Reliable homotopy continuation[*]

Third preliminary version

Joris van der Hoeven

LIX, CNRS
École polytechnique
91128 Palaiseau Cedex
France

Email: `vdhoeven@lix.polytechnique.fr`
Web: `http://lix.polytechnique.fr/~vdhoeven`

January 2, 2015

In this paper, we present several algorithms for certified homotopy continuation. One typical application is to compute the roots of a zero dimensional system of polynomial equations. We both present algorithms for the certification of single and multiple roots. We also present several ideas for improving the underlying numerical path tracking algorithms, especially in the case of multiple roots.

Keywords: Homotopy continuation, polynomial system solving, ball arithmetic, interval arithmetic, reliable computing

A.M.S. subject classification: 65H20, 65H04, 65G20, 30C15, 13P15

Disclaimer. This paper is a preliminary version, which is mainly intended as a working program for an ongoing implementation in the Mathemagix system. It is expected that several adjustments and corrections will have to be made during the implementation of the algorithms in this paper and their proof reading by colleagues.

1. Introduction

Besides Gröbner basis computations, homotopy methods are a popular technique for solving systems of polynomial equations. In this paper, we will only consider zero dimensional systems. Given such a system

$$P(z) = 0, \tag{1}$$

with $P=(P_1,\ldots,P_n)$ and $z=(z_1,\ldots,z_n)$, the idea is to find a suitable starting system

$$Q(z) = 0 \tag{2}$$

of which all solutions are known, to introduce the homotopy

$$H(z,t) = (1-t)\,P(z)+t\,Q(z), \tag{3}$$

[*]. This work has been supported by the ANR-09-JCJC-0098-01 MaGiX project, the Digiteo 2009-36HD grant and Région Ile-de-France.

and to compute the solutions to $P(z)=0$ by following the solutions of $Q(z)=0$ from $t=1$ to $t=0$. Two main approaches exist:

Algebraic homotopies. In this setting, the polynomial equations have exact rational or algebraic coefficients. The homotopy continuation is done exactly using suitable resultants. At the end of the homotopy, the solutions of the system $P(z)=0$ are again given exactly, as the solutions of simpler systems. The theory was developed in [GHMP95, GHH+97, Lec01, Dur08] and a concrete implementation is available in the Kronecker system [Lec01].

Numeric homotopies. An alternative approach is to follow the solution paths using a numeric path tracking algorithm; see [Mor87, Ver96, SW05] and references therein. This approach is usually faster, partly because most of the operations can be done at a significantly lower precision. However, the end result is only approximate. In particular, it cannot be used for the reliable resolution of overdetermined systems. Several implementations exist for numeric path tracking [Ver99, BHSW06, Ley09].

It is surprising that little effort has been undertaken so far in order to bring both approaches closer together. Particularly interesting challenges are how to make numeric homotopy as reliable as possible and how to reconstruct exact end results from the numeric output. Part of this situation might be due to the fact that interval analysis [Moo66, AH83, Neu90, JKDW01, Kul08, MKC09, Rum10] is not so well-known in the communities where homotopy methods were developed, with the exception of one early paper [Kea94]. The main objective paper is to systematically exploit interval analysis techniques in the context of homotopy continuation. We will show how to certify homotopy continuations as well as single and multiple solutions of the polynomial system $P(z)=0$.

Section 3 is devoted to preliminaries from the area of reliable computation. In section 3.1, we start by recalling the basic principles of ball arithmetic [Hoe09], which is a more suitable variant of interval arithmetic for our purposes. In section 3.2, we pursue by recalling the concept of a Taylor model [MB96, MB04], which is useful in order to compute with reliable enclosures of multivariate analytic functions on polydisks. We also introduce a variant of Taylors models in section 3.3, which simultaneously encloses an analytic function and a finite number of its derivatives. In sections 3.4 and 3.5, we discuss the well known problem of overestimation which is inherent to ball arithmetic. We will provide some techniques to analyze, quantify and reduce overestimation.

Before attacking the topic of certified path tracking, it is useful to review the theory of numeric path tracking first. In section 4, we start with the case of non singular paths, in which case we use a classical predictor corrector approach based on Euler-Newton's method. The goal of a numeric path tracker is to advance as fast as possible on the solution path while minimizing the risk of errors. Clearly, the working precision has to be sufficiently large in order to ensure that function evaluations are reasonably accurate. In section 4.4, we show how to find a suitable working precision using ball arithmetic. We consider this approach to be simpler, more robust and more general than the one proposed in [BSHW08]. In order to reduce the risk of jumping from one path to another path, we also need a criterion for checking whether our numeric approximations stay reasonably close to the true solution path. A numerically robust way to do this is to ensure that the Jacobian of H does not change to rapidly during each step; see section 4.5 and [BSHW08] for a related approach. Another technique is to detect near collisions of paths and undertake special action in this case; see section 4.6.

In section 5, we turn our attention to homotopies (3) such that the end system (1) admits multiple solutions. We will see that Euler-Newton iterations only admit a linear convergence near multiple solutions. Therefore, it is useful to search for alternative iterations which admit a better convergence. Now the solution path near a multiple solution is given by a convergent Puiseux series in t. When letting $t \rightarrow \mathrm{e}^{2\pi\mathrm{i}} t$ turn around the origin, we thus fall on another solution path. The collection of paths which are obtained through repeated rotations of this kind is called a herd. In sections 5.2 and 5.3, we will describe a new path tracking method with quadratic convergence, which operates simultaneously on all paths in a herd. The remaining issue of how to detect clusters and herds will be described in sections 5.4, 5.5 and 5.6.

In section 6, we turn our attention to the certification of single roots of (1) and single steps of a path tracker. An efficient and robust method for the certification of solutions to systems of non linear equations is Krawczyk's method [Kra69], with several improvements by Rump [Rum80]. In section 6.1, we adapt this classical method to the setting of ball arithmetic. In section 6.2, we will see that an easy generalization of this method provides an algorithm for certified path tracking. An alternative such algorithm was given in [Kea94], but the present algorithm presents similar advantages as Krawczyk's method with respect to other methods for the certification of solutions to systems of non linear equations. However, both methods still suffer from overestimation due to the fact that error bounds are computed on a polydisk which contains the solution path. Using the technique of Taylor models, we will show in section 6.3 that it possible to compute the error bounds in small tubes around the actual solution path, thereby reducing the problem of overestimation.

In section 7, we first consider the more difficult problem of certifying multiple roots in the univariate case. We will describe two methods based on Rouché's theorem and a third method which rather aims at certifying a local factorization. The last method also serves as an important ingredient for making the Weierstrass preparation theorem effective in an analytic context.

Before we turn our attention to the certification of multiple roots in the multivariate case, it will be convenient to have a general toolbox for effective complex analysis in several variables. In sections 8.1 and 8.2, we first propose two ways how to formalize "computable analytic functions" We next propose several basic algorithms for computations with such functions.

In section 9, we consider the more difficult problem of certifying multiple roots in the multivariate setting. Several algorithms have been proposed for this problem [OWM83, Lec02, GLSY05, GLSY07, LVZ06, LVZ08, RG10, MM11]. However, most existing strategies require the computation of a large number of derivatives of the system, which becomes prohibitive for large clusters of solutions. In the simplest case when the Jacobian matrix of the polynomial system has corank at most one at the singularity, our solution is based on the simultaneous consideration of all solution paths in a herd. The general case will essentially be reduced to this case using analytic elimination techniques which are similar to the geometric resolution method of polynomial systems [GHMP95, GHH+97, Dur08]. In section 9.8, we will also outline a bridge between numeric and algebraic solutions which provides an alternative way to certify solutions.

In section 10, we study generalizations to the resolution of systems of analytic equations in a given polydisk. In section 10.2, we first consider a system of analytic equations as a perturbation of a system of polynomial equations. If, for *each* of the solutions to the system of polynomial equations, we can control the speed with which the solution moves under perturbations, then we can solve the perturbed system. Unfortunately, this forces us to keep track of solutions which are far away from the region of interest. In section 10.3, we present an alternative strategy based on incremental resolution, as in [GHMP95, GHH+97, Dur08]. Although this strategy also may require to work outside the region of interest, it does stay closer.

2. Notations

Positive elements. Given a subset $R \subseteq \mathbb{R} \cup \{\pm\infty\}$, we denote

$$\begin{aligned} R^{\geqslant} &= \{x \in R : x \geqslant 0\} \\ R^{>} &= \{x \in R : x \neq 0\} \end{aligned}$$

Vector notation. Unless stated otherwise, we will use the L_1-norm for vectors $u \in \mathbb{C}^n$:

$$\|u\| = |u_1| + \cdots + |u_n|. \qquad (4)$$

This norm should not be confused with taking componentwise absolute values

$$|u| = (|u_1|, \ldots, |u_n|)$$

For $u, v \in \mathbb{R}^n$ we also define

$$\begin{aligned} u \leqslant v &\Leftrightarrow u_1 \leqslant v_1 \wedge \cdots \wedge u_n \leqslant v_n \\ u < v &\Leftrightarrow u_1 < v_1 \wedge \cdots \wedge u_n < v_n \\ \max(u, v) &= (\max(u_1, v_1), \ldots, \max(u_n, v_n)) \\ u \cdot v &= u_1 v_1 + \cdots + u_n v_n \end{aligned}$$

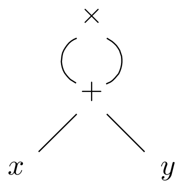
If $z_1, \ldots, z_n$ are formal variables, then we write

$$z^u = z_1^{u_1} \cdots z_n^{u_n}$$

Matrix notation. We write $\mathbb{K}^{r \times c}$ for the set of $r \times c$ matrices over a set $\mathbb{K}$. The matrix norm of a matrix $M \in \mathbb{C}^{r \times c}$ corresponding to the L_1-norm (4) for vectors

$$\begin{aligned} \|M\| &= \sup_{\|z\|=1} \|Mz\| \\ &= \sum_i \max_j |M_{i,j}|. \end{aligned}$$

Directed acyclic graphs. We recall that labeled directed acyclic graphs are often used for the representation of symbolic expressions with potential common subexpressions. For instance,



is a typical dag for the expression $(x + y)^2$. We will denote by s_f the size of a dag f. For instance, the size of the above dag is $s_f = 4$.

3. Reliable arithmetic

3.1. Ball arithmetic

Let us briefly recall the principles behind ball arithmetic. Given a normed vector space $\mathbb{K}$, we will denote by $\boldsymbol{\mathbb{K}}$ or $\mathcal{B}(\mathbb{K}, \mathbb{R})$ the set of closed balls with centers in $\mathbb{K}$ and radii in $\mathbb{R}^{\geqslant}$. Given such a ball $\boldsymbol{z} \in \mathcal{B}(\mathbb{K}, \mathbb{R})$, we will denote its center by $\operatorname{cen}(\boldsymbol{z})$ and its radius by $\operatorname{rad}(\boldsymbol{z})$. Conversely, given $z \in \mathbb{K}$ and $r \in \mathbb{R}$, we will denote by $z + \mathcal{B}(r)$ the closed ball with center z and radius r.

A continuous operation $f: \mathbb{K}^d \rightarrow \mathbb{K}$ is said to *lift* into an operation $f^{\text{lift}}: \boldsymbol{\mathbb{K}}^d \rightarrow \boldsymbol{\mathbb{K}}$ on balls, which is usually also denoted by f, if the *inclusion property*

$$f(x_1, \ldots, x_d) \;\in\; f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) \tag{5}$$

is satisfied for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \boldsymbol{\mathbb{K}}$ and $x_1 \in \boldsymbol{x}_1, \ldots, x_d \in \boldsymbol{x}_d$. We also say that $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ is an *enclosure* for the set $\{f(x_1, \ldots, x_d): x_1 \in \boldsymbol{x}_1, \ldots, x_d \in \boldsymbol{x}_d\}$, whenever (5) holds. For instance, if $\mathbb{K}$ is a Banach algebra, then we may take

$$\begin{aligned}
\boldsymbol{x} + \boldsymbol{y} &= \operatorname{cen}(\boldsymbol{x}) + \operatorname{cen}(\boldsymbol{y}) + \mathcal{B}(\operatorname{rad}(\boldsymbol{x}) + \operatorname{rad}(\boldsymbol{y}))\\
\boldsymbol{x} - \boldsymbol{y} &= \operatorname{cen}(\boldsymbol{x}) - \operatorname{cen}(\boldsymbol{y}) + \mathcal{B}(\operatorname{rad}(\boldsymbol{x}) + \operatorname{rad}(\boldsymbol{y}))\\
\boldsymbol{x}\,\boldsymbol{y} &= \operatorname{cen}(\boldsymbol{x}) \operatorname{cen}(\boldsymbol{y}) + \mathcal{B}(\operatorname{rad}(\boldsymbol{x})\,(|\operatorname{cen}(\boldsymbol{y})| + \operatorname{rad}(\boldsymbol{y})) + |\operatorname{cen}(\boldsymbol{y})| \operatorname{rad}(\boldsymbol{x})).
\end{aligned}$$

Similar formulas can be given for division and elementary functions. Certified upper and lower bounds for $|\boldsymbol{x}|$ will be denoted by $\lceil \boldsymbol{x} \rceil = |\operatorname{cen}(\boldsymbol{x})| + \operatorname{rad}(\boldsymbol{x})$ and $\lfloor \boldsymbol{x} \rfloor = \max\,\{0, |\operatorname{cen}(\boldsymbol{x})| - \operatorname{rad}(\boldsymbol{x})\}$.

It is convenient to extend the notion of a ball to more general radius types, which only carry a partial ordering. This allows us for instance to regard a vector $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) \in \mathcal{B}(\mathbb{K}, \mathbb{R})^d$ of balls as a "vectorial ball" with center $\operatorname{cen}(\boldsymbol{x}) = (\operatorname{cen}(\boldsymbol{x}_1), \ldots, \operatorname{cen}(\boldsymbol{x}_d)) \in \mathbb{K}^d$ and radius $\operatorname{rad}(\boldsymbol{x}) = (\operatorname{rad}(\boldsymbol{x}_1), \ldots, \operatorname{rad}(\boldsymbol{x}_d)) \in \mathbb{R}^d$. If $x = (x_1, \ldots, x_d) \in \mathbb{K}^d$, then we write $x \in \boldsymbol{x}$ if and only if $x_i \in \boldsymbol{x}_i$ for all $i \in \{1, \ldots, d\}$. A similar remark holds for matrices and power series with ball coefficients.

In concrete machine computations, numbers are usually approximated by floating point numbers with a finite precision. Let $\tilde{\mathbb{R}}$ be the set of floating point numbers at a given working precision, which we will assume fixed. It is customary to include the infinities $\pm\infty$ in $\tilde{\mathbb{R}}$ as well. The IEEE754 standard [ANS08] specifies how to perform basic arithmetic with floating point numbers in a predictable way, by specifying a rounding mode $\mathsf{R} \in \{\downarrow, \uparrow, \updownarrow\}$ among "down", "up" and "nearest". A multiple precision implementation of this standard is available in the Mpfr library [HLRZ00]. Given an operation $f: \mathbb{R}^d \rightarrow \mathbb{R}$, we will denote by $f^{\mathsf{R}}: \tilde{\mathbb{R}}^d \rightarrow \tilde{\mathbb{R}}$ its approximation using floating pointing arithmetic with rounding mode R. This notation extends to the case when $\mathbb{R}$ and $\tilde{\mathbb{R}}$ are replaced by their complexifications $\mathbb{C}$ and $\tilde{\mathbb{C}} = \tilde{\mathbb{R}}[\mathrm{i}]$.

Let $\mathbb{K} = \mathbb{R}$ and $\tilde{\mathbb{K}} = \tilde{\mathbb{R}}$ or $\mathbb{K} = \mathbb{C}$ and $\tilde{\mathbb{K}} = \tilde{\mathbb{C}}$. We will denote by $\tilde{\boldsymbol{\mathbb{K}}}$ or $\mathcal{B}(\tilde{\mathbb{K}}, \tilde{\mathbb{R}})$ the set of closed balls in $\mathbb{K}$ with centers in $\tilde{\mathbb{K}}$ and radii in $\tilde{\mathbb{R}}^{\geqslant}$. In this case, we will also allow for balls with an infinite radius. A continuous operation $f: \mathbb{K}^d \rightarrow \mathbb{K}$ is again said to *lift* to an operation $f: \tilde{\boldsymbol{\mathbb{K}}}^d \rightarrow \tilde{\boldsymbol{\mathbb{K}}}$ on balls if (5) holds for any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \tilde{\boldsymbol{\mathbb{K}}}$ and $x_1 \in \boldsymbol{x}_1, \ldots, x_d \in \boldsymbol{x}_d$. The formulas for the ring operations may now be adapted to

$$\begin{aligned}
\boldsymbol{x} + \boldsymbol{y} &= \operatorname{cen}(\boldsymbol{x}) +^{\updownarrow} \operatorname{cen}(\boldsymbol{y}) + \mathcal{B}(\operatorname{rad}(\boldsymbol{x}) +^{\uparrow} \operatorname{rad}(\boldsymbol{y}) +^{\uparrow} \epsilon_{+,\boldsymbol{x},\boldsymbol{y}})\\
\boldsymbol{x} - \boldsymbol{y} &= \operatorname{cen}(\boldsymbol{x}) -^{\updownarrow} \operatorname{cen}(\boldsymbol{y}) + \mathcal{B}(\operatorname{rad}(\boldsymbol{x}) +^{\uparrow} \operatorname{rad}(\boldsymbol{y}) +^{\uparrow} \epsilon_{-,\boldsymbol{x},\boldsymbol{y}})\\
\boldsymbol{x}\,\boldsymbol{y} &= \operatorname{cen}(\boldsymbol{x}) \times^{\updownarrow} \operatorname{cen}(\boldsymbol{y}) +\\
&\quad\ \mathcal{B}(\operatorname{rad}(\boldsymbol{x}) \times^{\uparrow} (|\operatorname{cen}(\boldsymbol{y})| +^{\uparrow} \operatorname{rad}(\boldsymbol{y})) +^{\uparrow} |\operatorname{cen}(\boldsymbol{y})| \times^{\uparrow} \operatorname{rad}(\boldsymbol{x}) +^{\uparrow} \epsilon_{\times,\boldsymbol{x},\boldsymbol{y}}),
\end{aligned}$$

where $\epsilon_{+,\boldsymbol{x},\boldsymbol{y}}$, $\epsilon_{-,\boldsymbol{x},\boldsymbol{y}}$ and $\epsilon_{\times,\boldsymbol{x},\boldsymbol{y}}$ are reliable bounds for the rounding errors induced by the corresponding floating point operations on the centers; see [Hoe09] for more details.

In order to ease the remainder of our exposition, we will avoid technicalities related to rounding problems, and compute with "idealized" balls with centers in $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and radii in $\mathbb{R}^{\geqslant}$. For those who are familiar with rounding errors, it should not be difficult though to adapt our results to more realistic machine computations.

Remark 1. In classical interval analysis so called interval lifts of operations $f: \mathbb{K}^d \rightarrow \mathbb{K}$ are sometimes required to satisfy the *inclusion monotonicity* property

$$\boldsymbol{x}_1 \subseteq \boldsymbol{y}_1 \wedge \cdots \wedge \boldsymbol{x}_d \subseteq \boldsymbol{y}_d \implies f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) \subseteq f(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d),$$

for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{K}$, which clearly implies the usual inclusion property (5). For floating intervals, it is easy to ensure this stronger property using correct rounding. In the ball setting, the exact ring operations in $\mathbf{R}$ and $\mathbb{C}$ are clearly inclusion monotonic, but it seems cumbersome to preserve this stronger property for floating balls. For this reason, we systematically develop our theory without assuming inclusion monotonicity.

3.2. Taylor models

If we are computing with analytic functions on a disk, or multivariate analytic functions on a polydisk, then Taylor models [MB96, MB04] provide a suitable functional analogue for ball arithmetic. We will use a multivariate setup with $z = (z_1, \ldots, z_d)$ as our coordinates and a polydisk $D = \mathcal{B}(\rho) = \{z, |z| \leqslant |\rho|\}$ for a fixed $\rho = (\rho_1, \ldots, \rho_d) \in (\mathbb{R}^>)^d$. Taylor models come in different blends, depending on whether we use a global error bound on D or individual bounds for the coefficients of the polynomial approximation. Individual bounds are sharper (especially if we truncate up to an small order such that the remainder is not that small), but more expensive to compute. Our general setup covers all possible blends of Taylor models.

