. The Cauchy endgame applies in each of the cases (1)—(4) above. (In (1), first change coordinates in $\mathbb{P}^n$ so that the path has a finite limit.) Simply stop the tracking of $x(t)$ at some fixed ϵ (e.g. $\epsilon = 0.1$), and then apply the Cauchy endgame. There will be r paths that converge to the endpoint $x(0)$. An additional check verifies that there are indeed $r-1$ other paths with this endpoint.

We deduce a strengthening of Theorem 4.2.2.

Theorem 4.2.2′ *Numerical homotopy continuation with endgames computes all isolated solutions to the start system $H(x;0) = 0$.*

A homotopy is *optimal* if every isolated solution of the target system is connected to a unique solution of the start system along a path of $C|_\tau$, for $\tau \subset \mathbb{C}_t \smallsetminus B$ a path connecting 0 to 1. By Proposition 4.2.3, the Bézout homotopy is optimal, for a general target system F. In a non-optimal homotopy some paths either diverge (Case(1)) or become singular (Cases (3) and (4)), all of which require expensive endgames.

The cost of using numerical homotopy continuation to solve a system of polynomials is dominated by path-tracking and engames, and is therefore minimized for optimal or near optimal homotopies. A significant advantage is that homotopy continuation algorithms are inherently massively parallelizable—once the initial precomputation of solving the start system and setting up the homotopies is completed, then each solution curve may be followed independently of all other solution curves.

Given a specific target system $F = (f_1, \ldots, f_n)$ with $\deg F_i = d_i$, using the Bézout homotopy (4.15) to compute its solutions may be problematic as neither the start (4.14) nor the target systems are necessarily generic. In practice the more robust Bézout homotopy algorithm overcomes this by using a two-step process.

Algorithm 4.2.7 (Bézout Homotopy Algorithm).
<u>INPUT:</u> A target system $F = (f_1, \ldots, f_n)$ with $\deg f_i = d_i$.
<u>OUTPUT:</u> Numerical approximations to the isolated zeroes of $\mathcal{V}(F)$.

1. Generate a random system $E = (e_1, \ldots, e_n)$ of polynomials with $\deg e_i = d_i$.
2. Use the Bézout homotopy (4.15) with start system (4.14) to compute all $d = d_1 \cdots d_n$ solutions to $\mathcal{V}(E)$.
3. Use the straight-line homotopy $tE + (1-t)F$ starting with the solutions to E to compute the isolated solutions to $\mathcal{V}(F)$, possibly employing endgames.

Theorem 4.2.8. *The Bézout homotopy algorithm computes all isolated zeroes of F.*

This is a probability one algorithm, as it requires that $\mathcal{V}(E)$ consist of d isolated solutions, which holds on an open dense set of such systems.

Proof. As E is generic, Proposition 4.2.3 implies that the first homotopy is optimal and it will compute all d solutions to $\mathcal{V}(E)$. For every t, the second homotopy is a system of polynomials in x of degrees $d_1, \ldots, d_n$ and so by Bézout's Theorem it has at most d isolated solutions, and when there are d solutions, all are nondegenerate. Since the start system has d solutions, this implies that $C|_{(0,1]}$ consists of d half-open paths beginning at $t = 1$ with $\mathcal{V}(E)$. By our discussion of endgames (and possibly using projective coordinates in Case (1)), all isolated solutions of $\mathcal{V}(F)$ will be found. □

Example 4.2.9. One source of homotopies are polynomial systems that depend upon parameters. For example, a bivariate polynomial $F_d(x; c)$ of degree d ($x \in \mathbb{C}^2$) has coefficients $c \in \mathbb{C}^{\binom{d+2}{2}}$. Thus a system consisting of a quadratic $F_2(x; a)$ and a cubic $F_3(x; b)$ depends upon $\binom{4}{2} + \binom{5}{2} = 6 + 10 = 16$ parameters. This gives a family

$$\Gamma \;:=\; \{(x; a, b) \in \mathbb{C}^2 \times \mathbb{C}^6 \times \mathbb{C}^{10} \mid F_2(x; a) = F_3(x; b) = 0\}\,,$$

with a map $\Gamma \to \mathbb{C}^{16}$ whose fiber over $(a, b) \in \mathbb{C}^{16}$ is the set of common zeroes to $F_2(x; a)$ and $F_3(x; b)$. There is a non-empty Zariski open set $U \subset \mathbb{C}^{16}$ consisting of pairs (a, b) for which the fiber has six solutions and its complement is the branch locus B. We call U the set of regular values of $\Gamma \to \mathbb{C}^{16}$. For a and b general, this has six solutions by Bézout's Theorem. We obtain a homotopy by parametrizing a line ℓ in the base, $f \colon \mathbb{C} \to \ell \subset \mathbb{C}^{16}$ where $f(t) = (a(t), b(t))$ with each component linear, and where $(a(1), b(1))$ gives a system with six solutions. Then $\Gamma|_\ell = f^{-1}(\Gamma)$ is given by the homotopy $H(x; t) := (F_2(x; a(t)), F_3(x; b(t)))$.