We first need some more definitions and notations. Assume that $\mathbb{N}^d$ is given the natural partial ordering. Let $\mathbb{e}_k$ denote the k-th canonical basis vector of $\mathbb{N}^d$, so that $(\mathbb{e}_k)_k = 1$ and $(\mathbb{e}_k)_l = 0$ for $l \neq k$. For every $i \in \mathbb{N}^d$, recall that $\|i\| = |i_1| + \cdots + |i_d|$. A subset $\mathcal{I} \subseteq \mathbb{N}^d$ is called an *initial segment*, if for any $i \in \mathcal{I}$ and $j \in \mathbb{N}^d$ with $j \leqslant i$, we have $j \in \mathcal{I}$. In that case, we write $\mathring{\mathcal{I}} = \{i \in \mathcal{I}: i + \{\mathbb{e}_1, \ldots, \mathbb{e}_d\} \subseteq \mathcal{I}\}$ and $\partial \mathcal{I} = \mathcal{I} \setminus \mathring{\mathcal{I}}$. In what follows, we assume that $\mathcal{I}$ and $\mathcal{J}$ are fixed initial segments of $\mathbb{N}^d$ with $\mathring{\mathcal{J}} \subseteq \mathcal{I}$. For instance, we may take $\mathcal{I} = \mathcal{T}_n = \{i \in \mathbb{N}^d: \|i\| \leqslant n\}$ and $\mathcal{J} = \mathcal{T}_{n+1}$ or $\mathcal{J} = \mathcal{T}_n$ or $\mathcal{J} = \{0\}$.

Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Given a series $f = \sum_{i \in \mathbb{N}^d} f_i z^i \in \mathbb{K}[[z]]$, we will write $\operatorname{supp} f = \{i \in \mathbb{N}^d: f_i \neq 0\}$ for its *support*. Given a subset $\mathbb{S} \subseteq \mathbb{K}[[z]]$ and a subset $\mathcal{S} \subseteq \mathbb{N}^d$, we write $f_{\mathcal{S}} = \sum_{i \in \mathcal{S}} f_i z^i$ and $\mathbb{S}_{\mathcal{S}} = \{g \in \mathbb{S}: \operatorname{supp} g \subseteq \mathcal{S}\}$. If f is analytic on D, then we denote its sup-norm by

$$\|f\|_D = \sup_{z \in D} |f(z)|.$$

A *Taylor model* is a tuple $\boldsymbol{P} = (\rho, \mathcal{I}, \mathcal{J}, \operatorname{cen}(\boldsymbol{P}), \operatorname{rad}(\boldsymbol{P}))$, where ρ, $\mathcal{I}$ and $\mathcal{J}$ are as above, $\operatorname{cen}(\boldsymbol{P}) \in \mathbb{K}[z]_{\mathcal{I}}$ and $\operatorname{rad}(\boldsymbol{P}) \in \mathbb{R}[z]_{\mathcal{J}}$. We will write $\mathbb{T} = \mathbb{T}_{D,\mathcal{I},\mathcal{J}} = \mathcal{B}_D(\mathbb{K}[z]_{\mathcal{I}}, \mathbb{R}[z]_{\mathcal{J}})$ for the set of such Taylor models. Given $\boldsymbol{P} \in \mathbb{T}$ and $i \in \mathbb{N}^d$, we will also denote $\boldsymbol{P} = \operatorname{cen}(\boldsymbol{P}) + \mathcal{B}_D(\operatorname{rad}(\boldsymbol{P}))$ and $\boldsymbol{P}_i = \operatorname{cen}(\boldsymbol{P})_i + \mathcal{B}(\operatorname{rad}(\boldsymbol{P})_i)$. Given an analytic function f on D, we write $f \in \boldsymbol{P}$, if there exists a decomposition

$$f = \operatorname{cen}(\boldsymbol{P}) + \sum_{i \in \mathcal{J}} \varepsilon_i z^i$$

with $\varepsilon_i \in \mathbb{C}[[z]]$ and $\|\varepsilon_i\|_D \leqslant \operatorname{rad}(\boldsymbol{P})_i$ for all i. In particular, if $f \in \boldsymbol{P}$, then

$$f(z) \in \sum_{i \in \mathcal{I} \cup \mathcal{J}} \boldsymbol{P}_i z^i,$$

for any $z \in D$. Given two Taylor models $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{T}$, we will say that $\boldsymbol{P}$ is *included* in $\boldsymbol{Q}$, and we write $\boldsymbol{P} \subseteq \boldsymbol{Q}$ if $f \in \boldsymbol{Q}$ for any $f \in \boldsymbol{P}$. This holds in particular if $\boldsymbol{P}_i \subseteq \boldsymbol{Q}_i$ for all i, in which case we say that $\boldsymbol{P}$ is *strongly included* in $\boldsymbol{Q}$ and write $P \sqsubseteq Q$. We finally define $\varpi(\boldsymbol{P}) \in \mathbb{C}$ by

$$\varpi(\boldsymbol{P}) = \boldsymbol{P}_0 + \sum_{i \neq 0} \boldsymbol{P}_i \mathcal{B}(\rho)^i,$$

so that $f(z) \in \varpi(\boldsymbol{P})$ for all $f \in \boldsymbol{P}$ and $z \in \mathcal{B}(\rho)$.

Addition, subtraction and scalar multiplication are defined in a natural way on Taylor models. For multiplication, we need a projection $\pi=\pi_{\mathcal{J}}\colon \mathbb{N}^d\to\mathcal{J}$ with $\pi(i)\leqslant i$ for all i and $\pi(i)=i$ if $i\in\mathcal{J}$. One way to construct such a mapping is as follows. For $i\in\mathcal{J}$, we must take $\pi(i)=i$. For $i\notin\mathcal{J}$, let k be largest such that $i_k\neq 0$. Then we recursively define $\pi(i)=\pi(i-\mathbb{e}_k)$. Given $\boldsymbol{P},\boldsymbol{Q}\in\mathbb{T}$, we now define their product by

$$\boldsymbol{P}\boldsymbol{Q} \;=\; \sum_{i,j\in\mathcal{I}} \boldsymbol{P}_i\,\boldsymbol{Q}_j\,\mathcal{B}_D(\rho)^{i+j-\pi(i+j)}\,z^{\pi(i+j)}.$$

Using the observation that $z^{i+j}\in\mathcal{B}_D(\rho)^{i+j-\pi(i+j)}\,z^{\pi(i+j)}$, this product satisfies the inclusion property that $fg\in\boldsymbol{P}\boldsymbol{Q}$ for any analytic functions $f\in\boldsymbol{P}$ and $Q\in\boldsymbol{Q}$ on D.

3.3. $\mathcal{D}$-stable Taylor models

For some applications, it is convenient to use Taylor models for enclosing both an analytic function and a certain number of its derivatives. Let us show how to incorporate this in our formalism. Throughout this section, we assume that $\mathcal{I}=\mathcal{J}$ and that $\mathcal{D}$ is an initial segment with $\mathcal{D}\subseteq\mathcal{I}$.

Given a Taylor model $\boldsymbol{P}\in\mathbb{T}_{D,\mathcal{I},\mathcal{I}}$ and $i\in\mathcal{D}$, we notice that $\partial^i\boldsymbol{P}/\partial z^i$ can be regarded as a Taylor model in $\mathbb{T}_{D,\mathcal{I}',\mathcal{I}'}$ with $\mathcal{I}'=\{j\in\mathbb{N}^n\colon i+j\subseteq\mathcal{I}\}$. Let $f\in D\to\mathbb{C}$ be an analytic function and $\boldsymbol{Q}\in\mathbb{T}_{D,\mathcal{I},\mathcal{I}}$. We define the relations $\in_{\mathcal{D}}$ and $\subseteq_{\mathcal{D}}$ by

$$\begin{aligned} f\in_{\mathcal{D}}\boldsymbol{P} &\iff \forall i\in\mathcal{D},\frac{\partial^i f}{\partial z^i}\subseteq\frac{\partial^i\boldsymbol{P}}{\partial z^i}\\ \boldsymbol{P}\subseteq_{\mathcal{D}}\boldsymbol{Q} &\iff \forall f\in_{\mathcal{D}}\boldsymbol{P}, f\in_{\mathcal{D}}\boldsymbol{Q}.\end{aligned}$$

Clearly, $\boldsymbol{P}\sqsubseteq\boldsymbol{Q}\Rightarrow\boldsymbol{P}\subseteq_{\mathcal{D}}\boldsymbol{Q}$ for all $\boldsymbol{P}$ and $\boldsymbol{Q}$.

Let $\omega\colon\mathbb{C}^d\to\mathbb{C}$ be an operation. Then ω is said to $\mathcal{D}$-lift to $\mathbb{T}_{D,\mathcal{I},\mathcal{I}}$, if for all $\boldsymbol{P}_1,\ldots,\boldsymbol{P}_d\in\mathbb{T}_{D,\mathcal{I},\mathcal{I}}$ and all $f_1\in_{\mathcal{D}}\boldsymbol{P}_1,\ldots,f_d\in_{\mathcal{D}}\boldsymbol{P}_d$, we have $\omega\circ(f_1,\ldots,f_d)\in_{\mathcal{D}}\omega(\boldsymbol{P}_1,\ldots,\boldsymbol{P}_d)$. Addition, subtraction and scalar multiplication $\mathcal{D}$-lift in the usual way. As to multiplication, we take

$$\boldsymbol{P}\times_{\mathcal{D}}\boldsymbol{Q} \;=\; \sum_{\substack{i,j\in\mathcal{I}\\ i+j\in\mathcal{I}}}\boldsymbol{P}_i\,\boldsymbol{Q}_j\,z^{i+j}+\sum_{\substack{i,j\in\mathcal{I}\\ i+j\notin\mathcal{I}\\ k\in\partial\mathcal{I}\\ k\leqslant i+j}} C_{i,j,k}\,\boldsymbol{P}_i\,\boldsymbol{Q}_j\,\mathcal{B}_D(\rho)^{i+j-k}\,z^k,$$

with

$$C_{i,j,k} \;=\; \max_{\substack{l\in\mathcal{D}\\ l\leqslant k}}\left(\frac{\partial^l z^{i+j}}{z^{i+j-l}\,\partial z^l}\middle/\frac{\partial^l z^k}{z^{k-l}\,\partial z^l}\right).$$

In order to see that $\times_{\mathcal{D}}$ satisfies the $\mathcal{D}$-inclusion property, it suffices to check that

$$z^{i+j}\;\in_{\mathcal{D}}\; z^i\times_{\mathcal{D}}z^j$$

for all $i,j\in\mathcal{I}$. This is clear if $i+j\in\mathcal{I}$. Otherwise,

$$z^i\times_{\mathcal{D}}z^j \;=\; \sum_{\substack{k\in\partial\mathcal{I}\\ k\leqslant i+j}} C_{i,j,k}\,\mathcal{B}_D(\rho)^{i+j-k}\,z^k.$$

For any $l\in\mathcal{D}$ with $l\leqslant i+j$, there exists a $k\in\partial\mathcal{I}$ with $l\leqslant k\leqslant i+j$. Hence,

$$\begin{aligned}\frac{\partial^l z^{i+j}}{\partial z^l} &= \frac{\partial^l z^{i+j}}{z^{i+j-l}\,\partial z^l}\,z^{i+j-l}\\ &\in \frac{\partial^l z^{i+j}}{z^{i+j-l}\,\partial z^l}\,\mathcal{B}_D(\rho)^{i+j-k}\,z^{k-l}\\ &= \frac{\partial^l}{\partial z^l}\left[\left(\frac{\partial^l z^{i+j}}{z^{i+j-l}\,\partial z^l}\middle/\frac{\partial^l z^k}{z^{k-l}\,\partial z^l}\right)\mathcal{B}_D(\rho)^{i+j-k}\,z^k\right]\\ &\subseteq \frac{\partial^l(z^i\times_{\mathcal{D}}z^j)}{\partial z^l}.\end{aligned}$$

In the particularly useful case when $\mathcal{I}=\mathcal{J}=\mathcal{T}_1=\{i\in\mathbb{N}^d\colon \|i\|\leqslant 1\}$, we notice that $C_{e_i,e_j,e_i}=C_{e_i,e_j,e_j}=1$ for all $i\neq j$ and $C_{e_i,e_i,e_i}=2$ for all i.

3.4. Overestimation

The major problem in the area of ball arithmetic is overestimation. For example, even though the expression $x-x$ evaluates to zero for any $x\in\mathbb{R}$, its evaluation at any ball in $\mathbf{R}$ with a non zero radius is not identically equal to zero. For instance,

$$(1+\mathcal{B}(0.1))-(1+\mathcal{B}(0.1)) \;=\; \mathcal{B}(0.2).$$

Algorithms which rely on ball arithmetic have to be designed with care in order to avoid this kind of overly pessimistic error bounds. In particular, if we evaluate a dag using ball arithmetic, then a symbolically equivalent dag might lead to better error bounds.

Consider a continuous function $f\colon\mathbb{K}^d\to\mathbb{K}$ with $\mathbb{K}$ as in section 3.1. We recall that f is said to lift into an operation $f^{\text{lift}}\colon\mathbf{K}^d\to\mathbf{K}$ if the inclusion property

$$f(x) \;\in\; f^{\text{lift}}(\boldsymbol{x})$$

is satisfied for all $\boldsymbol{x}\in\mathbf{K}^d$ and $x\in\boldsymbol{x}$. Clearly, such a lift is not unique: for any $\varepsilon\colon\mathbf{K}^d\to\mathbf{K}$ with $\operatorname{cen}\varepsilon(\boldsymbol{x})=0$ for all $\boldsymbol{x}$, the function $f^{\text{alt}}=f^{\text{lift}}+\varepsilon$ is also a lift of f. If we require that $\operatorname{cen} f^{\text{lift}}(\boldsymbol{x})=f(\operatorname{cen}\boldsymbol{x})$, then the best possible lift is given by

$$f^{\text{best}}(\boldsymbol{x}) \;=\; f(\operatorname{cen}\boldsymbol{x})+\mathcal{B}(\sup_{x'\in\boldsymbol{x}}|f(x')-f(\operatorname{cen}\boldsymbol{x})|).$$

In general, this lift may be expensive to compute. Nevertheless, its existence suggest the following definition of the quality of a lift. The *overestimation* $\chi_{f^{\text{lift}}}(\boldsymbol{x})$ of f^{lift} at $\boldsymbol{x}$ is defined by

$$\chi_{f^{\text{lift}}}(\boldsymbol{x}) \;=\; \frac{\operatorname{rad} f^{\text{lift}}(\boldsymbol{x})}{\operatorname{rad} f^{\text{best}}(\boldsymbol{x})}. \tag{6}$$

This quantity is easier to study if we let $\operatorname{rad}\boldsymbol{x}$ tend to zero. Accordingly, we also define the *pointwise overestimation* function $\chi_{f^{\text{lift}}}\colon\mathbb{K}^d\to\mathbb{R}^{\geqslant}$ by

$$\chi_{f^{\text{lift}}}(x) \;=\; \limsup_{\boldsymbol{\varepsilon}\to 0}\chi_{f^{\text{lift}}}(x+\boldsymbol{\varepsilon}). \tag{7}$$

Here $\boldsymbol{\varepsilon}\to 0$ means that $\operatorname{cen}\boldsymbol{\varepsilon}=0$ and $\operatorname{rad}\boldsymbol{\varepsilon}\to 0$.

If f^{lift} is computed by evaluating a dag f, then it would be nice to have explicit formulas for the pointwise overestimation. For $\operatorname{rad}\boldsymbol{x}\to 0$ and assuming that the lift f^{std} is evaluated using the default ball implementations of $+$, $-$ and $\times$ from section 3.1 , we claim that there exists a dag $\bar{\nabla} f$ with

$$\operatorname{rad} f^{\text{std}}(x+\mathcal{B}(\varepsilon)) \;=\; (\bar{\nabla} f)\cdot|\varepsilon|+\mathcal{O}(\varepsilon^2),$$

for $\varepsilon\to 0$. Indeed, we may compute $\bar{\nabla} f$ using the rules

$$\begin{aligned}
\bar{\nabla} c &= (0,\ldots,0) && (c\in\mathbb{K})\\
\bar{\nabla} X_k &= (0,\overset{k-1}{\ldots},0,1,0,\ldots,0) && (k\in\{1,\ldots,r\})\\
\bar{\nabla}(f\pm g) &= \bar{\nabla} f+\bar{\nabla} g\\
\bar{\nabla}(fg) &= (\bar{\nabla} f)\,|g|+|f|\,(\bar{\nabla} g),
\end{aligned}$$

where X_k stands for the k-th coordinate function. Now we also have

$$\operatorname{rad} f^{\text{best}}(x+\mathcal{B}(\varepsilon)) \;=\; |\nabla f|\cdot|\varepsilon|+\mathcal{O}(\varepsilon^2),$$

for $\varepsilon \rightarrow 0$. Consequently,

$$\chi_{f^{\text{std}}}(x) \;=\; \limsup_{\varepsilon \neq 0} \frac{(\bar{\nabla} f)(x) \cdot |\varepsilon|}{|(\nabla f)(x)| \cdot |\varepsilon|}.$$

If $d = 1$, then this formula simplifies to

$$\chi_{f^{\text{std}}}(x) \;=\; \frac{(\bar{\nabla} f)(x)}{|f'(x)|}.$$

Example 2. With $d = 1$, let us compare the dags $f = X^2 - 2X + 1$ and $g = (X - 1)^2$. We have $\bar{\nabla} f = 2X + 2$ and $\bar{\nabla} g = 2|X - 1|$, whence

$$\begin{aligned} \chi_{f^{\text{std}}}(x) &= \frac{|x| + 1}{|x - 1|} \\ \chi_{g^{\text{std}}}(x) &= 1. \end{aligned}$$

The example shows that we have an infinite amount of overestimation near double zeros, except if the dag is explicitly given as a square near the double zero. More generally, for the dag $f = X^n - n X^{n-1} + \binom{n}{2} X^{n-2} + \cdots + (-1)^n$ with an n-fold zero, we obtain

$$\chi_{f^{\text{std}}}(x) \;=\; \frac{(|x| + 1)^{n-1}}{|x - 1|^{n-1}}.$$

At a distance ε of the zero, ball arithmetic thus produces bounds which are $(2/\varepsilon)^{n-1}$ times too pessimistic.

Remark 3. An interesting problem is whether a good understanding of the pointwise overestimation also helps us to bound the overestimation on more general balls. One concrete question is whether we have

$$\operatorname{rad} f^{\text{std}}(\boldsymbol{x}) \;\leqslant\; \left(\sup_{x \in \boldsymbol{x}} \chi_{f^{\text{std}}}(x) \right) \operatorname{rad} f^{\text{best}}(\boldsymbol{x}),$$

for all polynomial dags f and balls $\boldsymbol{x}$. This inequality seems to hold in all easy cases that we have looked at, but we do not have a proof that it holds in general.

3.5. Reducing the overestimation

The example 2 shows that standard ball arithmetic generally produces an infinite amount of overestimation near double or multiple zeros. This raises the problem how to compute better ball lifts which do not present this drawback.

One possible remedy is to systematically compute the ball lifts using Taylor models. Indeed, assume that we want to evaluate f at the ball $\boldsymbol{x} = c + \mathcal{B}(\rho)$. Let $D = \mathcal{B}(\rho)$, $\mathcal{I}$ and $\mathcal{J}$ be as in section 3.2 and let $\mathbb{T} = \mathcal{B}_D(\mathbb{K}[\epsilon]_{\mathcal{I}}, \mathbb{R}[\epsilon]_{\mathcal{J}})$ be the corresponding domain of Taylor models in $\epsilon = (\epsilon_1, \ldots, \epsilon_d)$. Let $\boldsymbol{\xi} = (x_1 + \epsilon_1, \ldots, x_n + \epsilon_d) \in \mathbb{T}^d$ and consider the Taylor model evaluation of f at $\boldsymbol{\xi}$

$$f(\boldsymbol{\xi}) \;=\; P + \mathcal{B}_D(E).$$

Then

$$f^{\text{tay}}(\boldsymbol{x}) \;:=\; P^{\text{std}}(\mathcal{B}(\rho)) + \mathcal{B}(\lceil E^{\text{std}}(\mathcal{B}(\rho)) \rceil)$$

yields an enclosure of $\{f(x) : x \in \boldsymbol{x}\}$. Although the evaluation of $f^{\text{tay}}(\boldsymbol{x})$ is usually far more expensive than the evaluation of $f^{\text{std}}(\boldsymbol{x})$, let us now study how much the overestimation has been reduced.

Let $\mathcal{F} = \mathbb{N}^d \setminus \mathcal{I}$ and let us introduce the operator $\bar{D}^{\mathcal{F}}: \varepsilon \mapsto \bar{D}^{\mathcal{F}}(\varepsilon)$, which generalizes the mapping $\varepsilon \mapsto \varepsilon \cdot \bar{\nabla}$. The operator is defined by induction over the size of f:

$$\begin{aligned}
(\bar{D}^{\mathcal{F}} c)(\varepsilon) &= 0 && (c \in \mathbb{K})\\
(\bar{D}^{\mathcal{F}} X_k)(\varepsilon) &= \begin{cases} 0 & \text{if } (0, \overset{k-1}{\ldots}, 0, 1, 0, \ldots, 0) \in \mathcal{I}\\ \varepsilon_k & \text{otherwise}\end{cases} && (k \in \{1, \ldots, r\})\\
(\bar{D}^{\mathcal{F}}(f \pm g))(\varepsilon) &= (\bar{D}^{\mathcal{F}} f)(\varepsilon) + (\bar{D}^{\mathcal{F}} g)(\varepsilon)\\
(\bar{D}^{\mathcal{F}}(f g))(\varepsilon) &= ((\bar{D}^{\mathcal{F}} f)\,|g| + |f|\,(\bar{D}^{\mathcal{F}} g))(\varepsilon) + \sum_{\substack{i, j \in \mathcal{I} \setminus \{0\}\\ i + j \in \mathcal{F}}} \frac{\varepsilon^{i+j}}{i!\, j!}\,|f^{(i)}|\,|g^{(j)}|
\end{aligned}$$

For $\boldsymbol{\xi} = (x_1 + \epsilon_1, \ldots, x_n + \epsilon_d)$ as above, we then have

$$f(\boldsymbol{\xi}) \subseteq \left(\sum_{i \in \mathcal{I}} \frac{1}{i!} f^{(i)}(x)\,\epsilon^i \right) + \mathcal{B}_D((\bar{D}^{\mathcal{F}} f)(\rho)).$$

Now assume that $\mathcal{I} = \mathcal{J} = \mathcal{T}_n = \{i \in \mathbb{N}^d : \|i\| \leqslant n\}$ and let μ be the valuation of f at x. If $\mu < n$, then we have

$$\operatorname{rad} f^{\text{best}}(x + \mathcal{B}(\rho)) = \sup_{\varepsilon \in \mathcal{B}(r)} \left| \sum_{\|i\| = \mu} \frac{1}{i!} f^{(i)}(x)\,\rho^i \right| + \mathcal{O}(\rho^{\mu+1}) \tag{8}$$

$$\operatorname{rad} f^{\text{tay}}(x + \mathcal{B}(\rho)) = \sup_{\varepsilon \in \mathcal{B}(r)} \left| \sum_{\|i\| = \mu} \frac{1}{i!} f^{(i)}(x)\,\rho^i \right| + \mathcal{O}(\rho^{\mu+1}) \tag{9}$$

$$\chi_{f^{\text{tay}}}(x) = 1. \tag{10}$$

If $\mu = n$, then we still have (8), but (9) and (10) become

$$\begin{aligned}
\operatorname{rad} f^{\text{tay}}(x + \mathcal{B}(\rho)) &\leqslant (\bar{D}^{\mathcal{F}} f)(\rho)\\
\chi_{f^{\text{tay}}}(x) &= \operatorname*{limsup}_{\varepsilon \neq 0} \frac{(\bar{D}^{\mathcal{F}} f)(\varepsilon)}{\left| \sum_{\|i\| = n} \frac{1}{i!} f^{(i)}(x)\,\varepsilon^i \right|}.
\end{aligned}$$

If $\mu > n$, then we generally have

$$\chi_{f^{\text{tay}}}(x) = \infty,$$

although $\chi_{f^{\text{tay}}}(x) < \infty$ may occur in lucky cases.

4. Numeric path tracking

4.1. General framework

Let Ω be an open subset of $\mathbb{C}^n$ and $H: \Omega \times \mathbb{C} \rightarrow \mathbb{C}^n$ an analytic function. We consider H as a function $H(z, t)$ in z and the *time* t, where $z \in \Omega$ and $t \in \mathbb{C}$, and also call H a *homotopy*. Assuming that $H(z_1, 1) = 0$ for some $z_1 \in \Omega$ and that we are not in a “degenerate” case, there exists a unique analytic function $[0, 1] \rightarrow \Omega: t \mapsto z_t$ with $H(z_t, t) = 0$ for all t. We are interested in the value of z_t when $t \rightarrow 0$. More generally, given a vector $z_1 = (z_1^1, \ldots, z_1^k) \in \Omega^k$ of vectors, there exists a unique function $[0, 1] \rightarrow \Omega^k: t \mapsto z_t$ with $H(z_t^1, t) = \cdots = H(z_t^k, t) = 0$ for all t.

The goal of a numeric path tracker is to approximate the function $t \mapsto z_t$ as well and as quickly possible and, above all, to compute its value z_0 at the “end point” $t = 0$. In what follows, we will denote by $\tilde{\mathbb{R}} = \mathbb{R}_p$ the set of floating point numbers with p bit mantissas. We also define $\tilde{\mathbb{C}} = \mathbb{C}_p = \mathbb{R}_p[\mathrm{i}]$, $\tilde{\Omega} = \Omega \cap \tilde{\mathbb{C}}$ and assume that we have a program for computing a numeric approximation $\tilde{H}: \tilde{\Omega} \times \tilde{\mathbb{C}} \rightarrow \tilde{\mathbb{C}}^n$ of H. Given $z_1 \in \tilde{\Omega}$ with $\tilde{H}(z_1, 1) \approx 0$, we thus want to compute $z_0 \in \tilde{\Omega}$ with $\tilde{H}(z_0, 0) \approx 0$, by following the homotopy.

In many cases, we will be interested in homotopies for solving a system

$$P_1(z) = \cdots = P_n(z) = 0 \tag{11}$$

of polynomial equations. The number d of solutions to a generic system of this kind is given by the Bezout number $d = d_1 \cdots d_n$, where d_i is the total degree of P_i for each i. For suitable scaling parameters $\lambda_1, \ldots, \lambda_n, \sigma_1, \ldots, \sigma_n \in \mathbb{C}$, we now define $H: \mathbb{C}^{n+1} \to \mathbb{C}$ by

$$\begin{aligned} H_1(z,t) &= (1-t)\, P_1(z,t) + t\, \lambda_1\, (z^{d_1} - \sigma_1^{d_1}) \\ &\vdots \\ H_n(z,t) &= (1-t)\, P_n(z,t) + t\, \lambda_n\, (z^{d_n} - \sigma_n^{d_n}). \end{aligned}$$

Let

$$K = \{0, \ldots, d_1 - 1\} \times \cdots \times \{0, \ldots, d_n - 1\}$$

For any $k \in K$, the point

$$z_1^k = (\sigma_1\, \mathrm{e}^{2\pi \mathrm{i} k_1/d_1}, \ldots, \sigma_n\, \mathrm{e}^{2\pi \mathrm{i} k_n/d_n})$$

clearly satisfies $H(z_1^k) = 0$, whereas any z_0^k with $H(z_0^k, 0) = 0$ satisfies $P(z_0^k) = 0$.

If the system (11) is zero dimensional and the values $\lambda_1, \ldots, \lambda_n, \sigma_1, \ldots, \sigma_n$ are complex and sufficiently random (we also say that the homotopy is in general position), then the system $H_1(z, t) = \cdots = H_n(z, t) = 0$ is also zero dimensional for every $t \in [0, 1]$. In what follows we will always assume that the homotopy has been chosen in such a way.

4.2. Solutions at infinity

One classical difficulty with homotopy methods for solving a polynomial system (11) is that many of the solution paths z_t^k may tend to infinity in the sense that $(z_t^k)_i \to \infty$ for some i and $t \to 0$. Computations which infinities can be avoided by rewriting the equations in projective coordinates. More precisely, setting $z^{\mathrm{pr}} = (z_0, \ldots, z_n)$, the projectivation $A^{\mathrm{pr}} \in \mathbb{C}[z^{\mathrm{pr}}]$ of a polynomial $A \in \mathbb{C}[z]$ is defined by

$$A^{\mathrm{pr}}(z_0, \ldots, z_n) = z_0^{\deg A}\, A(\tfrac{z_1}{z_0}, \ldots, \tfrac{z_n}{z_0}).$$

Applying this to the system (11), we obtain a new system

$$P_1^{\mathrm{pr}}(z^{\mathrm{pr}}) = \cdots = P_n^{\mathrm{pr}}(z^{\mathrm{pr}}) = 0 \tag{12}$$

of homogeneous equations in z^{pr}. For a random hyperplane

$$\alpha_0\, z_0 + \cdots + \alpha_n\, z_n = \beta, \tag{13}$$

the composite system (12–13) is again zero dimensional, but without solutions at infinity. It is easy to reconstruct solutions to (11) from solutions to (12–13) and *vice versa*.

4.3. Predictor corrector methods

Assume that we have a way to approximate the Jacobian J_H of H by $\tilde{J}_H: \tilde{\Omega} \to \tilde{\mathbb{C}}^{n \times (n+1)}$. For instance, if H is given by a dag, then a dag for J_H can be computed using forward differentiation, and $\tilde{J}_H$ just corresponds to the approximated evaluation of this dag.

Assume that we are given $y = \tilde{H}(z,t)$ and $\tilde{J}_H(z,t)$ at a certain point where $\tilde{H}(z,t) \approx 0$. We may write $\tilde{J}_H(z, t) = (U, \dot{y})$ as the horizontal join of two matrices $U \in \tilde{\mathbb{C}}^{n \times n}$ and $\dot{y} \in \tilde{\mathbb{C}}^{n \times 1}$. Given $t' = t + d_t$ close to t, we may find a $z^{(1)}$ for which $\tilde{H}(z^{(1)}, t') \approx 0$ using Euler-Newton's method

$$z^{(1)} = z - U^{-1}(y + \dot{y}\, (t' - t)).$$

The replacement $(z,t) \rightsquigarrow (z^{(1)},t')$ is called a *prediction* step. We may still apply the formula when $t'=t$, in which case z' is usually a better approximation than z to a genuine zero of H at t than z. In this situation, the replacement $(z,t) \rightsquigarrow (z^{(1)},t)$ is called a *correction* step.

From the computational point of view, the evaluation of the Jacobian $J_{\tilde{H}}(z,t)$ is usually about n times more expensive than the evaluation of the function $\tilde{H}(z,t)$ itself (except for large n and sparse $J_{\tilde{H}}$). Instead of reevaluating the Jacobian after the prediction step at $(z^{(1)},t')$, it may therefore be worth it to perform a few correction steps using the Jacobian at (z,t) instead:

$$\begin{aligned} z^{(2)} &= z^{(1)} - U^{-1}\,\tilde{H}(z^{(1)},t') \\ &\vdots \\ z^{(\kappa)} &= z^{(\kappa-1)} - U^{-1}\,\tilde{H}(z^{(\kappa-1)},t'). \end{aligned}$$

Since the convergence of $z^{(1)}, z^{(2)}, \ldots$ is only linear, the number κ is typically chosen quite small ($\kappa \leqslant 3$). One full prediction-correction cyclus now just consists of the replacement $(z,t) \rightsquigarrow (z',t') = (z^{(\kappa)},t')$.

From the complexity point of view, the evaluation of $\tilde{H}$ and $J_{\tilde{H}}$ is usually far more expensive than the cost $\mathcal{O}(n^3)$ of linear algebra at size n, at least for the examples we will be interested in here. Therefore, it will not be necessary to device the linear algebra algorithms with special care (for instance, we may simply compute the inverse U^{-1} once and for all, instead of using LU decompositions). On the other hand, we typically want to increase the step size $t'-t$ as much as possible, while trying to stay reasonably close to the true solution path.

4.4. Precision control

One obvious source of numeric errors is when the numeric precision being used is insufficient for producing sensible results. In [BSHW08], a strategy has been proposed for selecting a sufficient precision for homotopy methods to be numerically reliable. We will now propose an alternative method for finding such a precision, whose justification is based on a simpler argument.

Let p be the current working precision. Our method is based on the following idea: when evaluating $y = \tilde{H}(z,t)$, the actual precision q of the result is usually smaller than p and of the form $q = p - c$ for some fixed constant. We will call q the *effective precision* and we may expect the numeric evaluations to be reliable as long as p is picked sufficiently large such that $q \geqslant \tau_0$ remains above a certain threshold $\tau_0 > 0$ (e.g. $\tau_0 = 10$).

We still need a more precise definition of the effective precision or a simple way to compute it. Assuming that $\tilde{H}$ admits a ball lift, we may evaluate $\boldsymbol{y} = \tilde{H}(\boldsymbol{z},\boldsymbol{t})$ at the ball $(\boldsymbol{z},\boldsymbol{t}) = (z + \mathcal{B}(0), t + \mathcal{B}(0)) \in \mathcal{B}(\mathbb{C}_p, \mathbb{R}_p)^{n+1}$. Then

$$q_{\mathrm{rel}}(z,t) \;=\; \min_i \left\lfloor \log_2 \frac{|\mathrm{cen}(\boldsymbol{y}_i)|}{\mathrm{rad}(\boldsymbol{y}_i)} \right\rfloor$$

provides an estimate for the relative precision of $\boldsymbol{y}$. If $\tilde{H}(z,\,t) \approx 0$, then this precision is potentially quite low. In that case, we may also consider $q_{\mathrm{rel}}(z,\,t')$ at the next time $t' = t + d_t$. Instead of performing one extra ball evaluation, we may also use the following approximation of $q_{\mathrm{rel}}(z,t')$:

$$q^*_{\mathrm{rel}}(z,t') \;=\; \min_i \left\lfloor \log_2 \frac{|\mathrm{cen}(\boldsymbol{y}_i) + \tilde{H}_t(z,t)_i\, d_t|}{\mathrm{rad}(\boldsymbol{y}_i)} \right\rfloor.$$

We now take

$$q=q(z,t,t') \;=\; \min_i \left\lfloor \log_2 \frac{\max \{|\mathrm{cen}(\boldsymbol{y}_i)|, |\mathrm{cen}(\boldsymbol{y}_i)+\tilde{H}_t(z,t)_i\, d_t|\}}{\mathrm{rad}(\boldsymbol{y}_i)} \right\rfloor.$$

for the current effective precision at (z,t) and assuming a current step size d_t.

4.5. Step size control

Since purely numeric homotopy methods are usually being designed for speed, the main focus is not on being 100% fool proof. Nevertheless, it remains worth it to search for cheap ways in order to detect errors and adapt the stepsize so as to avoid potential errors.

Now assume that we perform one full prediction correction cyclus $(z,t) \rightsquigarrow (z',t')$. We first need a criterion for when to accept such a step. The main problem with the design of numeric criteria is there is no way to decide whether a numeric quantity is small or large; such checks can only be performed with respect to other quantities. Instead of checking whether we remain close to the genuine solution path, it is therefore more robust to check that the Jacobian $\tilde{J}_H$ does not change not change to quickly on the interval $[t,t']$.

More precisely, let $y=\tilde{H}(z,t)$, $(U,\dot{y})=\tilde{J}_H(z,t)$, $y'=\tilde{H}(z',t')$ and $(U',\dot{y}')=\tilde{J}_H(z',t')$. Then it is natural to only accept steps for which

$$\|U^{-1}U'-1\| \;\leqslant\; \tau_1, \tag{14}$$

for a fixed threshold $\tau_1<1$ (e.g. $\tau_1=\frac{1}{4}$). Here we may use any matrix norm $\|\cdot\|$, so it is most convenient to chose one which is easy to compute:

$$\|M\| \;=\; \sum_i \max_j |M_{i,j}|.$$

The condition (14) is not fully satisfactory yet, since it relies on the expensive computation of a Jacobian U'. This is acceptable if the step has a good chance of being accepted (since we will need the Jacobian anyway for the next step), but annoying if the step is to be rejected. Before checking (14), it is therefore wise to perform a few cheaper checks in order to increase the probability that (14) will hold indeed. In particular, if $\kappa\geqslant 2$, then we may verify that

$$\|U^{-1}(y'-y^{(1)})-(z'-z^{(1)})\| \;\leqslant\; \tau_2\,\|z'-z^{(1)}\| \tag{15}$$

for the max-norm on vectors, where $\tau_2\leqslant\tau_1$ (e.g. $\tau_2=\frac{1}{2}\,\tau_1$) and $y^{(1)}=\tilde{H}(z^{(1)},t')$. This simplified check is linked to (14) by remarking that $y'-y^{(1)}\approx U'\,(z'-z)$. The new check (15) should not be applied when z' and $z^{(1)}$ are too close for y' and $y^{(1)}$ to be computed with sufficient precision. More precisely, it should really be replaced by the check

$$\left\{ \begin{array}{l} \|U^{-1}(y'-y^{(1)})-(z'-z^{(1)})\|\leqslant\tau_2\,\|z'-z^{(1)}\| \vee \\ \|z'-z^{(1)}\|\leqslant 2^{-\tau_3 q(z,t,t')}\,\|z^{(1)}\|, \end{array} \right. \tag{16}$$

where τ_3 is slightly smaller than one (e.g. $\tau_3=\frac{3}{4}$) and $q(z,t,t')$ stands for the "effective working precision" from section 4.4.

In addition to the above checks, one might wish to ensure that y' is reasonably small after each step. Unfortunately, there is no satisfactory reference with respect which smallness can be checked, except for $y^{(1)}$, ..., $y^{(\kappa-1)}$. The best we can do therefore consists of checking whether $y^{(1)}, y^{(2)}, \ldots$ tend to 0 at some indicated rate:

$$\left\{ \begin{array}{l} \|y^{(i+1)}\|\leqslant\tau_4\,\|y^{(i)}\| \vee \\ \|z^{(i+1)}-z^{(i)}\|\leqslant 2^{-\tau_4 q(z,t,t')}\,\|z^{(i)}\|, \end{array} \right. \tag{17}$$

for all $i<\kappa$, where $\tau_4<1$ (e.g. $\tau_4=\frac{1}{2}$). Again, we need to insert a safety exemption for the case when the convergence is exceptionally good.

Once that we have a criterion on whether a step $(z, t) \rightsquigarrow (z', t')$ should be accepted, an algorithm for automatic stepsize control is easily implemented: assuming that we are walking from $t = 1$ to $t = 0$, we start by setting $d_t := -1$. Given t and d_t, we try a step $(z, t) \rightsquigarrow (z', t')$ until $t' := t + d_t$. If the step fails, then we set $d_t := \lambda_{\text{fail}}\, d_t$ with $\lambda_{\text{fail}} < 1$ (e.g. $\lambda_{\text{fail}} = \frac{1}{2}$), and retry for the smaller stepsize. Otherwise, we accept the step $t := t'$ and set $d_t := \lambda_{\text{ok}}\, d_t$ for the next step, where $\lambda_{\text{ok}} > 1$ (e.g. $\lambda_{\text{ok}} = \sqrt{2}$).

4.6. Near collisions

Another way to look at the numerical error problem is to investigate what can actually go wrong. Theoretically speaking, around each true solution path z_t, there exists a small tube T_t of variable polyradius r_t, where Newton's method converges to the true solution z_t. As long as our current approximation z at time t remains in this tube T_t, no errors will occur. Now the Newton iterations have a strong tendency of projecting back into the tubes, especially if we use the additional safeguard (17). Nevertheless, it might happen that we jump from one tube into another tube, whenever two solution paths come close together.