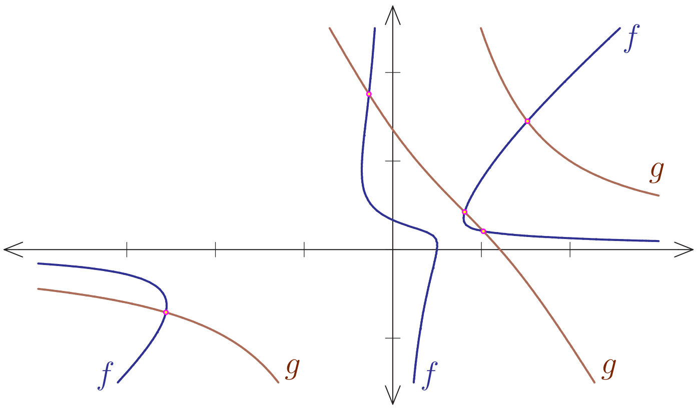
A homotopy arising from such a family where the start and target systems lie in the open set of regular values is a *parameter homotopy*. The Bézout homotopy for a general target system is a parameter homotopy. Nearly every homotopy we use will be a parameter homotopy, often with a propitious choice of line (or a rational curve) in the base.

Polynomial systems arising 'in nature' are rarely generic dense systems. The bound from Bézout's Theorem for the number of isolated solutions is typically not achieved.

For example, consider the system of cubic polynomials,

$$\begin{aligned} f\colon \quad & 1 - 2x - 3y + 4xy + 5x^2y - 6xy^2 = 0 \\ g\colon \quad & 23 - 17x - 19y - 13xy + 11x^2y + 7xy^2 = 0 \end{aligned} \tag{4.20}$$

This has the five solutions shown below, and not the 9 predicted by Bézout's Theorem.



Section ????? describes a method to solve structured systems of equations that are either not square or have fewer solutions than expected from Bézout's Theorem. Square systems of the form 4.20 which are general given the monomials in each polynomial are called *sparse*, and the polyhedral homotopy, based on ideas from the study of toric varieties, is an optimal homotopy for sparse systems. This will be developed in Section 8.6.

Exercises

1. Find the solutions to the system of equations (4.8) directly, and also by solving the initial value problem for the differential equation 4.12 starting at the solutions (4.10) using any of the iterative methods of Section 4.1.

2. Give a proof of Lemma 4.2.1.

3. Let $\gamma = a + b\sqrt{-1}$ with a, b real and $b \neq 0$. Show that the path in the complex plane $\tau_\gamma(t) := \gamma t/(\gamma t + (1-t))$ for $t \in [0,1]$ lies on the circle with centre $(\frac{1}{2}, -\frac{a}{2b})$ and radius $\frac{a^2+b^2}{4b^2}$, which contains the points 0 and 1. Show that the tangent direction of τ_γ at $t = 0$ is γ and that $\tau_\gamma[0,1]$ lies in the same half-plane as γ.

4. Discuss how to get equations for the branch locus in Example 4.2.9.

5. Explain the ramification of $y^2 = x^3$ in each coordinate projection.
6. Verify the claim in the text that the system (4.20) has exactly five solutions.
7. Needs more exercises.

4.3 Numerical Algebraic Geometry

In Section 4.2 on numerical homotopy continuation, we discussed path-tracking, presented the Bézout homotopy to compute all isolated solutions to a square system of polynomial equations, and mentioned improvements that are covered elsewhere in this text. Numerical algebraic geometry uses this ability to solve systems of polynomial equations to represent and study algebraic varieties on a computer. We first discuss overdetermined and undetermined systems of equations before introducing the notion of a witness set, which is one of the fundamental ideas in numerical algebraic geometry.

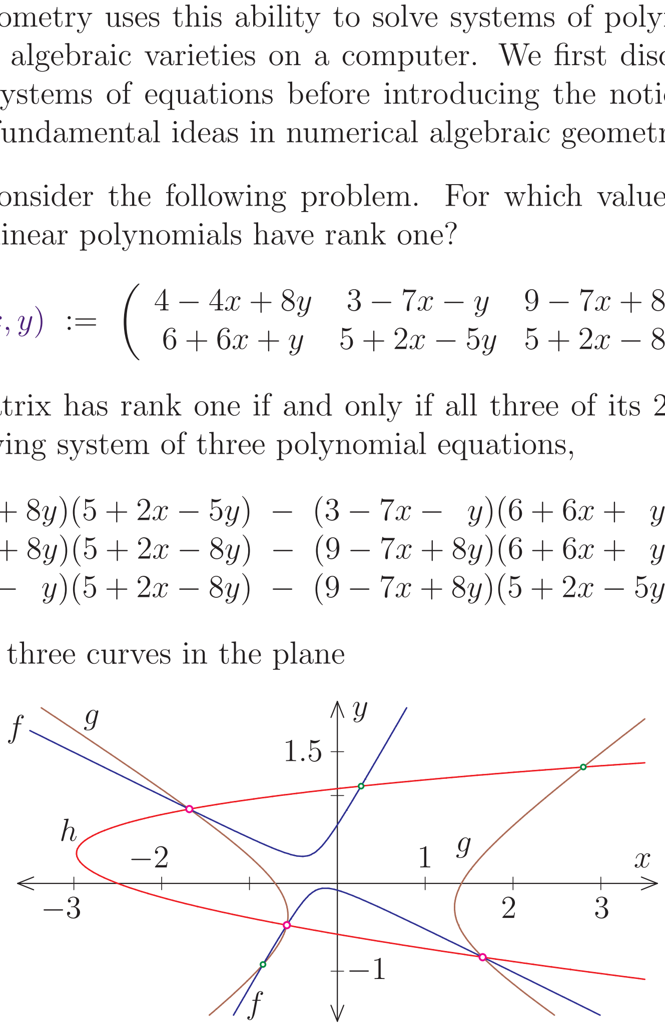
Example 4.3.1. Consider the following problem. For which values of (x, y) does the following matrix of linear polynomials have rank one?

$$M(x,y) \;:=\; \begin{pmatrix} 4-4x+8y & 3-7x-y & 9-7x+8y \\ 6+6x+y & 5+2x-5y & 5+2x-8y \end{pmatrix} \tag{4.21}$$

A nonzero 2×3 matrix has rank one if and only if all three of its 2×2 minors vanish. This gives the following system of three polynomial equations,

$$\begin{array}{rlcl} f: & (4-4x+8y)(5+2x-5y) - (3-7x-\;\,y)(6+6x+\;\,y) & = & 0 \\ g: & (4-4x+8y)(5+2x-8y) - (9-7x+8y)(6+6x+\;\,y) & = & 0 \\ h: & (3-7x-\;\,y)(5+2x-8y) - (9-7x+8y)(5+2x-5y) & = & 0 \end{array} \tag{4.22}$$

These minors define three curves in the plane



which appear to have three common solutions. We may verify this by computing a lexicographic Gröbner basis with $y < x$ from the minors (4.22),

$$G \;=\; \{213y^3 + 100y^2 - 152y - 72\,,\; 213y^2 - 136y - 68x - 152\}\,.$$

The eliminant in y has degree three. Finding its roots, substituting into the second polynomial, and solving for x gives the three solutions

$$(1.654564, -0.8399545)\,,\ (-0.5754011, -0.4756340)\,,\ (-1.685073, 0.8461049)\,. \qquad (4.23)$$

Each pair of minors vanishes at four points. The fourth point is where the column common to the two minors vanishes. It may be pruned from the other three by evaluating the third minor at all four points and retaining only those where the evaluation is below a predetermined threshold. For example, the minors f and g also vanish at $(-11/13, -12/13)$. Evaluating the third minor h at these four points gives the values $(-4.5 \times 10^{-6}, -1.3 \times 10^{-7}, 9.2 \times 10^{-7}, 45.6)$. The value of h at three of the points is approximately the working precision, so we (correctly) discard the fourth. [†]

As in Example 4.3.1, meaningful systems of equations are not necessarily square; they may have more equations than variables (are *overdetermined*), and yet define a zero-dimensional ideal. There may be no reasonable way to select a square subsystem whose solutions are only those of the original system. When faced with an overdetermined system, one approach is to randomly select a square subsystem ('squaring up' the original system). This square system is solved and polynomials from the original system are used to prune the excess solutions found in the square subsystem.

Algorithm 4.3.2 (Squaring up).
<u>INPUT:</u> An overdetermined system $F\colon \mathbb{C}^n \to \mathbb{C}^m$ $(m > n)$ of polynomials.
<u>OUTPUT:</u> A square system $G\colon \mathbb{C}^n \to \mathbb{C}^n$ whose solutions $\mathcal{V}(G)$ contain the solutions $\mathcal{V}(F)$ of F, with the nondegenerate solutions of F remaining nondegenerate for G.[†]
<u>DO:</u> Select a random linear map $\Lambda\colon \mathbb{C}^m \to \mathbb{C}^n$ and return $G := \Lambda \circ F$.

Algorithm 4.3.2 is a 'probablility one' algorithm. We give a proof of its correctness.

Theorem 4.3.3. *For any linear map $\Lambda\colon \mathbb{C}^m \to \mathbb{C}^n$, we have the containment $(\mathcal{F}) \subset \mathcal{V}(\mathcal{G})$ of solutions in Algorithm 4.3.2. There is a nonempty Zariski open subset U of $n \times m$ complex matrices consisting of linear maps Λ such that the nondegenerate solutions of F remain nondegenerate soutions of $G = \Lambda \circ F$.*

Proof. As Λ is linear, if $F(x) = 0$, then $\Lambda \circ F(x) = 0$, which proves the first statement.

For the second statement, let $x \in \mathcal{V}(F) \subset \mathbb{C}^n$ be nondegenerate. Then the Jacobian matrix $DF(x)$ at x gives an injective linear map from $\mathbb{C}^n = T_x\mathbb{C}^n \to \mathbb{C}^m$. A point $x \in \mathcal{V}(G)$ is nondegenerate if the composition $DG(x) = \Lambda \circ DF(x)\colon \mathbb{C}^n \to \mathbb{C}^n$ is injective. Geometrically, this means that the kernel of Λ is transverse to the image of $DF(x)$. This condition defines a nonempty open subset in the space of linear maps. Since $\mathcal{V}(F)$ has finitely many nondegenerate solutions, the set U is the intersection of these nonempty open subsets, one for each nondegenerate solution $x \in \mathcal{V}(F)$. □

[†] This needs to refer to an algorithm for solving in Chapter 2.

[†] Strengthen this to isolated solutions.

Remark 4.3.4. Choosing a random linear map $\Lambda\colon \mathbb{C}^m \to \mathbb{C}^{n-d}$ gives a subsystem $G = \Lambda\circ F$ of F with the following properites: Every component of $\mathcal{V}(G)$ has dimension at least d. These include all irreducible components of $\mathcal{V}(F)$ of dimension d or greater, together with possibly some other components of dimension d.