If we are considering a homotopy for solving a polynomial system $P_1 = \cdots = P_n$, then various solution paths will actually meet at $t = 0$ if the system admits multiple roots. Such multiple roots are an intrinsic difficulty and we will need dedicated "end game" strategies to ensure good numeric convergence in this case (see section 5 below).

For $t > 0$, and for suitably prepared functions H, the Lebesgue probability that two solutions paths meet at a point is zero. Nevertheless, we may have near collisions, which usually occur in pairs: the probability that more than two paths simultaneously pass close to a same point is extremely low.

So assume that we have a near collision of two solution paths. Then we have a true collision at (z_*, t_*) for some complex time t_* near the real axis. Locally around this collision point, the two paths are then given by

$$z_t^{\pm} \;=\; z_* \pm u \sqrt{t - t_*} + \mathcal{O}(t - t_*),$$

for some vector u. If we only know z_t^+ at a few points, then we may try to compute z_*, t_* and u, and also check whether the second path z_t^- indeed exists.

Now assume that we have approximated z_t^+ and derivative $\dot{z}_t^+ = \mathrm{d} z_t^+ / \mathrm{d} t$ at two times $t_1 > t_2$. Denote these approximations by $\tilde{z}_1 = \tilde{z}_1^+ \approx z_{t_1}^+$, $\tilde{\dot{z}}_1 \approx \dot{z}_{t_1}^+$, $\tilde{z}_2 = \tilde{z}_2^+ \approx z_{t_2}^+$ and $\tilde{\dot{z}}_2 \approx \dot{z}_{t_2}^+$. Then

$$\tilde{\dot{z}}_i \;\approx\; \frac{u}{2 \sqrt{t_i - t_*}}$$

for $i \in \{1, 2\}$, whence we may use the following approximations for z_*, t_* and u:

$$\begin{aligned} \tilde{t}_* &:= \frac{(\tilde{\dot{z}}_2)^2\, t_2 - (\tilde{\dot{z}}_1)^2\, t_1}{(\tilde{\dot{z}}_2)^2 - (\tilde{\dot{z}}_1)^2} \\ \tilde{u} &:= 2\, \tilde{\dot{z}}_2 \sqrt{t_2 - \tilde{t}_*} \\ \tilde{z}_* &:= \tilde{z}_2 - \tilde{u} \sqrt{t_2 - \tilde{t}_*}. \end{aligned}$$

We next perform several safety checks. First of all, we obtained $\tilde{t}_*$ as the division of two vectors; we may use the mean value of the componentwise divisions and check that the variance remain small. We next verify that $\tilde{z}_1 - \tilde{u} \sqrt{t_1 - \tilde{t}_*}$ and $\tilde{z}_*$ are reasonably close. We also verify that the Newton iteration starting at $\tilde{z}_2^- = \tilde{z}_* - u \sqrt{t_2 - \tilde{t}_*}$ converges to a solution close to $\tilde{z}_2^-$. We finally verify that the same thing holds for $\tilde{z}_{\bar{2}}^{\pm} = \tilde{z}_* \pm u \sqrt{t_{\bar{2}} - \tilde{t}_*}$ instead of $\tilde{z}_2^{\pm}$, where $t_{\bar{2}} = \operatorname{Re}(2\, \tilde{t}_* - t_2)$.

We will not go into technical details on the precise numerical checks here, since section 5.3 below contains a similar discussion for the case of multiple roots at $t=0$. We may also adapt the herd iteration from section 5.2 below to near collisions, which allows for the simultaneous continuation of z_t^+ and z_t^-. Contrary to the case when $t\to 0$, we also need to recompute better estimations of t_* at every step, which can be done *via* the simultaneous computation of $z_t^{\pm}$ and the two "conjugate" paths $z_{\bar{t}}^{\pm}$ with $\bar{t}=\mathrm{Re}\,(2\,\tilde{t}_*-t)$. Indeed, using the higher order expansion

$$z_t^{\pm} \;=\; z_* \pm u\,\sqrt{t-t_*}+v\,(t-t_*)+w\,(t-t_*)^{3/2}+\mathcal{O}((t-t_*)^2),$$

we get

$$\begin{aligned}
z_t^+ + z_t^- &= 2\,z_* + 2\,v\,(t-t^*)+\mathcal{O}((t-t^*)^2)\\
z_{\bar{t}}^+ + z_{\bar{t}}^- &= 2\,z_* + 2\,v\,(\bar{t}-t^*)+\mathcal{O}((\bar{t}-t^*)^2)\\
\dot{z}_t^+ + \dot{z}_t^- &= 2\,v+\mathcal{O}(t-t^*),
\end{aligned}$$

from which we may deduce high quality approximations of t^* and z^*. As soon as $\bar{t}-t$ is small with respect to $\mathrm{Im}\,t_*$, then the junction between paths and their conjugates occurs and we know how to traverse the near collision.

5. Multiple roots

5.1. Straightforward Euler-Newton type methods

Consider a homotopy induced by a polynomial system (11) with a zero dimensional set of solutions. It frequently occurs that some of the solutions are multiple roots, in which case the predictor corrector algorithm slows down significantly when t approaches 0. This is due to the fact that Newton's method only has a linear convergence if we are approaching a multiple root, whereas the convergence is quadratic for single roots.

In order to get a better understanding of this phenomenon, it is instructive to quantify the slow down in the case of an r-fold root of a univariate polynomial P, which is more or less representative for the general case. In the neighbourhood of the root α, we have

$$P_{+\alpha}(z) := P(\alpha+z) \;=\; c\,z^r+\mathcal{O}(z^{r+1}),$$

with $c=\frac{1}{r!}\,P^{(r)}(\alpha)$. Hence, the Newton iteration becomes

$$\begin{aligned}
z' &= z-\frac{P_{+\alpha}(z)}{P_{+\alpha}'(z)}\\
&= \left(1-\tfrac{1}{r}\right)z+\mathcal{O}(z^2).
\end{aligned}$$

In particular, we see that we need roughly r iterations in order to divide z by e. We also notice that $P(\alpha+z)$ is roughly divided by e at every iteration. For complexity measures, it is more reasonable to study the speed of convergence of $P(\alpha+z)$ rather than z itself. Indeed, the relative precision of an r-fold root is intrinsically r times smaller than the working precision.

If we are rather considering a homotopy $H(z,t)=(1-t)\,P(z)+t\,Q(z)$, then we usually have $q=Q(\alpha)\neq 0$. Locally, we may thus write

$$H(\alpha+z,t) \;=\; c\,z^r+q\,t+\mathcal{O}(z^{r+1})+\mathcal{O}(z\,t).$$

Assume that we have $H(\alpha+z,t)=0$ for small z and $t>0$, so that

$$z^r \;=\; -\tfrac{q}{c}\,t+\mathcal{O}(z^{r+1}).$$

Then the Euler-Newton iteration for step size d_t yields

$$\begin{aligned} z' &= z - \frac{q\,d_t}{r\,c\,z^{r-1}} \\ &= \left(1 - \frac{d_t}{r\,t}\right) z + \mathcal{O}(z^2). \end{aligned}$$

Following our criterion (14), we should have

$$\left|\left(1 - \frac{d_t}{r\,t}\right)^{r-1} - 1\right| \;\leqslant\; \tau_1.$$

Roughly speaking, this means that $d_t \leqslant \tau_1\,t$. Hence, t is multiplied by $1 - \tau_1$ at every step and z is multiplied by $1 - \tau_1$ every r steps.

5.2. The herd iteration

For high precision computations, it would be nice to have an algorithm with quadratic convergence in t. Before we give such an algorithm, let us first introduce some terminology and study the behaviour of the solutions paths when $t \to 0$.

By assumption, we are given a system (11) with an r-fold root $\alpha \in \Omega$. Consider a solution path z_t for the homotopy with $\lim_{t\to 0} z_t = \alpha$. Since z_t is algebraic in t, we may expand

$$z_t \;=\; \alpha + c_1\,t^{1/p} + c_2\,t^{2/p} + \cdots,$$

as a Puiseux series in t for a certain ramification index p (which we assume to be taken minimal). Now letting t turn around 0 once, we have

$$z_{\mathrm{e}^{2\pi\mathrm{i}}t} \;=\; \alpha + c_1\,\omega\,t^{1/p} + c_2\,\omega^2\,t^{2/p} + \cdots,$$

where $\omega = \mathrm{e}^{2\pi\mathrm{i}/p}$. When turning repeatedly, we thus obtain p pairwise distinct solutions paths $z_t^k := z_{\mathrm{e}^{2\pi\mathrm{i}k}t}$ with $k \in \{0, \ldots, p-1\}$. We will call such a family of solution paths a *herd*.

Contrary to the homotopy methods from section 4, which operate on individual paths, the iteration that we will present now simultaneously operates on all paths in a herd. Consider a solution path z_t with $\lim_{t\to 0}\, z_t = \alpha$ as above and the corresponding herd $z_t^k = z_{\mathrm{e}^{2\pi\mathrm{i}k}t}$ with $k \in \{0, \ldots, p-1\}$. We assume that both $\tilde{z}_t^k \approx z^k$ and $\dot{\tilde{z}}_t^k \approx \dot{z}_t^k$ are known for a given $t > 0$ and all $k \in \{0, \ldots, p-1\}$. Let $(F_0, \ldots, F_{p-1})$ and $(\dot{F}_0, \ldots, \dot{F}_{p-1})$ denote the FFT-transforms of the vectors $(\tilde{z}_t^0, \ldots, \tilde{z}_t^{p-1})$ and $(\dot{\tilde{z}}_t^0, \ldots, \dot{\tilde{z}}_t^{p-1})$ with respect to ω^{-1}. Then we have

$$\begin{aligned} F_k &= \sum_{i=0}^{p-1} \tilde{z}_t^i\,\omega^{-ik} \\ &= p\,t^{k/p}\,(c_k + c_{k+p}\,t + \mathcal{O}(t^2)) \\ t\,\dot{F}_k &= t^{k/p}\,(k\,c_k + (k+p)\,c_{k+p}\,t + \mathcal{O}(t^2)). \end{aligned}$$

for all k. We now compute $\tilde{c}_0, \ldots, \tilde{c}_{2p-1}$ using the formulas

$$\begin{aligned} \tilde{c}_{k+p} &:= \tfrac{1}{p\,t^{1+k/p}}\left(t\,\dot{F}_k - \tfrac{k}{p}\,F_k\right) \\ &= c_{k+p} + \mathcal{O}(t) \\ \tilde{c}_k &:= \tfrac{1}{p\,t^{k/p}}\,F_k - \tilde{c}_{k+p}\,t \\ &= c_k + \mathcal{O}(t^2). \end{aligned}$$

For $t' > 0$ of the order of t^2, we now have

$$\begin{aligned} z_{t'}^k &= \tilde{z}_{t'}^k + \mathcal{O}(t^2) \\ \tilde{z}_{t'}^k &:= \tilde{c}_0 + \tilde{c}_1\,\omega^k\,(t')^{1/p} + \cdots + \tilde{c}_{2p-1}\,\omega^{(2p-1)k}\,(t')^{(2p-1)/p}, \end{aligned} \tag{18}$$

for all $k \in \{0, \ldots, p-1\}$. We call (18) the *herd prediction*. This prediction may be corrected using κ conventional Newton iterations at time t', for a fixed constant $\kappa \in \mathbb{N} \setminus \{0\}$. A complete cyclus of this type will be called a *herd iteration*.

5.3. Step size control for herd iterations

Several technical details need to be settled in order to obtain a robust implementation of herd iterations. First of all, we need a numeric criterion for deciding when the approximations $\tilde{z}_t^k \approx z_t^k$ and $\tilde{\dot{z}}_t^k \approx \dot{z}_t^k$ are of a sufficient quality for starting our herd iteration. Clearly, the error of the approximation should be in $\mathcal{O}(t^2)$.

We may first ensure ourselves that the approximation can not substantially be improved using Newton iterations: let $(\tilde{z}_t^k)'$ be the result of applying one Newton iteration to $\tilde{z}_t^k$ at time t. Then we check whether

$$\operatorname{relerr}(\tilde{z}_t^k, (\tilde{z}_t^k)') \; := \; \frac{|\tilde{z}_t^k - (\tilde{z}_t^k)'|}{|\tilde{z}_t^k|} \leqslant \tau_5 \, t^2, \tag{19}$$

for some threshold τ_5, such as $\tau_5 = \frac{1}{2}$ (although this check becomes unstable if $\tilde{z}_t^k \approx 0$, we notice that this situation cannot arise systematically for $t \rightarrow 0$).

The check (19) for $k \in \{0, \ldots, p-1\}$ does not yet guarantee that the $\tilde{z}_t^k$ correspond to approximate evaluations of the Puiseux expansions. In order to check that this is indeed the case, we first compute the $\tilde{c}_k$ as described in the previous section. Defining

$$\tilde{c}(t) \; = \; \tilde{c}_0 + \tilde{c}_1 \, t^{1/p} + \cdots + \tilde{c}_{2p-1} \, t^{(2p-1)/p},$$

we next evaluate $\tilde{z}_t^{k+1/2} = \tilde{c}(\mathrm{e}^{2\pi \mathrm{i} k + \pi \mathrm{i}} \, t)$ for all $k \in \{0, \ldots, p\}$ and apply one Newton iteration at time t to the results, yielding $(\tilde{z}_t^{k+1/2})'$. We now check whether

$$\operatorname{relerr}(\tilde{z}_t^{k+1/2}, (\tilde{z}_t^{k+1/2})') \; \leqslant \; \tau_6 \, t^2, \tag{20}$$

for some threshold τ_6, such as $\tau_6 = 1$, and all k. Of course, this second check is more expensive than the first check (19). The thresholds should therefore be adjusted in such a way that the second check is likely to succeed whenever the first one does.

The above criteria can also be used for deciding whether a proposed herd iteration from t to t' should be accepted or not. We still have to decide how to chose t'. For a fixed constant $\gamma > 1$ and a positive integer s which may change at every step, we will take

$$t' \; = \; 2^{-\gamma^s} \, t.$$

If a step is accepted, then we increase s by one or a larger integer smaller than $1/\log_2 \gamma$. If a step is not accepted, then we decrease s by one and repeat the same procedure until acceptance or $s = 0$. If $s = 0$, then we have either reached the best possible accuracy for the current working precision, or our p paths did not really converge to the same point α. The first case occurs whenever the effective precision from section 4.4 drops below a given threshold. In the latter case, we revert to individual homotopies for further continuation.

5.4. Detection of clusters

Let us now go back to the initial polynomial system (11) and assume that we have computed numerical approximations of all $d = d_1 \cdots d_n$ individual homotopies $(z_t^k)_{k \in K}$ up till a certain time $t > 0$. We need a way to partition the individual paths into herds. One obvious way is to follow all solution paths from t to $\mathrm{e}^{2\pi \mathrm{i}} \, t$ and deduce the corresponding permutation of K. However, this computation is quite expensive, so it would be nice to have something faster.

A first step towards the detection of herds is to find all clusters, i.e. all groups of paths which tend to the same limit α. Here we notice that one cluster may contain several herds, as in the example

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

where all four solution paths $(x_t, y_t) = (\epsilon_x \sqrt{t}, \epsilon_y \sqrt{t})$ with $\epsilon_x, \epsilon_y \in \{-1, 1\}$ tend to the quadruple root $(0,0)$ of $x^2 = y^2 = 0$. This cluster contains two herds $(x_t, y_t) = (\pm\sqrt{t}, \pm\sqrt{t})$ and $(x_t, y_t) = (\pm\sqrt{t}, \mp\sqrt{t})$.

Now let $\tilde{z}_t^k \approx z_t^k$ and $\tilde{\dot{z}}_t^k \approx \dot{z}_t^k$ for all $k \in K$. For each $k \in K$, we consider the ball

$$\boldsymbol{z}_t^k = \tilde{z}_t^k + \mathcal{B}(2\,t\,\tilde{\dot{z}}_t^k).$$

The radii of these balls has been chosen with care, such that, with high probability, any two paths which belong to the same herd are also in the same connected component of $\mathcal{Z} := \bigcup_{k \in K} \boldsymbol{z}_t^k$. This is best verified on the case of path $z_t = \alpha + c\,t^{1/p} + \cdots$. Then the next path in the cluster is $z_{\mathrm{e}^{2\pi\mathrm{i}}t} = \alpha + c\,\omega\,t^{1/p} + \cdots$ and

$$\begin{aligned} \tfrac{1}{2}\,|z_{\mathrm{e}^{2\pi\mathrm{i}}t} - z_t| &\approx \tfrac{c}{2}\,|\omega - 1|\,t^{1/p} \\ &\leqslant \frac{2\,c}{p}\,t^{1/p} \\ &\approx 2\,t\,\dot{z}_t. \end{aligned}$$

An efficient way to separate different connected components of $\mathcal{Z}$ is via projection. Let $\lambda \in \mathbb{R}^{2n}$ be a random vector of real numbers of length $\|\lambda\| = 1$. Then any point $z \in \mathbb{C}^n$ may be projected to the vector product $\pi_\lambda(z) = \lambda \cdot (\operatorname{Re} z, \operatorname{Im} z) \in \mathbb{R}$. Applying this projection to our balls $\boldsymbol{z}_t^k$, we obtain intervals $\boldsymbol{x}^k$. We may sort the $\boldsymbol{x}^k$ (and the corresponding $\boldsymbol{z}_t^k$) on their centers in time $\mathcal{O}(d \log d)$ and compute the various connected components of $\mathcal{X} := \bigcup_{k \in K} \boldsymbol{x}^k$ using a linear pass. Whenever $\boldsymbol{x}^k$ and $\boldsymbol{x}^l$ are in different connected components, then so are $\boldsymbol{z}_t^k$ and $\boldsymbol{z}_t^l$. Assuming that t is sufficiently small, application of this procedure for $2\,n$ random vectors λ results with probability one in the separation of all connected components corresponding to different clusters.

5.5. Detection of herds

Let $K' \subseteq K$ be a set of indices such that the z^k with $k \in K'$ form a cluster with limit α. We still need a way to find the various herds inside the cluster. In a similar way as in section 5.3, we may improve the quality of our approximations $\tilde{z}^k$ and $\tilde{\dot{z}}^k$ *via* Newton iteration until $\tilde{z}_t^k = z_t^t + \mathcal{O}(t^2)$ and $\tilde{\dot{z}}_t^k = \dot{z}_t^k + \mathcal{O}(t)$. From now on, we assume that we have done this.

For each $k \in K'$ and $i \in \{1, \ldots, n\}$, we may write

$$(z_t^k)_i = \alpha_i + c_i^k\,t^{\beta_i^k} + \cdots,$$

for some $c_i^k \in \mathbb{C} \setminus \{0\}$ and $\beta_i^k \in \mathbb{Q}^>$. We obtain a good approximation $A \approx \alpha + \mathcal{O}(t)$ using

$$\tilde{\alpha} = \frac{1}{|K'|} \sum_{k \in K'} \tilde{z}_t^k. \tag{21}$$

If $|K'|$ is not too large (so that β_i^k has a small numerator and denominator), then we also obtain reasonably accurate approximations $\tilde{\beta}_i^k \approx \beta_i^k$ and $\tilde{c}_i^k \approx c_i^k$ by

$$\begin{aligned} \tilde{\beta}_i^k &= \frac{t\,(\tilde{\dot{z}}_t^k)_i}{(\tilde{z}_t^k)_i - \alpha_i} \\ \tilde{c}_i^k &= (\tilde{z}^k - \alpha)\,t^{-\tilde{\beta}^k}. \end{aligned}$$

and check whether

$$z^k_{\mathrm{e}^{2\pi\mathrm{i}}t} \;\approx\; \tilde{\alpha} + \tilde{c}^k\,\mathrm{e}^{2\pi\mathrm{i}\tilde{\beta}^k}\,t^{\tilde{\beta}^k}$$

is indeed close to some $\tilde{z}^{k'}$ with $k' \in K'$. Doing this for all $k \in K'$, we thus obtain a candidate permutation $\sigma: K' \to K'$ with $z^k_{\mathrm{e}^{2\pi\mathrm{i}}t} = z^{\sigma(k)}_t$ for all $k \in K'$. Each cycle in this permutation induces a candidate herd. Using the criteria from 5.3, we may next check whether the quality of the candidate herd is sufficient. If not, then we may always resort to the more expensive computation of the solution path from t to $\mathrm{e}^{2\pi\mathrm{i}}\,t$.