Example 4.3.5. Let us view Example 4.3.1 in a different light. Suppose that we want a rank one 2×3 matrix M, that is, a solution to the equations

$$M_{1,1}M_{2,2} - M_{1,2}M_{2,1} \;=\; M_{1,2}M_{2,3} - M_{1,3}M_{2,2} \;=\; M_{1,1}M_{2,3} - M_{1,3}M_{2,1} \;=\; 0\,. \qquad (4.24)$$

With three linearly independent equations on a six-dimensional space, this defines a subvariety of either of dimension three or of dimension four. We reduce this to a zero-dimensional problem by slicing the set of solutions to (4.24), adding the (successive) linear equations

$$\begin{aligned} 5M_{1,1} - 2M_{1,3} + 3M_{2,3} &= 17\\ 40M_{1,2} - 58M_{1,3} - 63M_{2,3} &= -717\\ 8M_{2,1} + 10M_{1,3} + 11M_{2,3} &= 193\\ 40M_{2,2} + 6M_{1,3} - 19M_{2,3} &= 159 \end{aligned} \qquad (4.25)$$

Only after these four linear equations are added to (4.24) do we obtain a zero-dimensional system with three solutions. This shows that the dimension of the set of rank one 2×3 matrices is four. The local dimension test, which we will describe later, is another way to determine the dimension of a variety, given a point on it and its defining equations.

A system of equations as in (4.24) that defines a positive-dimensional variety V whose points are of interest is an *underdetermined* system. As in Example 4.3.5, adding further equations to reduce its dimension to zero will give a system of polynomials to solve, obtaining points of V. By Bézout's Theorem, this system is expected to have the fewest number of solutions when the additional equations are linear. In this case, the linear equations define a linear subspace L whose codimension equals the dimension of V (L is *complimentary* to V), and the points are the linear section $V \cap L$.

These two techniques of squaring up and slicing down reduce any system of equations to a square system whose solutions may be further processed to obtain solutions to the original problem. When the system is underdetermined and defines a variety $V = \mathcal{V}(F)$, we obtain points of V in the complimentary linear section $V \cap L$. Numerical algebraic geometry uses this to study the algebraic variety V.

Remark 4.3.6. In Examples 4.3.1 and 4.3.5, the variety of rank one 2×3 matrices were sliced by the same linear subspace. The linear equations (4.25) define a four-dimensional linear subspace L of $\mathrm{Mat}_{2\times 3}\mathbb{C}$, and the family of matrices $M(x, y)$ (4.21) for $x, y \in \mathbb{C}$ is a parametrization of L. You are asked to verify this in Exercise 1.

Any linear subspace of $\mathbb{C}^n$ has an extrinsic description as the vanishing set of some linear equations, and an intrinsic description as the image of a linear map (a parametrization). Either description may be used for slicing. This flexibility may be used to improve the efficiency of an algorithm.

The first question numerical algebraic geometry addresses is how to represent an algebraic variety $V \subset \mathbb{C}^n$ on a computer? In symbolic computation, this is answered by giving a finite set of polynomials $F = \{f_1, \ldots, f_m\}$ such that $V = \mathcal{V}(F)$, perhaps with the good algorithmic property of being a Gröbner basis for the ideal of V. In numerical algebraic geometry, if V is zero-dimensional, then the list V of (approximations to) its points, together with the polynomials that define V is a reasonable representation. When the dimension of V is at least one, we will slice V to obtain a collection of points and use these points and the slice as our representation.

Definition 4.3.7. Let $V \subset \mathbb{C}^n$ be an irreducible variety of dimension d. A *witness set* for V is a set W of the form $V \cap L$, where L is a general linear subspace of $\mathbb{C}^n$ complimentary to V, so that it has codimension d. The generality of V ensures that the intersection is transverse and by Bézout's Theorem[†], W consists of $\deg V$ points. For computational/algorithmic purposes, we will represent a witness set for V by a triple (W, F, L). Here, $W = V \cap L$ with V an irreducible component of $\mathcal{V}(F)$ where $F = \{f_1, \ldots, f_m\}$ a system of polynomials on $\mathbb{C}^n$ and $L = \{\ell_1, \ldots, \ell_d\}$ is d general linear polynomials. (We write L for both the linear subspace and its given equations.)

The same definition makes sense when V is a reducible variety, all of whose components have the same dimension d. While a witness set is certainly a representation of a variety, this definition is justified by its utility. We shall see that a witness sets is the central notion in numerical algebraic geometry and is the input for many of its algorithms. We describe some of the more elementary algorithms that use witness sets. (Needs Examples)

Sampling. A witness set (W, F, L) for a variety $V \subset \mathbb{C}^n$ includes a collection of points of V. If $L' = \{\ell'_1, \ldots, \ell'_d\}$ is another collection of d linear polynomials, we may form the straight-line homotopy

$$H(x;t) \ := \ (F(x)\,,\, tL(x) + (1-t)L'(x))\,. \tag{4.26}$$

For almost all $t \in \mathbb{C}$, the d linear polynomials $tL(x) + (1-t)L'(x)$ define a codimension d linear subspace $L_t \subset \mathbb{C}^n$ with $L_1 = L$ and $L_0 = L'$. As $V \cap L$ is transverse, for any point $w \in W = V \cap L$, the homotopy (4.26) defines a path $w(t)$ in V for $t \in (0, 1]$ with $w(1) = w$ and well-defined endpoint $w(0) = \lim_{t \to 0} w(t)$ (perhaps lying at infinity in V). We may use a witness set as the input for an algorithm to sample points of V.

Algorithm 4.3.8 (Sampling)**.**
<u>Input:</u> A witness set (W, F, L) for $V \subset \mathbb{C}^n$.
<u>Output:</u> Point(s) of V.
<u>Do:</u>

1. Choose d linear polynomials $L' = \{\ell'_1, \ldots, \ell'_d\}$ and form the homotopy $H(x;t)$ (4.26).
2. Follow one or more points w of W along the homotopy $H(x;t)$ from $t = 1$ to $t = 0$ and return the endpoints $w(0)$ of the homotopy paths.

[†] Make sure to state it this way in Chapter 3

Proof of correctness. By the generality of L, all homotopy paths $w(t)$ for $w \in W$ are smooth for $t \in (0,1]$. In particular, each lies in the smooth locus of $\mathcal{V}(F)$. As the initial point $w(1)$ of each path is a point of $W = V \cap L$, the path lies on V, and in particular, its endpoint $w(0)$ is a point of V (possibly singular in $\mathcal{V}(F)$ or at infinity). □

Moving a witness set. If the linear polynomials $L' = \{\ell'_1, \dots, \ell'_d\}$ in (4.26) are general, then $V \cap L'$ is transverse and consists of $\deg(V)$ points, by Corollary 3.6.1. Thus $W' = (V \cap L', F, L')$ is another witness set for V. This justifies the following algorithm.

Algorithm 4.3.9 (Moving a Witness Set)**.**
<u>Input:</u> A witness set (W, F, L) for $V \subset \mathbb{C}^n$.
<u>Output:</u> A second witness set $(V \cap L', F, L')$ for V.
<u>Do:</u> Choose general linear polynomials $L' = (\ell'_1, \dots, \ell'_d\}$ on $\mathbb{C}^n$. Run Algorithm 4.3.8 on all points $w \in W$, using this choice of L'. Set W' to be the collection of endpoints obtained and output (W, F, L').

Remark 4.3.10. Note that this algorithm only needs that $W = V \cap L$ is transverse and consists of $\deg(V)$ points, for then (4.26) is a homotopy defining paths that start at points w of W. It is only needed that L' be a general complimentary linear subspace.

Similarly, the Sampling Algorithm 4.3.8 only needs a complimentary linear space L and a point $w \in V \cap L$ where $V \cap L$ is transverse at w to sample points of V.

Membership. Suppose that $x \in \mathbb{C}^n$ is a point of $\mathcal{V}(F)$. We may use a witness set (W, F, L) for a variety V that is a component of $\mathcal{V}(F)$ to determine if $x \in V$.

Algorithm 4.3.11 (Membership Test)**.**
<u>Input:</u> A witness set (W, F, L) for $V \subset \mathbb{C}^n$ and a point $x \in \mathcal{V}(F)$.
<u>Output:</u> True (if $x \in V$) or False (if $x \notin V$).
<u>Do:</u>

1. Choose d linear polynomials $L' = \{\ell'_1, \dots, \ell'_d\}$ that are general given that $\ell'_i(x) = 0$.

2. Call Algorithm 4.3.8 using L' in the homotopy (4.26) and follow homotopy paths from every point $w \in W$.

3. If x is an endpoint of one of these paths, return True, otherwise, return False.

Proof of correctness. By the genericity of L', the set $W' := V \cap L'$ consists of $\deg(V)$ points, counted with multiplicity. All points of W', except possibly x (if $x \in W$), are smooth points of V with none at infinity, and all points of W' are endpoints of homotopy paths.[†] Thus $x \in V$ if and only if it is an endpoint of a path given by the homotopy (4.26) that starts at some point of W. □

[†] There is something to prove here that should have been proved earlier

Inclusion. Suppose that X and V are irreducible subvarieties of $\mathbb{C}^n$. If $X \not\subset V$, then their set-theoretic difference $X \smallsetminus V$ is open and dense in X. Furthermore, if L is a general linear subspace complimentary to X, then $X \cap L \subset X \smallsetminus V$. This observation leads to the following probability one algorithm to test if $X \subset V$.

Algorithm 4.3.12 (Inclusion).
INPUT: Witness sets (W_X, F_X, L_X) for $X \subset \mathbb{C}^n$ and (W_V, F_V, L_V) for $V \subset \mathbb{C}^n$.
OUTPUT: True (if $X \subset V$) or False (if $X \not\subset V$).
DO:

1. Call the Sampling Algorithm 4.3.8 using the witness set (W_X, F_X, L_X) for X to obtain a point $x \in X \cap L'$, where L' is a general linear subspace complimentary to X.

2. Call the Membership Test Algorithm 4.3.11 to test if $x \in V$.

Proof of correctness. The point $x \in X$ lies in $X \cap L'$, where L' is a general linear subspace complimentary to X. If $X \subset V$, then $x \in V$, and the algorithm returns True. If $X \not\subset V$, then with probability one $(X \cap L') \cap V = \emptyset$, so that $x \notin V$ and the algorithm returns False. □

The first step in this algorithm is precautionary because L_X may not be sufficiently general to avoid points of $X \cap V$ when $X \not\subset V$.

Witness set of a product. Suppose that $A \subset \mathbb{C}^n$ and $B \subset \mathbb{C}^m$ are irreducible varieties and that (W_A, F_A, L_A) and (W_B, F_B, L_B) are witness sets for A and B, respectively. Here $F_A = F_A(x)$ is a system of polynomials on $\mathbb{C}^n$ and $F_B = F_B(y)$ is a system of polynomials on $\mathbb{C}^m$. The product $A \times B$ is an irreducible component of the concatenation $F = F(x, y) = (F_A(x), F_B(y))$ of the two systems. Also, the degree of $A \times B$ is the product of the degree of A and the degree of B. Furthermore, $L_A \times L_B$ is a linear subspace of $\mathbb{C}^n \times \mathbb{C}^m$ complimentary to $A \times B$ and $W_A \times W_B = (A \times B) \cap (L_A \times L_B)$ is transverse and consists of $\deg(A) \cdot \deg(B)$ points. While we would like for $(W_A \times W_B, (F_A, F_B), L_A \times L_B)$ to be a witness set for $A \times B$, it is not a witness set as $L_A \times L_B$ is not a general linear subspace of $\mathbb{C}^n \times \mathbb{C}^m$. For example, $L_A \times L_B$ is not in general position with respect to the coordinate projections to $\mathbb{C}^n$ and to $\mathbb{C}^m$.