5.6. Synchronization

Our algorithms for the previous sections for cluster and herd detection rely on the availability of approximations $\tilde{z}^k_t \approx z^k_t$ on all paths at the *same* time t. Usually the individual homotopies are launched in parallel and advance at different speeds. Consequently, the synchronization of all paths at the same time t is a non trivial matter.

Strictly speaking, we notice that it is not necessary to synchronize all paths, but rather those paths which belong to the same cluster or herd. In particular, we will concentrate on those paths which tend to multiple roots.

So consider a path z^k_t which tends to a multiple root α. As long as z^k_t is approximated using an individual continuation, we have seen that the convergence to $t \to 0$ is linear. For a fixed $\gamma < 1$ (such as $\gamma = \frac{1}{2}$), the computation of z^k_t at all "checkpoints" $t = \gamma, \gamma^2, \gamma^3, \ldots$ thus only requires a constant overhead. At every checkpoint, we may now launch the algorithm for the detection of clusters. For every candidate cluster K', we next determine the checkpoint γ^i with highest i at which $z^k_{\gamma^i}$ is available for all $k \in K'$. We launch our algorithm for the detection of herds at this checkpoint $t = \gamma^i$.

In addition, it is a good practice to check that we still have points on all $d = d_1 \cdots d_n$ paths at every checkpoint. For paths z^k_t which tend to a single root, we may approximate $z^k_{\gamma^i}$ for large i using a single step continuation from $t = 0$ to $t = \gamma^i$. For the approximation of α using (21), we notice that it important that no paths of the cluster are missing or counted twice. Indeed, in the contrary case, we only have $A = \alpha + \mathcal{O}(t^\beta)$ with $\beta_i = \min_{k \in K'} \beta^k_i$ for all i, which is insufficient for the computation of accurate approximations of β^k_i and c^k_i.

6. CERTIFIED HOMOTOPIES

6.1. Certification of Newton's method

Consider an analytic function $f: \Omega \to \mathbb{C}^n$ on some open subset Ω of $\mathbb{C}^n$ and assume that f admits a ball lift. Given an isolated root z of f, it is well known that Newton's method converges to z in a small neighbourhood of z. It is a natural question to explicitly compute a ball neighbourhood for which this is the case (where we notice that a "ball" in $\mathbb{C}^n$ is really a compact polydisk). One method which is both efficient and quite tight was proposed by Krawczyk [Kra69]. Recall that J_f denotes the Jacobian of f.

THEOREM 4. *Let $\boldsymbol{u} \in \mathbb{C}^n$, $u = \operatorname{cen} \boldsymbol{u}$ and let g: $\boldsymbol{u} \to \mathbb{C}^n$ be an analytic function. Let $J_g(\boldsymbol{u}) \in \mathbb{C}^{n\times n}$ be a ball enclosure of the set* $\operatorname{im} J_g$*. If*

$$g(u) - J_g(\boldsymbol{u})\,\mathcal{B}(\operatorname{rad}\boldsymbol{u}) \;\subseteq\; \boldsymbol{u},$$

then g admits a fixed point $z \in \boldsymbol{u}$.

Proof. For any $z \in \boldsymbol{u}$, we have

$$g(z) \;=\; g(u) + \int_0^1 J_g(u + (z - u)\, t)\, (z - u)\, \mathrm{d}t.$$

Since $J_g(\boldsymbol{u})$ is convex, we also have

$$\int_0^1 J_g(u + (z - u)\, t)\, \mathrm{d}t \;\in\; J_g(\boldsymbol{u}).$$

Hence

$$\begin{aligned} g(z) &\in g(u) + J_g(\boldsymbol{u})\, (\boldsymbol{u} - u) \\ &\subseteq \boldsymbol{u}. \end{aligned}$$

It follows that g is an analytic function from the compact ball $\boldsymbol{u}$ into itself. By Brouwer's fixed point theorem, we conclude that there exists a $z \in \boldsymbol{u}$ with $g(z) = z$. □

Corollary 5. *Let $\boldsymbol{u} \in \mathbb{C}^n$, $u = \operatorname{cen} \boldsymbol{u}$ and let $V \in \mathbb{C}^{n \times n}$ be an invertible matrix with $V J_f(\operatorname{cen} \boldsymbol{u}) \approx 1$. If $\Omega \supseteq \boldsymbol{u}$ and*

$$u - V f(u) + (1 - V J_f(\boldsymbol{u}))\, \mathcal{B}(\operatorname{rad} \boldsymbol{u}) \;\subseteq\; \boldsymbol{u},$$

then the equation $f(z) = 0$ admits a root $z \in \boldsymbol{u}$.

Proof. We apply the theorem for $g(z) = z - V f(z)$. □

The above method is still a bit unsatisfactory in the sense that it does not guarantee the uniqueness of the solution. Denoting by $\operatorname{int}(X)$ the interior of a subset X of $\mathbb{R}^n$, the following sharpening of the method is due to Rump [Rum80].

Theorem 6. *With the notations from theorem 4, if*

$$g(u) - J_g(\boldsymbol{u})\, \mathcal{B}(\operatorname{rad} \boldsymbol{u}) \;\subseteq\; \operatorname{int}(\boldsymbol{u}),$$

then g admits a unique fixed point in $\boldsymbol{u}$.

Proof. Let us first show that the spectral norm (i.e. the norm of the largest eigenvalue) of any $M \in J_g(\boldsymbol{u})$ is <1. Indeed, our assumption implies

$$\operatorname{rad}(J_g(\boldsymbol{u})\, \mathcal{B}(\operatorname{rad} \boldsymbol{u})) \;<\; \operatorname{rad} \boldsymbol{u}.$$

Now consider the norm $\|v\| = \max(|v_1| / \operatorname{rad} \boldsymbol{u}_1, \ldots, |v_n| / \operatorname{rad} \boldsymbol{u}_n)$ on $\mathbb{C}^n$. Then, for any $M \in J_g(\boldsymbol{u})$ and v with $\|v\| = 1$, we have

$$\begin{aligned} |M v| &\leqslant \operatorname{rad}(M \mathcal{B}(|v|)) \\ &\leqslant \operatorname{rad}(M \mathcal{B}(\operatorname{rad} \boldsymbol{u})) \\ &\leqslant \operatorname{rad}(J_g(\boldsymbol{u})\, \mathcal{B}(\operatorname{rad} \boldsymbol{u})) \\ &< \operatorname{rad} \boldsymbol{u}, \end{aligned}$$

whence $\|M v\| < 1$. This is only possible if the spectral norm of M is <1.

Now consider $\varphi(z) = z - g(z)$. By what precedes, any matrix M in $J_\varphi(\boldsymbol{u}) = 1 - J_g(\boldsymbol{u})$ is invertible. For any two distinct points $z, z' \in \boldsymbol{u}$, we have

$$\varphi(z') - \varphi(z) \;=\; \int_0^1 J_\varphi(z + (z' - z)\, t)\, (z' - z)\, \mathrm{d}t.$$

Since $J_\varphi(\boldsymbol{u})$ is convex, there exists a matrix $M \in J_\varphi(\boldsymbol{u})$ with

$$M = \int_0^1 J_\varphi(z + (z' - z)\, t)\, \mathrm{d}t.$$

By what precedes, it follows that $\varphi(z') - \varphi(z) = M\,(z' - z) \neq 0$. We conclude that $g(z) \neq z$ or $g(z') \neq z'$. The existence of a fixed point follows from theorem 4. □

Corollary 7. *With the notations of corollary 5, if*

$$u - V f(u) + (1 - V J_f(\boldsymbol{u}))\, \mathcal{B}(\operatorname{rad} \boldsymbol{u}) \subseteq \operatorname{int}(\boldsymbol{u}),$$

then the equation $f(z) = 0$ admits a unique root $z \in \boldsymbol{u}$.

Proof. Application of theorem 6 for $g(z) = z - V f(z)$. □

Assuming that we have computed a numeric approximation $\tilde{z}$ to a root z of f, a second question is how to find a suitable ball $\boldsymbol{z} \ni \tilde{z}$ for which the corollaries apply. Starting with $\boldsymbol{z}_0 := \tilde{z} + \mathcal{B}(0)$, a simple solution is consider the sequence defined by

$$\begin{aligned} \boldsymbol{z}_{i+1} &= \operatorname{cen} \boldsymbol{z}_i + \mathcal{B}(\max(\operatorname{rad} \boldsymbol{z}_i, \operatorname{rad}(K(\boldsymbol{z}_i) - \operatorname{cen} \boldsymbol{z}_i))) \qquad (22)\\ &\supseteq \boldsymbol{z}_i \cup K(\boldsymbol{z}_i), \end{aligned}$$

where

$$K(\boldsymbol{u}) = \operatorname{cen} \boldsymbol{u} - V f(\operatorname{cen} \boldsymbol{u}) + (1 - V J_f(\boldsymbol{u}))\, \mathcal{B}(\operatorname{rad} \boldsymbol{u})$$

Whenever $K(\boldsymbol{z}_i) \subseteq \operatorname{int}(\boldsymbol{z}_i)$, then we are done. In order to ensure the convergence of this method, we need to tweak the recurrence (22) and replace it by

$$\boldsymbol{z}_{i+1} = \operatorname{cen} \boldsymbol{z}_i + \mathcal{B}((1 + \varepsilon) \max(\operatorname{rad} \boldsymbol{z}_i, \operatorname{rad}(K(\boldsymbol{z}_i) - \operatorname{cen} \boldsymbol{z}_i)) + \eta), \qquad (23)$$

for suitable small positive constants ε and η. We refer to [Rum80] for more details on this technique, which is called ε-inflation.

6.2. Certification of a numeric homotopy continuation

Assume that the polynomial system (11) admits only simple roots and that we have obtained numeric approximations $\tilde{z}^k = \tilde{z}_0^k$ for all these roots using a numeric path tracker. Then theorem 5 suffices for the joint certification of the numeric approximations $\{\tilde{z}^k\}_{k \in K}$. Indeed, using the above technique, we first compute balls $\boldsymbol{z}^k \ni \tilde{z}^k$ for which theorem 5 applies. To conclude, it then suffices to check that these balls are pairwise disjoint. This can be done using the same algorithm as for the detection of clusters, which was described in section 5.4.

In the case when two balls $\boldsymbol{z}^k$ and $\boldsymbol{z}^{k'}$ do intersect, then we recompute approximations for the paths z_t^k and $z_t^{k'}$ using a smaller step size, that is, by lowering the constant τ_1 in (14). We keep doing so until none of the balls $\boldsymbol{z}^k$ intersect; even if some of the *paths* z_t^k may have been permuted due to numerical errors, the final *set* of all $\boldsymbol{z}^k$ is correct if none of the balls intersect. Indeed, each of the balls contains a solution and there can be no more solutions than the number predicted by the Bezout bound.

If $\boldsymbol{z}^k$ and $\boldsymbol{z}^{k'}$ intersect then, instead of recomputing the paths z_t^k and $z_t^{k'}$ using smaller and smaller step sizes, we may also search for a way to certify the entire homotopy computations. This will be the topic of the remainder of this section. Let us first show how to adapt the theory from the previous section to certified path tracking. From now on, we assume that $H: \Omega \times \mathbb{C} \rightarrow \mathbb{C}^n$ is an analytic function which admits a ball lift.

Theorem 8. *Let $(\boldsymbol{u}, \boldsymbol{t}) \in \mathbb{C}^n \times \mathbb{C}$ be such that $\boldsymbol{u} \subseteq \Omega$. Let $J = (\partial H / \partial z)(\operatorname{cen} \boldsymbol{u}, \operatorname{cen} \boldsymbol{t})$ and let $V \in \mathbb{C}^{n \times n}$ be an invertible matrix with $VJ \approx 1$. If*

$$\operatorname{cen} \boldsymbol{u} - VH(\operatorname{cen} \boldsymbol{u}, \boldsymbol{t}) + \left(1 - V \frac{\partial H}{\partial z}(\boldsymbol{u}, \boldsymbol{t})\right) \mathcal{B}(\operatorname{rad} \boldsymbol{u}) \;\subseteq\; \operatorname{int}(\boldsymbol{u}),$$

then the equation $H(z, t) = 0$ admits a unique root $z \in \boldsymbol{u}$ for each $t \in \boldsymbol{t}$.

Proof. Let $t \in \boldsymbol{t}$ and consider the function $g: \boldsymbol{u} \rightarrow \mathbb{C}^n; z \mapsto z - H(z, t)$. Then $\boldsymbol{u} - VH(\boldsymbol{u}, \boldsymbol{t})$ encloses $\operatorname{im} g$ and $1 - V \frac{\partial H}{\partial z}(\boldsymbol{u}, \boldsymbol{t})$ encloses $\operatorname{im} J_g$, and we conclude by theorem 6. □

Clearly, for any $t, t' \in \boldsymbol{t}$, theorem 8 ensures the existence of a unique solution path from t to t' in the tube $\boldsymbol{u} \times [t, t']$. As at the end of the previous section, the question again arises how to compute balls $\boldsymbol{u}$ and $\boldsymbol{t}$ for which the conditions of the theorem are likely to be satisfied. Since the computation of $\frac{\partial H}{\partial z}(\boldsymbol{u}, \boldsymbol{t})$ is expensive, it is important to keep down the number of iterations of the type (22) or (23) as much as possible (say at most one iteration).

Now assume that we performed a numeric homotopy computation from (z, t) to (z', t'). Then a reasonable first guess is to take

$$\begin{aligned}
\boldsymbol{u} &= {}^1\!/_2 (z + z') + \mathcal{B}(c (z' - z)) \\
\boldsymbol{t} &= {}^1\!/_2 (t + t') + \mathcal{B}({}^1\!/_2 (t' - t)),
\end{aligned}$$

for some $c > {}^1\!/_2$, say $c = 1$. Unfortunately, if one of the components of $z' - z$ tends to zero, then this guess turns out to be inadequate. Therefore, it is recommended to use an additional inflation proportional to the norm of $z' - z$:

$$\boldsymbol{u} = {}^1\!/_2 (z + z') + \mathcal{B}(c (z' - z) + c' \|z' - z\|),$$

for some small $c' > 0$, say $c' = {}^1\!/_{10}$. Another idea is to use the radius of the previous step as a reference (except for the very first step, of course). For instance, if our previous step went from $(\grave{z}, \grave{t})$ to (z, t), then we may take

$$\boldsymbol{u} = {}^1\!/_2 (z + z') + \mathcal{B}\left(c (z' - z) + c'' (z - \grave{z}) \frac{t' - t}{t - \backslash t}\right),$$

for some small $c'' > 0$, say $c'' = {}^1\!/_{10}$.

6.3. Certification *via* tubular models

One important disadvantage of the method from the previous section for the certification of one path tracking step is that we use global error bounds on the tube $\boldsymbol{u} \times \boldsymbol{t}$. Consequently, the inaccuracy $\operatorname{rad} \boldsymbol{u}$ of $\boldsymbol{u}$ is proportional to the step size $2 \operatorname{rad} \boldsymbol{t}$, whence any overestimation in the evaluation of H or J_H due to the inaccuracy in $\boldsymbol{u}$ requires a reduction of the step size.

For this reason, it is much better to follow the solution path as closely as possible instead of enclosing it in a "square tube". This can be achieved *via* the use of Taylor models. Using $\mathcal{D}$-stable Taylor models, it is possible to simultaneously compute of accurate enclosures for H and J_H on the tube.

More precisely, let $r_\epsilon \in (\mathbb{R}^>)^n$, $r_\delta \in \mathbb{R}^{\geqslant}$ and $D = \mathcal{B}(r_\epsilon) \times \mathcal{B}(r_\delta)$. For a fixed k in $\mathbb{N} \setminus \{0\}$, let $\mathcal{I} = \mathcal{J}$ be an initial segment of $\mathbb{N}^{n+1}$ of the form

$$\mathcal{I} = 0 \times \{0, \ldots, k\} \cup \{\mathbb{e}_1, \ldots, \mathbb{e}_n\} \times \{0\}$$

and let $\mathcal{D} = \mathcal{T}_1 = \{i \in \mathbb{N}^{n+1}: \|i\| \leqslant 1\}$. A $\mathcal{D}$-stable Taylor model in $\mathcal{B}_D(\mathbb{C}[\epsilon, \delta]_{\mathcal{I}}, \mathbb{R}[\epsilon, \delta]_{\mathcal{I}})$ will also be called a *tubular model*. We will write $\mathbb{T}_{D,\mathcal{I}}$ for the set of tubular models. Given $\boldsymbol{y} \in \mathbb{T}_{D,\mathcal{I}}^n$, we let $\boldsymbol{y}_{\text{cst}} \in \mathbb{C}^n$ and $\boldsymbol{y}_{\text{lin}} \in \mathbb{C}^{n \times n}$ be such that

$$\begin{aligned}
(\boldsymbol{y}_{\text{cst}})_i &= \varpi(\boldsymbol{y}_i) \\
(\boldsymbol{y}_{\text{lin}})_{i,j} &= \varpi\left(\frac{\partial \boldsymbol{y}_i}{\partial \epsilon_j}\right),
\end{aligned}$$

for all $i, j \in \{1, \ldots, n\}$.

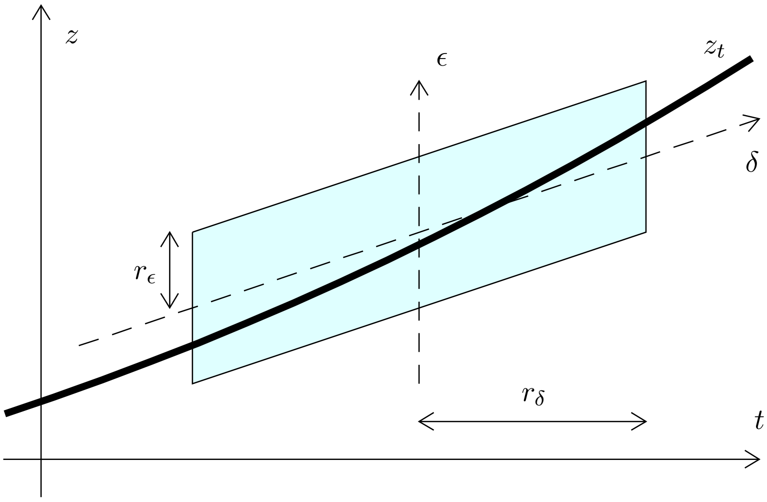


Figure 1. Illustration of a solution path z_t in a tube.

Theorem 9. *Let $c = (c_\epsilon, c_\delta) \in \Omega \times \mathbb{C}$, $r = (r_\epsilon, r_\delta) \in (\mathbb{R}^>)^n \times \mathbb{R}^>$, $D = \mathcal{B}(r)$ and let*

$$E(\delta) \;=\; E_0 + \cdots + E_{k_1} \delta^k$$

be an approximation of the solution ε_δ to $H(c_\epsilon + \varepsilon_\delta, c_\delta + \delta) = 0$. For instance, if $k = 1$ and $H(c) \approx 0$, then we may take $E(\delta) \approx -VF\delta$, with $V \approx (\partial H / \partial z)(c)^{-1}$ and $F \approx (\partial H / \partial t)(c)$. Consider $\boldsymbol{u} \in \mathbb{T}_{D,\mathcal{I}}^n$ and $\boldsymbol{v} \in \mathbb{T}_{D,\mathcal{I}}$ with

$$\begin{aligned} \boldsymbol{u}_i &= (c_\epsilon)_i + \epsilon_i + E(\delta)_i \\ \boldsymbol{v} &= c_\delta + \delta. \end{aligned}$$

Let $g(z, t) = z - VH(z, t)$, $\boldsymbol{x} = g(c_\epsilon + E(\delta), \boldsymbol{v})$, $\boldsymbol{y} = g(\boldsymbol{u}, \boldsymbol{v})$. If

$$\boldsymbol{x}_{\mathrm{cst}} + \boldsymbol{y}_{\mathrm{lin}} \mathcal{B}(r_\epsilon) \;\subseteq\; \mathrm{int}(c_\epsilon + \mathcal{B}(r_\epsilon)), \tag{24}$$

then the equation $H(z, t) = 0$ admits a unique solution $z \in c_\epsilon + E(t - c_\delta) + \mathcal{B}(r_\epsilon)$, for every $t \in c_\delta + \mathcal{B}(r_\delta)$.