Algorithm 4.3.13 (Witness Set of a Product).
INPUT: Witness sets (W_A, F_A, L_A) for $A \subset \mathbb{C}^n$ and (W_B, F_B, L_B) for $B \subset \mathbb{C}^m$.
OUTPUT: A witness set (W, F, L) for the product $A \times B \subset \mathbb{C}^n \times \mathbb{C}^m$.
DO:

1. Set $F := (F_A, F_B)$ and choose general linear forms $L = (\ell_1, \ldots, \ell_d)$ on $\mathbb{C}^n \times \mathbb{C}^m$ where $d = \dim(A) + \dim(B)$.

2. Call the Moving Algorithm 4.3.9 with input $(W_A \times W_B, F, (L_A, L_B))$ to move the set $W_A \times W_B$ to the set $W := (A \times B) \cap L$.

Proof of correctness. The collection (L_A, L_B) defines $L_A \times L_B$. We observed that while $W_A \times W_B = (A \times B) \cap (L_A \times L_B)$ is transverse and consists of $\deg(A \times B) = \deg(A) \cdot \deg(B)$ points, it is not a witness set as $L_A \times L_B$ is not a general complimentary linear subspace. By Remark 4.3.10, the Moving Algorithm 4.3.9 only needs its input to be a transverse intersection with a complimentary linear subspace to compute a witness set. This implies that (W, F, L) will be a witness set for the product $A \times B$. □

Witness sets for projections. Suppose that $X \subset \mathbb{C}^n \times \mathbb{C}^k$ is a variety that is an irreducible component of a system $F = F(x, y)$ of polynomials with $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^k$. Let $\pi \colon \mathbb{C}^n \times \mathbb{C}^k \to \mathbb{C}^n$ be the coordinate projection and set $V := \overline{\pi(X)}$, and irreducible subvariety of $\mathbb{C}^n$. By Lemma 2.1.7, this coordinate projection corresponds to the elimination of the y variables from $F(x, y)$, which may be accomplished by resultants or Gröbner bases. Besides the potential complexity of this computation, there is a very real and practical problem with symbolic elimination: If X' is an irreducible component of $\mathcal{V}(F)$, eliminating y from F gives polynomials that vanish on $\pi(X')$. If it happens that $\overline{\pi(X)} \subset\neq \overline{\pi(X')}$, then we could not study $V = \overline{\pi(X)}$ using an eliminant.

In this setting of a projection, we instead use a variant of a witness set for V. For this, let $M \subset \mathbb{C}^k$ be a general linear subspace complimentary to V. As M is general, $V \cap M$ is transverse and each of its $\deg(V)$ points lies in $\pi(X)$. The intersection $X \cap (\mathbb{C}^n \times M)$ is a collection of $\deg(V)$ fibers of the projection $X \to V$. Write X_v for the fiber over a point $v \in V \cap M$. The genericity of M implies that these fibers all have the same dimension, $\dim(X) - \dim(V)$, and they all have the same degree.

Let $L \subset \mathbb{C}^n$ be a general linear subspace of codimension $\dim(X) - \dim(V)$. As it is general, L meets each of the fibers X_v for $v \in V \cap M$ transversally in $\deg(X_v)$ points, and we have

$$X \cap (L \times M) \;=\; \bigcup \{X_v \cap L \mid v \in V \cap M\}\,.$$

The quadruple $(X \cap (L \times M), F, L, M)$ is a *pseudowitness set* for $V = \overline{\pi(X)}$. This representation of V does not require knowing polynomials that vanish on V.

A pseudowitness set for the image $\overline{\pi(X)}$ of a projection may be computed from a witness set (W, F, L_X) for X in the same way as Algorithm 4.3.13, but run in reverse. That is, given L and M, we compute $X \cap (L \times M)$ from $X \cap L_X$ using a homotopy between the general linear subspace L_X and the linear subspace $L \times M$.

As the complimentary linear subspace $L \times M$ in a pseudowitness set is not in general position, algorithms that involve manipulating a pseudowitness set for V to obtain a pseudowitness set for another variety $V' = \overline{\pi(X')}$ will have two additional steps (1 and 3 below) as described in the following outline.

1. Use the pseudowitness set $(X \cap (L \times M), F, L, M)$ for $\overline{\pi(X)}$ to compute a witness set (W, F, L_X) for X (using Algorithm 4.3.13).

2. Apply an appropriate construction or algorithm on X using (W, F, L_X) to obtain a witness set (W', F', L'_X) for a variety X' with projection $\overline{\pi(X')} = V'$.

3. Using Algorithm 4.3.13 move the witness set (W', F', L'_X) for X' to a pseudowitness set $(X' \cap (L' \times M'), F', L', M')$ for $V' = \overline{\pi(X')}$.

Studying $V = \overline{\pi(X)}$ using a pseudowitness set is a numerical version of elimination theory.

Local dimension test. Suppose that x is a point lying on a variety V that is an irreducible component of $\mathcal{V}(F)$ where $F = \{f_1, \ldots, f_m\}$ is a system of polynomials on $\mathbb{C}^n$. We would like a numerical method to compute the dimension of V. We discuss one that assumes V is smooth at x. While this does not use a witness set as an input, it has a similar flavor and is needed in subsequent sections.