Proof. For an illustration of the proof, see figure 1. Let $u = \mathrm{cen}(\boldsymbol{u}) \in \mathbb{C}[\epsilon, \delta]^n$ and $v = \mathrm{cen}(\boldsymbol{v}) \in \mathbb{C}[\delta]$. By construction, and using the facts that $\partial u / \partial \epsilon = 1$ and $\partial v / \partial \epsilon = \partial E / \partial \epsilon = 0$, we have

$$\begin{aligned} g(u(0, \delta), v(\delta)) &\in \boldsymbol{x}_{\mathrm{cst}} \\ \frac{\partial g}{\partial u}(u(\epsilon, \delta), v(\delta)) &\in \boldsymbol{y}_{\mathrm{lin}} \end{aligned}$$

for any $\epsilon \in \mathcal{B}(r_\epsilon)$ and $\delta \in \mathcal{B}(r_\delta)$. For a fixed $t \in c_\delta + \mathcal{B}(r_\delta)$, it follows that $\boldsymbol{y}_{\mathrm{lin}}$ encloses $(\partial g / \partial u)(\cdot, t)$ on the disk $\mathcal{U} := c_\epsilon + E(t - c_\delta) + \mathcal{B}(r_\epsilon)$. Our hypothesis (24) also implies that

$$g(c_\epsilon + E(t - c_\delta), t) + \boldsymbol{y}_{\mathrm{lin}} \mathcal{B}(r_\epsilon) \;\subseteq\; \mathrm{int}(\mathcal{U}).$$

From theorem 6, we conclude that $g(\cdot, t)$ admits a unique fixed point $z \in \mathcal{U}$. □

In order to apply the theorem, it remains to be shown how to find a good tube, i.e. how to choose c_ϵ, c_δ, r_ϵ, r_δ and $E(\delta)$. For a fixed order k of the approximation, the idea is to adjust c_ϵ and $E(\delta)$ such that r_ϵ can be chosen minimal.

Let us first consider the first order case $k = 1$. Assume that we performed a numeric path continuation from $(z_t,\ t)$ to $(z_{t'}, t')$ and that both $\dot{z}_t$ and $\dot{z}_{t'}$ are approximatively known. Then there exists a unique curve $\tilde{z}_s$ of degree three with $\tilde{z}_t = z_t$, $\tilde{z}_{t'} = z_{t'}$, $\dot{\tilde{z}}_t = \dot{z}_t$ and $\dot{\tilde{z}}_{t'} = \dot{z}_{t'}$. Let $\tilde{\tilde{z}}_s$ be a linear curve which minimizes the maximum μ_i of $|(\tilde{z}_s - \tilde{\tilde{z}}_s)_i|$ on $[t, t']$ for every i. Then we take $c_\delta = (t + t')/2$, $r_\delta = (t' - t)/2$, $c_\epsilon = \hat{z}_{c_t}$ and $E(\delta) = \dot{\tilde{\tilde{z}}}_{c_t}\,\delta$. We may also take $r_\epsilon = c\,\mu$ for some fixed $c > 1$ such as $c = 2$. However, for better performance it is recommended to apply an additional inflation to r_ϵ, similar to what we did in the previous section.

For higher orders k, we proceed in an essentially similar way. We first compute a high order numeric polynomial approximation $\tilde{z}_s$ of z_s. For orders >3, this may require the accurate approximation of additional points $(z_{t''},\ t'')$ with $t'' \in (t,\ t')$ on the solution path. We next find a k-th order polynomial $\hat{z}_s$ which approximates $\tilde{z}_s$ as good as possible and choose our tube in a similar way as above. It should be noticed that the evaluation $g(\boldsymbol{u}, \boldsymbol{v})$ in theorem 9 is at least thrice as expensive as the numeric evaluation of J_H. This makes it worth it to improve the quality of the numeric approximations of points $z_t, z_{t'}, z_{t''}$ on the curve using one or more additional Newton iterations. The use of higher order approximations makes it possible to choose r_ϵ very small, thereby avoiding a great deal of the overestimation due to the use of ball arithmetic.

6.4. Numeric spearhead computations and certification

We insist once more on the importance of performing all certifications as late as possible rather than along with the numeric computations themselves. One should regard the numeric computation (the spearhead) as an educated guess of what is happening and the certification as an independent problem at a second stage. In particular, only the numeric results which interests us (i.e. the solutions of the polynomial system) need to be certified and not the way we obtained them (i.e. the homotopies).

Even in the case when we are interested in certifying the homotopies themselves, it is best to do so once the numeric computations have already been completed. This allows for instance for more parallelism in the computations. Indeed, we may cut the entire homotopy path in several pieces and certify these pieces in parallel. More numeric data may also be available once all numeric computations have been completed, which might be useful for guiding and accelerating the certification stage.

Similar remarks apply more generally for certified integration of dynamical systems. In that setting, one is often interested in the flow in the neighbourhood of some initial condition. The sequential part of such a computation resides in the numeric integration for a particular initial condition. Once the corresponding numeric trajectory is known numerically, we may again cut it in pieces which and compute the corresponding flows and certifications in parallel. Whenever the dependence on the initial conditions is very strong, the actual numeric integration should be done using accurate high order Taylor methods using a multiple working precision.

7. Certification of multiple univariate roots

In section 5.1, we have studied in detail the numeric determination of a multiple root of a univariate polynomial. It is instructive to take up this study and examine how we certify such multiple roots. Since the property of being an r-fold root is lost under small perturbations, this is actually impossible using ball arithmetic. The best we can hope for is to certify the existence of r roots in a small ball, or the existence of an r-fold root of a small perturbation of the polynomial (see also [Rum10]).

7.1. Applying Rouché's theorem

So consider a polynomial P with an approximate r-fold root at $c \in \mathbb{C}$ and assume that we wish to certify that P admits exactly r roots in the ball $c + \mathcal{B}(\rho)$, for some $\rho > 0$. One first strategy is to make use of the Taylor series expansion of P at c. More precisely, let $\mathbb{T} = \mathbb{T}_{D,\mathcal{I},\mathcal{I}}$ be the set of univariate Taylor models in ϵ with $D = \mathcal{B}(\rho)$ and $\mathcal{I} = \{0, \ldots, s\}$ for some $s \geqslant r$. Evaluating P at $\boldsymbol{c} = c + \epsilon$, we obtain a Taylor model $\boldsymbol{Q} = P(\boldsymbol{c})$ with the property that $P(c + z) \in \boldsymbol{Q}_0 + \cdots + \boldsymbol{Q}_s z^s$ for any $z \in \mathcal{B}(\rho)$. It remains to be shown that any $Q \in \boldsymbol{Q}$ admits r roots in $\mathcal{B}(\rho)$. We claim that this is the case if

$$\lceil \boldsymbol{Q}_0 + \cdots + \boldsymbol{Q}_{r-1} \mathcal{B}(\rho)^{r-1} + \boldsymbol{Q}_{r+1} \mathcal{B}(\rho)^{r+1} + \cdots + \boldsymbol{Q}_s \mathcal{B}(\rho)^s \rceil \; < \; \lfloor \boldsymbol{Q}_r \rfloor \, \rho^n. \qquad (25)$$

Indeed, assume that we have (25) and let $Q \in \boldsymbol{Q}$. Then

$$|Q(z) - Q_r z^r| \; < \; |Q_r z^n|$$

for all z with $|z| = \rho$. By Rouché's theorem [Lan76, page 158], it follows that $Q(z)$ and $Q_r z^r$ admit the same number of roots on $\mathcal{B}(\rho)$. If r becomes large, or if P admits other roots close to $\mathcal{B}(\rho)$, then the bound (25) often does not hold. In that case, one may use more sophisticated techniques from [Sch82, Hoe11] in order to certify that $\boldsymbol{Q}$ admits r roots in $\mathcal{B}(\rho)$. From the complexity point of view, the series expansion method requires $\mathcal{O}(\mathsf{M}(r))$ evaluations of P, where $\mathsf{M}(r)$ denotes the cost of multiplying two polynomials of degrees $\leqslant r$.

7.2. Computing the winding number

Another approach is to apply Rouché's theorem in a more direct way by computing P on a path γ starting at ρ and which circles around the origin once. If the reliable image $P \circ \gamma$ of this path avoids the origin, then the number of roots of P coincides with the number of times that $P \circ \gamma$ turns around the origin. More precisely, let $\omega = \mathrm{e}^{2\pi\mathrm{i}/R}$ for a suitable $R > r$ (see also below) and let $\boldsymbol{z}_i = \omega^i + \mathcal{B}(|\sqrt{\omega} - 1|)$ for $i \in \{0, \ldots, R-1\}$. Then we evaluate $\boldsymbol{y}_i = P(\boldsymbol{z}_i)$ and check whether $0 \notin \boldsymbol{y}_i$ for all i. If this is the case, then

$$r' \; = \; \frac{1}{2\pi} \sum_{i=0}^{R-1} \arg \frac{\operatorname{cen}(\boldsymbol{y}_{i+1 \bmod R})}{\operatorname{cen}(\boldsymbol{y}_i)}$$

yields the exact number of roots of P inside $\mathcal{B}(\rho)$. This method requires R evaluations of P, but R needs to be sufficiently large if we want to ensure a reasonable chance of success for the method.

Let us investigate the choice of an appropriate R in more detail on the simplest example when $c = 1$ and

$$P(z) \; = \; z^r - r z^{r-1} + \binom{r}{2} z^{r-2} + \cdots + (-1)^r.$$

Consider the evaluation of P at $\boldsymbol{z} = 1 + \rho + \mathcal{B}(\epsilon)$. We have

$$\begin{aligned} P(\boldsymbol{z}) &= P(1+\rho) + \mathcal{B}(\sigma) \\ &= \rho^r + \mathcal{B}(\sigma) \\ \sigma &= \sum_{k=0}^{r} \binom{r}{k} ((1+\rho+\epsilon)^k - (1+\rho)^k) \\ &= (2+\rho+\epsilon)^r - (2+\rho)^r \end{aligned}$$

For small ϵ, the condition $0 \notin P(\boldsymbol{z})$ thus implies

$$r\,(2+\rho)^{r-1}\epsilon \; \approx \; (2+\rho+\epsilon)^r - (2+\rho)^r \; < \; \rho^r.$$

Roughly speaking, for $\rho \rightarrow 0$, this means that

$$\begin{aligned}\epsilon &< \tfrac{1}{r}\left(\tfrac{\rho}{2}\right)^{r-1}\rho\\ R &> \frac{\rho}{\pi\epsilon} > r\left(\frac{2}{\rho}\right)^{r-1}.\end{aligned}$$

We recall from example 2 that $(\rho/2)^{r-1}$ also corresponds to the punctual overestimation of the ball evaluation of P at $1+\rho$. If we want to reduce R to a quantity which does not depend on ρ, then it follows from the considerations in section 3.5 that we need to evaluate P using Taylor models of order at least r. However, in that case, we might just as well use the first method based on a direct series expansion of P at c.

7.3. Certifying a local factorization

Whenever a polynomial $P(z)$ admits r roots $\alpha_1, \ldots, \alpha_r$ in a disk, then the polynomial $Q = (z-\alpha_1)\cdots(z-\alpha_r) = z^r + Q_{r-1} z^{r-1} + \cdots + Q_0$ is a monic factor of P. Instead of trying to determine and certify the roots $\alpha_1, \ldots, \alpha_r$ individually, another idea is to determine and certify this monic factor Q of P.

Assuming that $P(z)$ admits no other roots near $\{\alpha_1, \ldots, \alpha_r\}$, we will show in this section that Q forms an isolated zero of $P \operatorname{rem} Q = 0$, when regarding this equation as a system of r equations $(P \operatorname{rem} Q)_0 = \cdots = (P \operatorname{rem} Q)_{r-1} = 0$ in r variables $Q_0, \ldots, Q_{r-1}$. On one hand, this approach has the advantage that we may apply Corollary 7 in order to certify the factorization. On the other hand, we may use fast polynomial arithmetic for the actual evaluation of $P \operatorname{rem} Q$.

Theorem 10. *Let $\alpha_1, \ldots, \alpha_\ell \in \mathbb{C}$ be roots of a polynomial P of multiplicities $\mu_1, \ldots, \mu_\ell \in \mathbb{N}^>$, and $r = \mu_1 + \cdots + \mu_\ell$. Then $Q = (z-\alpha_1)^{\mu_1} \cdots (z-\alpha_\ell)^{\mu_\ell}$ is an isolated zero of the system $P \operatorname{rem} Q = 0$, when considered as a system of r equations $(P \operatorname{rem} Q)_0 = \cdots = (P \operatorname{rem} Q)_{r-1} = 0$ in r variables $Q_0, \ldots, Q_{r-1}$.*

Proof. Given the euclidean division $P = A\,Q + B$ of P by Q, we have $A\,\mathrm{d}Q + Q\,\mathrm{d}A + \mathrm{d}B = \mathrm{d}P = 0$ for small perturbations $\mathrm{d}Q$ of Q, with $\deg(\mathrm{d}B) < r$. Consequently,

$$\mathrm{d}(P \operatorname{rem} Q) = -((P \operatorname{quo} Q)\,\mathrm{d}Q) \operatorname{rem} Q.$$

Computing in the quotient algebra $\mathbb{C}[x]/(Q)$, this equation can be reread as

$$\frac{\partial(P \operatorname{rem} Q)}{\partial Q} = -(P \operatorname{quo} Q) \operatorname{rem} Q. \tag{26}$$

Since $P = Q\,R$ with $\gcd(R, Q) = 1$, the multiplication mapping with $R = P$ quo Q in $\mathbb{C}[x]/(Q)$ is invertible. By (26), the Jacobian matrix of

$$(Q_0, \ldots, Q_{r-1}) \longmapsto ((P \operatorname{rem} Q)_0, \ldots, (P \operatorname{rem} Q)_{r-1})$$

is precisely the matrix of this multiplication mapping with respect to the canonical basis $(1, x, \ldots, x^{r-1})$. □

8. Effective complex analysis

For the sequel, it will sometimes be convenient to consider the more general context of analytic functions instead of mere polynomials. We provide two formalizations of computable multivariate analytic functions: an abstract one relying on analytic continuation in section 8.1, as well as a more concrete evaluation based one in section 8.2. In the remainder of the section, we will discuss evaluation of analytic functions in zero dimensional algebras, effective Weierstrass preparation, and computable meromorphic functions.

8.1. Computable multivariate analytic functions

We recall [Wei00] that a real number $x \in \mathbb{R}$ is said to be *left* (resp. *right*) *computable* if there exists an increasing (resp. decreasing) computable sequence $(x_n) \in \mathbb{Q}^>$ with $x = \lim_{n \to \infty} x_n$. We say that $x \in \mathbb{R}$ is *computable* if x is both left and right computable. We denote the sets of computable, left computable and right computable real numbers by $\mathbb{R}^{\text{com}}$, $\mathbb{R}^{\text{lcom}}$ and $\mathbb{R}^{\text{rcom}}$. We define $\mathbb{C}^{\text{com}} = \mathbb{R}^{\text{com}}[\mathrm{i}]$ to be the set of computable complex numbers. The definitions also adapt in a straightforward way to extended real numbers $x \in \bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. The theory of computable real numbers provides a suitable abstract framework for studying which analytic problems can be solved.

In [Hoe05, Hoe07], we proposed a similar concept of computable analytic functions. Given an analytic function f at the origin, we say that f is *computable* if there exists methods for computing the power series expansion of f, a lower bound for its convergence radius, an upper bound for f on any closed disk on which f converges, and a method for the analytic continuation of f. Formally speaking, denoting by $\mathbb{F}^{\text{com}}$ the set of such functions, this means that we may compute

- The computable power series expansion $\mathsf{series}(f) \in \mathbb{C}^{\text{com}}[[z]]^{\text{com}}$ of f (this means that we have an algorithm for the computation of the coefficients of $\mathsf{series}(f)$).
- A lower bound $\underline{r}_f \in \bar{\mathbb{R}}^{\text{lcom},>}$ for the radius of convergence r_f of f.
- A computable partial function $\lceil\!\lceil f \rceil\!\rceil_\cdot \colon \mathbb{R}^{\text{com},>} \rightharpoonup \mathbb{R}^{\text{rcom}}$, which yields an upper bound $\lceil\!\lceil f \rceil\!\rceil_\rho \geqslant \|f\|_\rho = \sup_{|z| \leqslant \rho} |f(z)|$ for every $\rho < \underline{r}_f$.
- A computable partial function $f_{+\cdot} \colon \mathbb{C}^{\text{com}} \rightharpoonup \mathbb{F}^{\text{com}}$, which yields the analytic continuation $f_{+\delta}$ of f (with $f_{+\delta}(z) = f(z+\delta)$) as a function of $\delta \in \mathbb{C}^{\text{com}}$ with $|\delta| < \underline{r}_f$.

Given $f \in \mathbb{F}^{\text{com}}$, we call $\underline{r}_f$ its *computable radius of convergence*. Usually, $\underline{r}_f$ is smaller than the genuine radius of convergence of $\mathsf{series}(f)$.

This definition admits several variants. In practice, it is usually most convenient to provide a method for the computation of bounds $\lceil\!\lceil f \rceil\!\rceil_\rho$ using ball arithmetic and allow for infinite bounds. In that case, we automatically obtain an algorithm for the computation of $\underline{r}_f$, by taking $\underline{r}_f := \max\{\rho \in \mathbb{R}^{\text{com},>} \colon \lceil\!\lceil f \rceil\!\rceil_\rho < \infty\}$. This definition is also convenient to extend to the case of multivariate analytic functions f in $z_1, \ldots, z_n$. In this case, we require algorithms for the computation of:

- The computable power series expansion $\mathsf{series}(f) \in \mathbb{C}^{\text{com}}[[z_1, \ldots, z_n]]^{\text{com}}$ of f.
- A computable partial function $\lceil\!\lceil f \rceil\!\rceil_\cdot \colon (\mathbb{R}^{\text{com},>})^n \rightharpoonup \bar{\mathbb{R}}^{\text{rcom}}$, which yields a possibly infinite upper bound $\lceil\!\lceil f \rceil\!\rceil_\rho \geqslant \|f\|_\rho = \sup_{|z| \leqslant \rho} |f(z)|$.
- A computable partial function $f_{+\cdot} \colon (\mathbb{C}^{\text{com}})^n \rightharpoonup \mathbb{F}^{\text{com}}$, which yields the analytic continuation $f_{+\delta}$ of f as a function of $\delta \in \mathbb{C}^{\text{com}}$ with $\lceil\!\lceil f \rceil\!\rceil_{|\delta|} < \infty$.

Recall that the function $\lceil\!\lceil f \rceil\!\rceil_\cdot$ is necessarily upper continuous (e.g. [Hoe07, Theorem 2.3]). In particular, for every ρ with $\lceil\!\lceil f \rceil\!\rceil_\rho < \infty$ there exists an $\varepsilon \in \mathbb{Q}^>$ with $\lceil\!\lceil f \rceil\!\rceil_{(1+\varepsilon)\rho} < \infty$.

8.2. Multivariate analytic functions as evaluable functions

In practice, multivariate analytic functions such as $\exp(x \log(1-y)) \operatorname{erf}(x^2 - y^2)$ are often built up as dags from univariate analytic functions such as log, exp and erf. In that case, it would be very expensive to systematically use power series expansions in several variables in order to compute with such functions. Instead, it would be better to represent such multivariate analytic functions as objects which can be *evaluated* at analytic functions in an arbitrary number of variables, or even at points in more general algebras.

More precisely, let $\mathbb{A}^{\text{com}} \supseteq \mathbb{C}^{\text{com}}$ be an effective Banach algebra over $\mathbb{C}^{\text{com}}$. This means that $\mathbb{A}^{\text{com}}$ is the set of computable points in a Banach algebra $\mathbb{A}$ over $\mathbb{C}$, and that the operations $+, -, \times$ and the norm $\|\cdot\|: \mathbb{A} \to \mathbb{R}^{\geqslant}$ can be computed by algorithm. Recall that

$$\begin{aligned} \|a+b\| &\leqslant \|a\| + \|b\| \\ \|a\,b\| &\leqslant \|a\|\,\|b\|, \end{aligned}$$

for all $a, b \in \mathbb{A}$. We do not necessarily assume $\mathbb{A}$ to be commutative. Given a multivariate analytic function f in $z_1, \ldots, z_n$ with $\lceil\!\lceil f \rceil\!\rceil_\rho < \infty$ and pairwise commuting $a_1, \ldots, a_n \in \mathbb{A}$ with $\|a_i\| \leqslant \rho_i$ for $i = 1, \ldots, n$, the evaluation

$$b = f(a_1, \ldots, a_n) \;=\; \sum_{k \in \mathbb{N}^n} f_k\, a_1^{k_1} \cdots a_n^{k_n} \tag{27}$$

is well defined. What is more: if $f \in \mathbb{F}^{\text{com}}$ and $a_1, \ldots, a_n \in \mathbb{A}^{\text{com}}$, then $f(a_1, \ldots, a_n) \in \mathbb{A}^{\text{com}}$. Indeed, we start by computing $\varepsilon \in \mathbb{Q}^{>}$ and $M = \lceil\!\lceil f \rceil\!\rceil_{(1+\varepsilon)\rho} < \infty$ such that $M < \infty$. For fixed $K \in \mathbb{N}$ and

$$\tilde{b} \;=\; \sum_{k_1, \ldots, k_n < K} f_k\, a_1^{k_1} \cdots a_n^{k_n},$$

we then have

$$\begin{aligned} \|\tilde{b} - b\| &\leqslant \sum_{i=1}^{n} \sum_{k} \|f_k\, a_1^{k_1} \cdots a_n^{k_n}\, a_i^K\|, \\ &\leqslant n\,M \left(\frac{1+\varepsilon}{\varepsilon}\right)^n \left(\frac{1}{1+\varepsilon}\right)^{Kn}. \end{aligned}$$

By choosing K sufficiently large, we may thus make $\|\tilde{b} - b\|$ as small as desired.

Conversely, assume now that, in the definition of computable multivariate analytic functions, we replace the method series: $\mathbb{F}^{\text{com}} \to \mathbb{C}^{\text{com}}[[z_1, \ldots, z_n]]^{\text{com}}$ by an evaluation method $\mathsf{eval}_{\mathbb{A}^{\text{com}}}: \mathbb{F}^{\text{com}} \times (\mathbb{A}^{\text{com}})^n \to \mathbb{A}^{\text{com}}$ for *any* effective Banach algebra $\mathbb{A}^{\text{com}}$ over $\mathbb{C}^{\text{com}}$. In particular, given $\rho \in (\mathbb{R}^{\text{com}})^n$, we may take $\mathbb{A}$ to be the algebra of all formal power series $\varphi \in \mathbb{C}[[z_1, \ldots, z_n]]$ for which

$$\|\varphi\| \;=\; \sup_{k \in \mathbb{N}^n} \|\varphi_k\, \rho^k\|$$

is finite. Given $f \in \mathbb{F}_{\text{com}}$ with $r_f > \rho$, it follows that the evaluation $\mathsf{eval}_{\mathbb{A}^{\text{com}}}(f, z_1, \ldots, z_n)$ is well defined, and this evaluation yields the power series expansion of f at the origin. This shows that providing an evaluation method $\mathsf{eval}_{\mathbb{A}^{\text{com}}}$ is essentially equivalent to providing a method series for series expansion.

Remark 11. In fact, analytic continuation and bound computation can also be regarded as evaluations in suitable "Banach algebras". Indeed, the analytic continuation at $\delta \in \mathbb{C}^{\text{com}}$ corresponds to the evaluation at the analytic function $\delta + z$. The computation of the bound $\lceil\!\lceil f \rceil\!\rceil_\rho$ can be done by evaluating f at the ball $\mathcal{B}(0, \rho)$ and taking $\lceil\!\lceil f \rceil\!\rceil_\rho = \|f(\mathcal{B}(0, \rho))\|$, where $\|\mathcal{B}(c, r)\| = |c| + r$. Nevertheless, explicit methods for analytic continuation and bound computations are usually of a better quality. They may also be needed for the implementation of more general evaluation methods.

Following the same line of ideas, it may also be useful to consider the evaluation of analytic functions at broken line paths (or more general piecewise analytic paths) instead of ordinary points, thereby combining analytic continuation and ordinary evaluation in a single method. One might even consider evaluations at "paths" $\mathbb{A}_1, \ldots, \mathbb{A}_k$ of successive Banach algebras of a similar type, such as $\mathbb{A}_i = \mathbb{C}[z] / (z^2 - \alpha_i\, z - \beta_i)$ with the α_i and the β_i sufficiently close, and $\alpha_1 = \beta_1 = 0$.

8.3. Evaluation in commutative zero dimensional algebras

In order to simplify notations, we will now stop our digression on the abstract notions of computability and drop the superscripts "com". In view of what we have seen in section 9.2, it is particularly interesting to evaluate multivariate analytic functions in commutative zero dimensional algebras $\mathbb{A}$ over $\mathbb{C}$. Therefore, we will now study this special case in more detail.

Let $\mathbb{A}$ be a finite dimensional $\mathbb{C}$-algebra. Given any basis for $\mathbb{A}$, the elements of $\mathbb{A}$ can be represented by matrices, so $\mathbb{A}$ may be identified with a commutative subalgebra of the algebra $\mathbb{C}^{k\times k}$ of $k\times k$ matrices for some k. In particular, any matrix norm on $\mathbb{C}^{k\times k}$ induces a norm on $\mathbb{A}$.

Given a multivariate analytic function f at the origin which is convergent on a polydisk of polyradius $\rho\in(\mathbb{R}^>)^n$, we may use (27) in order to evaluate f at any points $a_1,\ldots,a_n\in\mathbb{A}$ with $\|a_i\|\leqslant\rho_i$ for all i. Since the a_i commute, it is actually possible to do a bit better. Indeed, it is classical that there exists an invertible matrix U (corresponding to a base change), such that

$$U a_i U^{-1} = \begin{pmatrix} T_{i,1} & & \\ & \ddots & \\ & & T_{i,l} \end{pmatrix} = \begin{pmatrix} \lambda_{i,1}+E_{i,1} & & \\ & \ddots & \\ & & \lambda_{i,l}+E_{i,l} \end{pmatrix},$$

where $\lambda_{i,j}\in\mathbb{C}$, each $E_{i,j}$ is a nilpotent triagonal $k_j\times k_j$ matrix, and $k=k_1+\cdots+k_l$. We may thus compute $f(a_1,\ldots,a_n)$ using

$$f(a_1,\ldots,a_n) = U\begin{pmatrix} f(T_{1,1},\ldots,T_{n,1}) & & \\ & \ddots & \\ & & f(T_{1,l},\ldots,T_{n,l}) \end{pmatrix}U^{-1}.$$

For any j with $k_j=1$, we notice that the evaluation $f(T_{1,j},\ldots,T_{n,j})$ reduces to an evaluation $f(\lambda_{1,j},\ldots,\lambda_{n,j})$ at an ordinary point. If $n=1$, then we also notice that

$$f(T_{1,j}) = \sum_{i=0}^{r} f^{(i)}(\lambda_{1,j})\,E_{1,j}^i,$$

where r is minimal with $E_{1,j}^{r+1}=0$.

8.4. Weierstrass preparation

8.4.1. Implicit functions

Let us start with a special case of Weierstrass preparation: the implicit function theorem. Consider a computable analytic function f in $z_1,\ldots,z_n$ in the neighbourhood of a point $\alpha=(\alpha_1,\ldots,\alpha_n)$. Assume that $f(\alpha)=(\partial f/\partial z_n)(\alpha)=0$ but $(\partial f/\partial z_n)(\alpha)\neq 0$. Then the implicit function theorem states that there exists a unique analytic function $g=\mathsf{solve}(f,\alpha)$ in $z_1,\ldots,z_{n-1}$ such that $f(z_1,\ldots,z_{n-1},g(z_1,\ldots,z_{n-1}))=0$ in a neighbourhood of $(\alpha_1,\ldots,\alpha_{n-1})$ and $g(\alpha_1,\ldots,\alpha_{n-1})=\alpha_n$. It is not hard to check that g is computable, for instance in the sense of section 8.1:

- The coefficients of g can be computed using the classical formulas for implicit functions or using relaxed power series evaluation [Hoe02].
- Bound computations can be done using Corollary 7.
- Given sufficiently small $\delta_1,\ldots,\delta_{n-1}$ and $\delta_n=g(\alpha_1+\delta_1,\ldots,\alpha_{n-1}+\delta_{n-1})-\alpha_n$, the analytic continuation of g from $(\alpha_1,\ldots,\alpha_{n-1})$ to $(\alpha_1+\delta_1,\ldots,\alpha_{n-1}+\delta_{n-1})$ is simply $g_{+\delta}=\mathsf{solve}(f_{+\delta},\alpha+\delta)$.

More generally, given k analytic functions $f=(f_1,\ldots,f_k)$ such that the Jacobian matrix $J=\partial f/\partial z$ has full rank k at α, then it can be shown in a similar way that the system of k equations in k unknowns $g_1,\ldots,g_k$

$$\begin{aligned} f(z_1,\ldots,z_{n-k},g_1(z_1,\ldots,z_{n-k}),\ldots,g_k(z_1,\ldots,z_{n-k})) &= 0\\ g_i(\alpha_1,\ldots,\alpha_{n-k}) &= \alpha_{n-k+i}\end{aligned}$$

admits a unique analytic solution at $\alpha_1,\ldots,\alpha_{n-k}$ which is again computable. Notice that this more general case also follows through a k-fold application of the case of a single function.

8.4.2. Weierstrass preparation

Let us still consider a computable analytic function f in $z_1,\ldots,z_n$ in the neighbourhood of a point $\alpha=(\alpha_1,\ldots,\alpha_n)$. Assume now that $f(\alpha)=(\partial f/\partial z_n)(\alpha)=\cdots=(\partial^{r-1}f/\partial z_n^{r-1})(\alpha)=0$ but $(\partial^r f/\partial z_n^r)(\alpha)\neq 0$. Then the Weierstrass preparation theorem states that there exist unique analytic functions $g_0,\ldots,g_{r-1}$ in $z_1,\ldots,z_{n-1}$ and an invertible analytic function u in $z_1,\ldots,z_n$ such that

$$f = (z_n^r+g_{r-1}z_n^{r-1}+\cdots+g_0)\,u. \tag{28}$$

We may regard (28) as a local factorization of the analytic function f in z_n, which depends on $z_1,\ldots,z_{n-1}$ as parameters. Now Theorem 10 for the computation of local factorizations readily generalizes to computable analytic functions. More precisely, $g=z_n^r+g_{r-1}z_n^{r-1}+\cdots+g_0$ is the solution of the system $f\operatorname{rem}_{z_n}g=0$, which we consider as a system of r analytic equations $(f\operatorname{rem}_{z_n}g)_0=\cdots=(f\operatorname{rem}_{z_n}g)_{r-1}=0$ in $n-1+r$ unknowns $z_1,\ldots,z_{n-1},g_0,\ldots,g_{r-1}$. It can be checked that this system of equations admits full rank, so that we may compute $g_0,\ldots,g_{r-1}$ by applying the effective implicit function.

8.4.3. Certifying the multiplicity in a small neighbourhood

Effective Weierstrass preparation as described in the previous section can only be applied on rare occasions. Indeed, the assumption that we have an exact multiple cancellation $f(\alpha)=(\partial f/\partial z_n)(\alpha)=\cdots=(\partial^{r-1}f/\partial z_n^{r-1})(\alpha)=0$ is too ambitious, and needs to be replaced by the weaker condition that f has a multiplicity r in z_n in a small neighbourhood of α. This condition is still enough for the existence of a unique solution to the system $f\operatorname{rem}_{z_n}g$ in a neighbourhood of α.

This naturally leads us to the question how to certify that f has multiplicity r in z_n, in a small neighbourhood of α, and uniformly in $z_1,\ldots,z_{n-1}$. In fact, paying some care, all methods from section 7 can be generalized to this setting. First of all, when using ball arithmetic, the parameters $z_1,\ldots,z_{n-1}$ are simply replaced by small balls $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_{n-1}$ around $\alpha_1,\ldots,\alpha_{n-1}$, which reduces the problem to a univariate one. Secondly, when working with Taylor models at any chosen truncation order T, the tail $f_T\,(z_n-\alpha_n)^T+f_{T+1}\,(z_n-\alpha_n)^{T+1}+\cdots$ of an analytic function near α_n is replaced by a ball of the form $\mathcal{B}_{\boldsymbol{\alpha}_n}(\varepsilon)$ or $\mathcal{B}_{\boldsymbol{\alpha}_n}(\varepsilon\,z_n^T)$. This reduction allows us to work with polynomials instead of series.

We also notice that the above certification methods can actually prove something slightly stronger than a uniform multiplicity: they can usually be used to verify that the zero set $\mathcal{Z}=\{z\in\boldsymbol{\alpha}: f(z)=0\}$ does not intersect $\boldsymbol{\alpha}_1\times\cdots\times\boldsymbol{\alpha}_{n-1}\times\partial\boldsymbol{\alpha}_n$. If such is the case, then we say that the equation $f(z)=0$ is *equisolvable* in z_n on $\boldsymbol{\alpha}$.

8.4.4. Analytic elimination on small balls

Given a ball $\boldsymbol{\alpha}\in\mathbb{C}^n$ (and where we recall that such a "ball" is really a polydisk) such that we can certify a constant multiplicity r of f in z_n on $\boldsymbol{\alpha}$, the unique (and computable) analytic function $g=z_n^r+g_{r-1}z_n^{r-1}+\cdots+g_0$ such that f/g is an analytic unit on $\boldsymbol{\alpha}$ will be called the Weierstrass normalization of f on $\boldsymbol{\alpha}$.

Given two analytic functions f and φ on $\boldsymbol{\alpha}$, assume that we can certify a constant multiplicity for at least one of them, say for f, and let g be as above. Then Weierstrass division of φ by g yields unique analytic functions χ and ρ such that $\varphi = \chi g + \rho$ and ρ is polynomial in z_n with $\deg_{z_n} \rho < r$. We may obtain (the residue class of) ρ as a computable analytic function through evaluation of φ in the quotient algebra $\mathbb{F}^{\text{com}} / (g)$ of computable analytic functions in $z_1, \ldots, z_n$ by the relation $g = 0$.

Having computed the remainder $\rho = \varphi \operatorname{rem} g$ of the Weierstrass division of φ by g, we may finally form the resultant $R = \operatorname{Res}_{z_n}(\rho, g)$ of the polynomials ρ and g in z_n. This resultant is a computable analytic function in $z_1, \ldots, z_{n-1}$ whose zeros are precisely the projections of the common zeros of f and φ on $\mathbb{C}^{n-1}$. We will also call $R = \operatorname{Res}_{z_n}(f, \varphi)$ the *analytic resultant* of f and g on $\boldsymbol{\alpha}$. Whenever $\operatorname{Res}_{z_n}(\varphi, f)$ is also defined, it can be checked that $\operatorname{Res}_{z_n}(\varphi, f) = w \operatorname{Res}_{z_n}(f, \varphi)$ for some analytic unit w on $\boldsymbol{\alpha}_1 \times \cdots \times \boldsymbol{\alpha}_{n-1}$. Indeed, writing $f = u g$ and $\varphi = v \psi$ for analytic units u and v, $\operatorname{Res}_{z_n}(f, \varphi) \equiv \operatorname{Res}_{z_n}(g, \varphi) = \operatorname{Res}_{z_n}(g, \varphi \operatorname{rem} g) = \operatorname{Res}_{z_n}(g, v \operatorname{rem} g) \operatorname{Res}_{z_n}(g, \psi \operatorname{rem} g) = \operatorname{Res}_{z_n}(g, v \operatorname{rem} g) \operatorname{Res}_{z_n}(g, \psi)$ and $\operatorname{Res}_{z_n}(g, v \operatorname{rem} g)$ is a unit. Similarly, $\operatorname{Res}_{z_n}(\varphi, f) = \operatorname{Res}_{z_n}(\psi, u \operatorname{rem} \psi) \operatorname{Res}_{z_n}(\psi, g)$.

8.5. Simplifying computable meromorphic functions

A computable meromorphic function is simply the quotient of two computable analytic functions such that the denominator is non zero. The computable meromorphic functions on a given domain form an effective field (but without a zero test). Using an external argument, we sometimes know that a computable meromorphic function f / g is actually analytic on some domain. In this case, we would like to conclude that f / g is computably analytic on this domain.

8.5.1. A reliable version of l'Hôpital's rule

Let us first consider the case when f and g are both univariate computable analytic functions in z, in the neighbourhood of a single isolated zero α. We start with the subcase when $g = z - \alpha$ on a ball $\boldsymbol{\alpha}$ with $\alpha \in \boldsymbol{\alpha}$. Let $\boldsymbol{\sigma} = f'(\boldsymbol{\alpha})$ be the ball evaluation of f' at $\boldsymbol{\alpha}$. For all $z \in \boldsymbol{\alpha}$, we then we have

$$f(z) = \int_{\alpha}^{z} f'(u) \, \mathrm{d}u \subseteq \int_{\alpha}^{z} \boldsymbol{\sigma} \, \mathrm{d}u \subseteq \boldsymbol{\sigma} \, (z - \alpha).$$

Consequently, $(f / (z - \alpha))(\boldsymbol{\alpha}) \subseteq f'(\boldsymbol{\alpha})$. For general g, we obtain in a similar way that $(g / (z - \alpha))(\boldsymbol{\alpha}) \subseteq g'(\boldsymbol{\alpha})$. We conclude that

$$\frac{f}{g}(\boldsymbol{\alpha}) \;\subseteq\; \frac{f'}{g'}(\boldsymbol{\alpha}). \tag{29}$$

8.5.2. Higher multiplicities

It would be interesting to know whether there exist generalizations of (29) to higher multiplicities or involving the successive derivatives of f / g. In this section we will present a more *ad hoc* strategy for computing a reliable Taylor series expansion of f / g in the case when f, g and f / g are all analytic.

Let $\boldsymbol{\alpha}$ and $\mathbf{A}$ be all small and a larger ball with the same center c and radii ε and R. Assume that f admits exactly r zeros $\alpha_1, \ldots, \alpha_r$ both on $\boldsymbol{\alpha}$ and on $\mathbf{A}$. Let $g = (z - \alpha_1) \cdots (z - \alpha_r)$ and $h = f / g$. Writing $\boldsymbol{f}_i = f^{(i)}(\boldsymbol{\alpha}) / i!$, $\boldsymbol{g}_i = g^{(i)}(\boldsymbol{\alpha}) / i!$ and $\boldsymbol{h}_i = h^{(i)}(\boldsymbol{\alpha}) / i!$ for each i, we have

$$\boldsymbol{h}_i \;\subseteq\; \boldsymbol{f}_i - (\boldsymbol{h}_{i+1} \boldsymbol{g}_{r-1} + \cdots + \boldsymbol{h}_{i+r} \boldsymbol{g}_0). \tag{30}$$

Let M be an upper bound for $|f(\mathbf{A})|$. Since h achieves its maximum on $\mathbf{A}$ at its border, we also have

$$|h(\mathbf{A})| \leqslant \frac{M}{(R-\varepsilon)^r}.$$

From Cauchy's formula, it follows that

$$\boldsymbol{h}_i \subseteq \frac{\mathcal{B}(0,M)}{(R-\varepsilon)^{i+r}} \tag{31}$$

for all i. Plugging this into (30), we obtain

$$\boldsymbol{h}_i \subseteq \boldsymbol{f}_i + \frac{\mathcal{B}(0,M)}{(R-\varepsilon)^{i+r}}\left[\frac{1}{(1-\varepsilon/R)^r}-1\right]. \tag{32}$$

For the small balls, we may thus use (32) for the computation of an enclosure for $\boldsymbol{h}_i$, and switch to (31) whenever this enclosure becomes better.