As explained in Section 3.4, given a point $x \in \mathcal{V}(F)$, the differentials $d_x f_i$ of the polynomials $f_i \in F$ at x define the Zariski tangent space $T_x\mathcal{V}(F)$ of $\mathcal{V}(F)$ at x. Thus $\dim T_x\mathcal{V}(F) = n - \operatorname{rank}(DF(x))$, the corank of the Jacobian matrix of F at x. When x is a smooth point of $\mathcal{V}(F)$, the Zariski tangent space is the ordinary tangent space and its dimension is the dimension of V.

The problem with this calculation is that in practice x is only a numerical approximation to a point of $\mathcal{V}(F)$ and the Jacobian may likely have full rank $\min\{m, n\}$ at x. The notion of numerical rank from numerical analysis suggests a resolution of this problem. We begin with an example.

Example 4.3.14. If we solve the linear equations (4.25) on the set of 2×3 rank one matrices, as in Example 4.3.5, there are three solutions. Here is one,

$$M \;=\; \begin{pmatrix} -9.33788 & -7.74194 & -9.30154 \\ 15.0874 & 12.5089 & 15.0288 \end{pmatrix} .$$

This corresponds to the first solution in (4.23), substituted into the matrix $M(x, y)$. The Jacobian matrix of the three 2×2 minors (4.24) is

$$\begin{pmatrix} M_{2,2} & -M_{2,1} & 0 & -M_{1,2} & M_{1,1} & 0 \\ 0 & M_{2,3} & -M_{2,2} & 0 & -M_{1,3} & M_{1,2} \\ M_{2,3} & 0 & -M_{2,1} & -M_{1,3} & 0 & M_{1,1} \end{pmatrix} ,$$

and evaluating it at the point M gives

$$\begin{pmatrix} 12.5089 & -15.0874 & 0 & 7.74194 & -9.33788 & 0 \\ 0 & 15.0288 & -12.5089 & 0 & 9.30154 & -7.74194 \\ 15.0288 & 0 & -15.0874 & 9.30154 & 0 & -9.33788 \end{pmatrix} . \tag{4.27}$$

This has full rank 3, but the 3×3-minors are all at most 0.026 in absolute value, so the Jacobian matrix is nearly singular.

Numerical analysis furnishes a method to estimate the rank of such a matrix, by determining a nearby singular matrix. A *singular value decomposition* of a complex $m \times n$ matrix M is a factorization $M = UDV^*$, where U and V are unitary matrices of sizes $m \times m$ and $n \times n$, respectively, and D is a $m \times n$ matrix whose only nonzero entries are nonnegative real numbers on the diagonal. The diagonal entries of D are the *singular values* of M. The columns of U are orthonormal eigenvectors of MM^*, the columns of V are the orthonormal eigenvectors of M^*M, and the nonzero singular values are the square roots of the non-zero eigenvalues of both M^*M and MM^*. The *numerical rank* of M is the number of singular values that, when divided by the maximum singular value, exceed a pre-determined thershold.

For example, the matrix (4.27) has singular values $29.045, 29.045, 8.35 \times 10^{-5}$. As the ratios are $1, 1$, and 2.6×10^{-6} with the third about the working precision, we declare its numerical rank to be 2. The singular values for the Jacobian matrices at the other two solutions are $36.12, 36.12$, and 3.85×10^{-5} and $15.79, 15.79$, and 3.06×10^{-5}, so these likewise have numerical rank 2. Refining the approximate solutions to 12 significant digits does not affect the first two singular values, but the third shrinks to about 10^{-11}, so the ratio is again the working precision.

Algorithm 4.3.15 (Local Dimension Test)**.**
Input: A point x on an irreducible component V of $\mathcal{V}(F) \subset \mathbb{C}^n$ and a threshold ϵ.
Output: An estimate for $\dim(V)$.
Do: Compute the singular value decomposition of the Jacobiaan $DF(x)$ of F at x. Let $\sigma_{\max}$ be the maximal singular value and return

$$n \;-\; \#\{\sigma \text{ is a singular value of } DF(x) \text{ with } \sigma > \epsilon\sigma_{\max}\}\,.$$

Explain how this can be used to determine the dimension of an image.
Make an exercise involving 2×4 matrices ?

Exercises

1. Verify the claim in Example 4.3.5 that the system of minors and four linear equations has three solutions. Relate this to Example 4.3.1: Show that the matrix $M(x, y)$ of linear polynomials parametrizes the zero locus of the linear equations (4.25), and that the three rank one matrices obtained in each example are the same.

2. Explain why three linearly independent equations on a six-dimensional space define a subvariety either of dimension three or of dimension four.

3. Verify the claim that in Example 4.3.1 and Example 4.3.5, the variety of rank-one 2×3 matrices was sliced by the same linear subspace.

4. More exercises

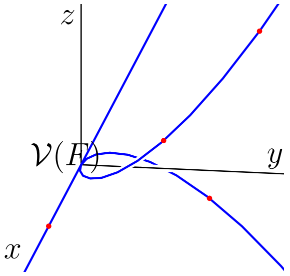
4.4 Numerical Irreducible Decomposition

Section 4.3 introduced witness sets to represent varieties numerically and discussed some algorithms that use witness sets to manipulate varieties. It did not address how to compute a witness set. We offer one method in this section. Here, we explain numerical irreducible decomposition, which begins with a witness set for a (possibly) reducible but equidimensional variety V, decomposing that witness set into witness sets for each irreducible component of V.

Example 4.4.1. Suppose that F is the system of polynomials

$$xy - z \ = \ xz - y^2 \ = \ 0\,,$$

in variables x, y, z for $\mathbb{C}^3$. Substituting the first into the second gives $x^2y - y^2 = 0$ or $y(x^2 - y) = 0$. If $y = 0$ then $z = 0$ and we see that the x-axis $\mathcal{V}(y, z)$ is a subset of $\mathcal{V}(F)$. If $y \neq 0$, then $y = x^2$ so that $z = x^3$, which shows that the moment curve is also a subset of $\mathcal{V}(F)$. In fact, $\langle F\rangle = \langle y, z\rangle \cap \langle x^2 - y, x^3 - z\rangle$, so that $\mathcal{V}(F)$ is the union of these two curves.



Suppose that we were not able to decompose $\mathcal{V}(F)$ by hand, but only had access to a witness set for it. To be specific, as the two polynomials in F are each irreducible, Exercise 9 of Section 3.2 implies that that $\dim(\mathcal{V}(F)) = 1$. If we add the linear equation $x + 2y - 2z = 1$ to F and solve, we obtain the four points

$$(1, 0, 0)\,, \quad (1, 1, 1)\,, \ (\tfrac{1}{\sqrt{2}}, \tfrac{1}{2}, \tfrac{1}{2\sqrt{2}})\,, \ (-\tfrac{1}{\sqrt{2}}, \tfrac{1}{2}, -\tfrac{1}{2\sqrt{2}})\,,$$

which constitute a witness set W for $\mathcal{V}(F)$. A numerical irreducible decomposition of $\mathcal{V}(F)$ is the partition of these points

$$\{(1, 0, 0)\} \ \sqcup \ \{(1, 1, 1)\,, \ (\tfrac{1}{\sqrt{2}}, \tfrac{1}{2}, \tfrac{1}{2\sqrt{2}})\,, \ (-\tfrac{1}{\sqrt{2}}, \tfrac{1}{2}, -\tfrac{1}{2\sqrt{2}})\}$$

into two parts with each part being a witness set for one component of $\mathcal{V}(F)$.

Suppose that $V \subset \mathbb{C}^n$ is a (possibly) reducible variety, all of whose irreducible components have the same dimension, and that (W, F, L) is a witness set for V. Let $V = V_1 \cup V_2 \cup \cdots \cup V_s$ be the decomposition of V into irreducible components. This induces a partition

$$W \ = \ W_1 \sqcup W_2 \sqcup \cdots \sqcup W_s \tag{4.28}$$

of the witness set W with each part the witness set of the corresponding component of W, $W_i = V_i \cap L$. We call this partition (4.28) of a witness set W for V a *numerical irreducible decomposition* of V. We will describe methods to compute and verify a numerical irreducible decomposition for V, given a witness set (W, F, L) for V.

Before describing these methods, let us briefly discuss an algorithm to obtain such a witness set. (We will describe more sophisticated methods in the nest section.) Let $F = \{f_1, \ldots, f_m\}$ be a system of polynomials on $\mathbb{C}^n$. The variety $\mathcal{V}(F)$ many have many components of different dimensions, and we would like to obtain a witness set for the union V of its components of a given dimension d. The following algorithm furnishes such a method.

Algorithm 4.4.2.
<u>Input:</u> A system $F = \{f_1, \ldots, f_m\}$ of polynomials on $\mathbb{C}^n$ and a positive intgeger $d < n$.
<u>Output:</u> A witness set for the union V of components of $\mathcal{V}(F)$ of dimension d.
<u>Do:</u>

1. Select a random subsystem F' of F consisting of $n-d$ polynomials. This uses a variant of Algorithm 4.3.2.

2. Choose d general linear polynomials, L.

3. Use the Bézout Homotopy Algorithm 4.2.7 (or any other method) to compute all isolated solutions W' to the square system $\mathcal{V}(F', L)$.

4. Let $W \subset W'$ be those points of W' that lie on $\mathcal{V}(F)$ and have local dimension d in $\mathcal{V}(F)$. Return (W, F, L).

Proof of correctness. Rewrite this As F' consists of $n - d$ polynomials, the components of $\mathcal{V}(F')$ all havbe dimension d or more. As this is a subsystem of F, $\mathcal{V}(F) \subset \mathcal{V}(F')$, and as it is a random subsystem, with probability one, the components of $\mathcal{V}(F')$ of dimension more than d are components of $\mathcal{V}(F)$ Need a proof of this. Thus every d-dimensional component of $\mathcal{V}(F)$ is a component of $\mathcal{V}(F')$.

The endpoints of the homotopy paths in Step (3) will all be points of $'calV(F') \cap L$, and will include all isolated points. As L is general, it will meet the union of all d-dimensional components of $\mathcal{V}(F')$ transversally in the set of isolated points. Those points that lie in $\mathcal{V}(F)$ anbds for which $\mathcal{V}(F)$ has local dimension d will thus be the points of a witness set $W = V \cap L$ of the union V of the d-dimensional components of $\mathcal{V}(F)$. □

- **Monodromy** Explain how to apply to the example. Find a single loop that permutes the three points.

Lemma Homotopy paths remain on the same irreducible component.

Theorem Monodromy on an irreducible component = full symmetric group.

⟶ The idea is that connectivity of smooth points implies that monodromy is transitive, and the existence of a simple tangent (reduce to plane curves) implies a simple transposition.

Discuss the coarsening algorithm using monodromy, but note that it lacks a stopping criterion when V has two or more components.

- **Trace Test**

Explain the need, apply to the example, and then perhaps to the example from Frank's paper with Anton and Jose.

Prove that trace is linear on a witness set, and it is not linear if the wotness set is incomplete.

Discuss how there is currently no way to certify linearity.

Give the numerical irreducible decomposition algorithm, together with a proof of its correctness.

Perhaps give a meatier example ?

- Witness set of an intersection.

Exercises