8.5.3. Recovering analytic functions from their values on a circle

More generally, it is in principle possible to recover any analytic function f on a closed disk D from its values on its boundary using Cauchy's formula. We will now describe an efficient method to do this effectively. When applied to computable meromorphic functions as above, this method has the advantage that we do not need to know the numerator and denominator explicitly; it is sufficient that we can compute its ball values on the boundary circle. Without loss of generally, we may assume that D is the unit disk.

For a given number N of evaluation points (preferably a power of two) and with $\delta = \left|\mathrm{e}^{\pi \mathrm{i}/N}-1\right|$, we evaluate f at the disks $\boldsymbol{\omega}_k = \mathcal{B}\left(\mathrm{e}^{2\pi \mathrm{i}k/N}, \delta\right)$ for $k \in \{0,\ldots,N-1\}$, yielding $\boldsymbol{v}_k = f(\boldsymbol{\omega}_k) = \mathcal{B}(v_k, \varepsilon_k)$. We next compute ball enclosures $\boldsymbol{P}_i$ for the coefficients of the unique polynomial P with $\deg P < n$ and $P\left(\mathrm{e}^{2\pi \mathrm{i}k/N}\right) = v_k$ for all k using one reliable discrete Fourier transform. The remainder $R = f - P$ is bounded by $\varepsilon = \max_k \varepsilon_k$ on the unit circle, whence on the unit disk. This yields a Taylor model $\boldsymbol{P} + \mathcal{B}_D(\varepsilon)$ for f. Taking N sufficiently large, the so computed Taylor model can be made as accurate as needed.

8.5.4. Multivariate meromorphic functions

Let us now consider a meromorphic function f/g in several variables $z_1, \ldots, z_n$ which is actually analytic on some small ball. Modulo a linear change of coordinates and shrinking the domain to a smaller ball $\boldsymbol{\alpha}$, we may first ensure that g has constant multiplicity in z_n on $\boldsymbol{\alpha}$. We may then use any of the methods from the previous section for computing f/g as an analytic function in z_n, which depends analytically on the parameters $z_1, \ldots, z_{n-1}$.

9. Certification of multiple multivariate roots

9.1. Jacobians of corank one

Consider a system (11) and assume that we have numerically isolated a solution α of multiplicity r. Based on numerical computations, assume also that the Jacobian $J = \partial P/\partial z$ is expected to have corank at most one at α.

In a small ball $\boldsymbol{\alpha}$ around α (and where we recall that $\boldsymbol{\alpha}$ is really a polydisk) we may certify the corank assumption by computing $J(\boldsymbol{\alpha})$ using ball arithmetic and then checking that an $(n-1) \times (n-1)$ submatrix admits a non zero determinant (still using ball arithmetic).

At the next stage, we would like to certify that there exist indeed exactly r solutions in a suitable neighbourhood of α, when counted with multiplicities. Now using our effective version of the implicit function theorem from section 8.4.1, we may analytically eliminate at least $n-1$ variables from the equations (11), say $z_2,\ldots,z_n$ from the equations $P_2(z)=\cdots=P_n(z)=0$. This yields computable analytic functions $\Phi_2(z_1),\ldots,\Phi_n(z_1)$ such that

$$P_i(z_1,\Phi_2(z_1),\ldots,\Phi_n(z_1)) \;=\; 0$$

for all $i\in\{2,\ldots,n\}$. We next compute the computable analytic function

$$R(z_1) \;=\; P_1(z_1,\Phi_2(z_1),\ldots,\Phi_n(z_1)).$$

Using the techniques from section 8.4.3, we finally check that R has multiplicity r in a suitable neighbourhood $\boldsymbol{\alpha}_1$ of α_1.

Remark 12. One of the simplest examples when the corank one hypothesis fails is the following system:

$$\begin{aligned} z_1^2 &= 0\\ z_2^2 &= 0. \end{aligned}$$

9.2. Certification of herd homotopies

Let us still consider the system (11) and assume that we are given a herd of numeric solution paths $z_t^1,\ldots,z_t^r$ which all tend to the same limit α. In this section, we will investigate the problem of simultaneously certifying each of the paths in the herd.

We would like the follow the same strategy as in section 7.3, where we certified a single factor of order r instead of r separate roots. In the current setting, instead of viewing the z_t^i as distinct individual paths, this amounts to regarding the whole herd $t\mapsto\mathcal{Z}_t=\{z_t^1,\ldots,z_t^r\}$ as a single multivalued path. From the algebraic point of view, this means that we need to consider the ideal $\mathfrak{I}_t$ which annihilates $\mathcal{Z}_t$. There are several ways to represent this ideal $\mathfrak{I}_t$ by a system Σ_t of polynomial equations. We will use so called univariate representations.

Now recall that each z_t^i can be considered as a vector $z_t^i=(z_{t,1}^i,\ldots,z_{t,n}^i)$ of Puiseux series $z_{t,j}^i$ in t of valuations $\geqslant 0$. Setting

$$A_t(z_1) \;=\; (z_1-z_{t,1}^1)\cdots(z_1-z_{t,1}^r),$$

we notice that $A_t(z_1)$ is invariant if we turn t once around the origin. This means that $A_t(z_1)$ is really an analytic function in t at the origin. Assuming general position, $A_t(z_1)$ is actually the minimal annihilator of $\{z_{t,1}^1,\ldots,z_{t,1}^r\}$. Furthermore, for $t\neq 0$ sufficiently small, the roots $z_{t,1}^i$ are pairwise distinct, so there are unique interpolating polynomials $U_{t,2},\ldots,U_{t,n}\in\mathbb{C}[u]$ of degree $<r$ with

$$z_{t,i} \;=\; U_{t,i}(z_{t,1})$$

for all $i\in\{2,\ldots,n\}$. In what follows, we will represent $\mathfrak{I}_t$ by the system of polynomials

$$\begin{cases} A_t(z_{t,1})\\ z_{t,2}-U_{t,2}(z_{t,1})\\ \vdots\\ z_{t,n}-U_{t,n}(z_{t,1}), \end{cases} \tag{33}$$

where A_t is monomic of degree r and $\deg U_{t,i} < r$ for each i.

Solving the original system Σ for $z_1, \ldots, z_n$ of this form now amounts to a new system $\tilde{\Sigma}$ of $n\,r$ equations in the unknowns $(A_t)_j$ and $(U_{t,i})_j$. This system can also be considered as a system of n equations in n unknowns $z_1, \ldots, z_n$ over the algebra $\mathbb{A} = \mathbb{C}[u] / (A_t(u))$, augmented with one special unknown u^r and one special equation $z_1 = u$. In analogy with Theorem 10, we may hope that this new system is non singular near $t = 0$, so that we can certify the herd using Corollary 7. It can be checked that this is always so in the corank one case from the previous section. Unfortunately, Theorem 10 fails to generalize in general. In particular, the new system remains singular whenever several herds converge to the same singularity.

9.3. Equisolvability

Let us now return to the general case when the corank of the Jacobian matrix at the solution may be larger than one. First of all, it will be useful to generalize the notion of equisolvability from section 8.4.3 to the case of several equations. Consider the system

$$P_k(z) = \cdots = P_n(z) = 0 \tag{34}$$

on $\boldsymbol{\alpha} \in \mathbb{C}^n$, where $P_k, \ldots, P_n$ might actually be analytic functions instead of mere polynomials. We say that (34) is *equisolvable* in $z_k, \ldots, z_n$ on $\boldsymbol{\alpha}$ if the zero set

$$\mathcal{Z} = \mathcal{V}_{\boldsymbol{\alpha}}(P_{\geqslant k}) = \{z \in \boldsymbol{\alpha} : P_k(z) = \cdots = P_n(z) = 0\} \tag{35}$$

does not intersect $\boldsymbol{\alpha}_{<k} \times \partial \boldsymbol{\alpha}_{\geqslant k}$, where $\boldsymbol{\alpha}_{<k} = \boldsymbol{\alpha}_1 \times \cdots \times \boldsymbol{\alpha}_{k-1}$ and $\boldsymbol{\alpha}_{\geqslant k} = \boldsymbol{\alpha}_k \times \cdots \times \boldsymbol{\alpha}_n$. In particular, this means that the number of solutions $(z_k, \ldots, z_n) \in \boldsymbol{\alpha}_{\geqslant k}$ of (34) as a function of $(z_1, \ldots, z_{k-1}) \in \boldsymbol{\alpha}_{<k}$ does not depend on $z_1, \ldots, z_{k-1}$. Furthermore, the following lemma shows that there are no exceptional fibers with an infinite number of solutions.

Lemma 13. *Assume that (34) is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$ and that $\mathcal{X} \subseteq \boldsymbol{\alpha}_{<k}$ is the zero set of a non zero analytic function on $\boldsymbol{\alpha}_{<k}$. Then none of the components of $\mathcal{Z}$ can be contained entirely in the set $\mathcal{X} \times \boldsymbol{\alpha}_{\geqslant k}$.*

Proof. Assume for contradiction that $(\mathcal{X} \times \boldsymbol{\alpha}_{\geqslant k}) \cap \mathcal{Z}$ has dimension $\geqslant k - 1$. Taking $u \in \mathcal{X}$, the intersection $(\{u\} \times \boldsymbol{\alpha}_{\geqslant k}) \cap \mathcal{Z}$ has dimension $\geqslant 1$, whence it contains an analytic solution curve $t \mapsto \varphi(t) = (u, \varphi_k(t), \ldots, \varphi_n(t)) \in \mathcal{Z}$ whose image has dimension one. Since (34) is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$, the curve $(\varphi_k(t), \ldots, \varphi_n(t))$ is entirely contained in $\boldsymbol{\alpha}_{\geqslant k} \setminus \partial \boldsymbol{\alpha}_{\geqslant k}$. Consequently, the analytic functions $\varphi_i(t)$ are all bounded, whence constant. This contradicts our assumption that $\operatorname{im} \varphi$ has dimension one. □

Example 14. The equation

$$P(x, y) = 4\,y^2 - x \ = \ 0$$

is equisolvable in y on $\mathcal{B}(0,1)^2$, since there are always two solutions, but

$$Q(x, y) = y - 2\,x \ = \ 0$$

is not equisolvable, since the equation has one solution for $x = 0$, but no solutions for $x = 1$. Of course, Q is equisolvable on $\mathcal{B}(0, 1)^2$ for y on a larger disk (or x on a smaller disk). Graphically speaking, an equation $P(x, y) = 0$ is equisolvable if and only if all solution curves $y(x)$ are “nicely horizontal” in $\mathcal{B}(0,1)^2$.

9.4. Kronecker representations

With the notations from the previous section, consider polynomials

$$\begin{array}{rcl} Q & = & z_k^d + Q_{k-1} z_k^{d-1} + \cdots + Q_0 \\ V_i & = & V_{i,d-1} z_k^{d-1} + \cdots + V_{i,0}, \qquad i > k, \end{array}$$

where the coefficients Q_j and $V_{i,j}$ are computable analytic functions in $z_1, \ldots, z_{k-1}$ on $\boldsymbol{\alpha}_{<k}$, such that $\operatorname{Res}_{z_k}(Q, Q') \neq 0$ (with $Q' = \partial Q / \partial z_k$) and the system (34) is equivalent to

$$\left\{ \begin{array}{rcl} Q(z_{\leqslant k}) & = & 0 \\ Q'(z_{\leqslant k})\, z_i & = & V_i(z_{\leqslant k}), \quad i > k, \end{array} \right. \tag{36}$$

for all $z \in \boldsymbol{\alpha} \setminus (\mathcal{X} \times \boldsymbol{\alpha}_{\geqslant k})$, where $\mathcal{X} = \mathcal{V}_{\boldsymbol{\alpha}_{<k}}(\operatorname{Res}_{z_k}(Q, Q'))$. Then we call (36) the *Kronecker representation* for (34) on $\boldsymbol{\alpha}$.

LEMMA 15. *Assume that the system $P_1(z) = \cdots = P_n(z) = 0$ admits a finite number of solutions in $\boldsymbol{\alpha}$ and that the z_1-coordinates of these solutions are pairwise distinct. Then there exists a Kronecker representation for $P_1(z) = \cdots = P_n(z) = 0$ on $\boldsymbol{\alpha}$.*

Proof. Let $z^1, \ldots, z^k \in \mathbb{C}^n$ be such that $z_1^1, \ldots, z_1^k$ are pairwise distinct. Then we define the Kronecker representation for $\mathcal{Z} = \{z^1, \ldots, z^k\}$ to be the unique n-tuple $(Q, V_2, \ldots, V_n) = \mathrm{K}^{\mathcal{Z}} = (Q^{\mathcal{Z}}, V_2^{\mathcal{Z}}, \ldots, V_n^{\mathcal{Z}})$ of univariate polynomials with $\deg Q = k$, $Q_k = 1$, $\deg V_2 < k, \ldots, \deg V_n < k$, such that $\mathcal{Z}$ is annihilated by the system

$$\begin{array}{rcl} Q(z_1) & = & 0 \\ Q'(z_1)\, z_2 & = & V_2(z_1) \\ & \vdots & \\ Q'(z_1)\, z_n & = & V_n(z_1). \end{array}$$

We may compute such a Kronecker representation as follows. If $k = 1$, then we simply take $Q(z_1) = z_1 - z_1^1$ and $V_i = z_i^1$ for all $i > 1$. For $k > 1$, we decompose $\mathcal{Z} = \mathcal{X} \amalg \mathcal{Y}$ with $|\mathcal{X}| = \lfloor |\mathcal{Z}| / 2 \rfloor$ and compute $\mathrm{K}^{\mathcal{Z}}$ in terms of $\mathrm{K}^{\mathcal{X}}$ and $\mathrm{K}^{\mathcal{Y}}$ using

$$\begin{array}{rcl} Q^{\mathcal{Z}} & = & Q^{\mathcal{X}} Q^{\mathcal{Y}} \\ V_i^{\mathcal{Z}} & = & V_i^{\mathcal{X}} Q^{\mathcal{Y}} + Q^{\mathcal{X}} V_i^{\mathcal{Y}}, \quad i > 1. \end{array}$$

This yields an efficient dichotomic algorithm for the computation of $\mathrm{K}^{\mathcal{Z}}$. The result follows by applying the method to the solutions of $P_1(z) = \cdots = P_n(z) = 0$ in $\boldsymbol{\alpha}$. □

THEOREM 16. *Assume that (34) is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$. For the zeroset $\mathcal{X}$ of some non zero analytic function on $\boldsymbol{\alpha}_{<k}$, assume that for any $u \in \mathcal{U} := \boldsymbol{\alpha}_{<k} \setminus \mathcal{X}$, the solutions of (34) with $z_{<k} = u$ admit pairwise distinct z_k-coordinates. Then there exists a Kronecker representation for (34).*

Proof. For any $u \in \mathcal{U}$, the above lemma yields a Kronecker representation

$$\begin{array}{rcl} Q(u, z_k) & = & 0 \\ Q'(u, z_k)\, z_i & = & V_i(u, z_k), \quad i > k \end{array}$$

for the system (34) with $z_{<k} = u$. Moreover, the equisolvability assumption implies that the degree d of Q^u does not depend on u. Furthermore, since $Q'(u, z_k) \neq 0$ for $u \in \mathcal{U}$, the way we compute the functions $Q, V_{k+1}, \ldots, V_n$ depends analytically on u. We thus obtain a "Kronecker representation" for (34) on the subset $\mathcal{U} \times \boldsymbol{\alpha}_{\geqslant k}$ instead of $\boldsymbol{\alpha}$.

Let $M>0$ be such that $|\boldsymbol{\alpha}_k|\leqslant M$. Expanding $Q(u,z_k)=z_k^d+Q_{d-1}(u)\,z_k^{d-1}+\cdots+Q_0(u)$ and $(z_k-M)^d=z_k^d+\bar{Q}_{d-1}\,z_k^{d-1}+\cdots+\bar{Q}_0$ with $\bar{Q}_i=M^{d-i}\binom{d}{i}$, we have $|Q_i(u)|\leqslant\bar{Q}_i$ for all $u\in\mathcal{U}$. Consequently, $Q(u,z_k)$ remains bounded on $\mathcal{U}\times\boldsymbol{\alpha}_k$. Since $\mathcal{X}$ is a removable isolated singularity, it follows that $Q(u,z_k)$ admits a unique analytic continuation on $\mathcal{U}$. It also follows that we may also choose $\mathcal{X}$ to be the zero set of $\operatorname{Res}_{z_k}(Q,Q')$.

It remains to be shown that the V_i with $i>k$ can also extended analytically from $\mathcal{U}\times\boldsymbol{\alpha}_k$ to $\boldsymbol{\alpha}_{\leqslant k}$. For this, we consider a new coordinate $z_k^\lambda=z_k+\lambda_{k+1}\,z_{k+1}+\cdots+\lambda_n\,z_n$ where $\lambda_{k+1},\ldots,\lambda_n$ are parameters. For $(\lambda_{k+1},\ldots,\lambda_n)$ sufficiently close to $(0,\ldots,0)$, we may apply the above result to the system $P_k(z)=\cdots=P_n(z)$ rewritten as equations $P_k^\lambda(z^\lambda)=\cdots=P_n^\lambda(z^\lambda)$ in $z^\lambda=\left(z_{<k},z_k^\lambda,z_{>k}\right)$. Notice that this new system is defined on a domain $\boldsymbol{\alpha}^\lambda$ which is slightly smaller than $\boldsymbol{\alpha}$. We thus obtain a Kronecker representation

$$\begin{aligned}Q^\lambda\left(z_{\leqslant k}^\lambda\right)&=0\\(Q^\lambda)'\left(z_{\leqslant k}^\lambda\right)z_i&=V_i^\lambda\left(z_{\leqslant k}^\lambda\right),\quad i>k,\end{aligned}$$

which is equivalent to $P_k^\lambda(z^\lambda)=\cdots=P_n^\lambda(z^\lambda)$ for z^λ with $R^\lambda\left(z_{<k}^\lambda\right)\neq0$, where $R^\lambda=\operatorname{Res}_{z_k^\lambda}(Q^\lambda,(Q^\lambda)')$. The function Q^λ is defined on $\boldsymbol{\alpha}^\lambda$ and analytic both in λ and z^λ. Differentiating the equation $Q^\lambda\left(z_{\leqslant k}^\lambda\right)=0$ with respect to λ_i with $i>k$, we obtain

$$(Q^\lambda)'\left(z_{\leqslant k}^\lambda\right)z_i=-\frac{\partial Q^\lambda}{\partial\lambda_i}\left(z_{\leqslant k}^\lambda\right).$$

Letting λ tend to 0, we obtain that $V_i(z_{\leqslant k})=-(\partial Q^0/\partial\lambda_i)(z_{\leqslant k})$. □

Proposition 17. *Assume that $P_{k+1}(z)=\cdots=P_n(z)$ is equisolvable in $z_{>k}$ on $\boldsymbol{\alpha}$. Assume that we are given a Kronecker representation (36) for (34) such that $Q(z_{\leqslant k})=0$ is equisolvable in z_k on $\boldsymbol{\alpha}_{\leqslant k}$ and such that Q vanishes on $\mathcal{V}_{\boldsymbol{\alpha}}(P_{>k})$. Then (34) is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$.*

Proof. Assume for contradiction that $z\in\mathcal{Z}\cap\boldsymbol{\alpha}_{<k}\times\partial\boldsymbol{\alpha}_{\geqslant k}$. Since $P_{k+1}(z)=\cdots=P_n(z)=0$ is equisolvable in $z_{k+1},\ldots,z_n$ on $\boldsymbol{\alpha}$, we cannot have $z\in\mathcal{Z}\cap\boldsymbol{\alpha}_{\leqslant k}\times\partial\boldsymbol{\alpha}_{>k}$. Consequently, $z\in\boldsymbol{\alpha}_{<k}\times\partial\boldsymbol{\alpha}_k\times\boldsymbol{\alpha}_{>k}$. Hence, we both have $Q(z_{\leqslant k})=0$ and $z_{\leqslant k}\in\boldsymbol{\alpha}_{<k}\times\partial\boldsymbol{\alpha}_k$. This contradicts the assumption that Q is equisolvable in z_k on $\boldsymbol{\alpha}_{\leqslant k}$. □

9.5. Univariate representations

The *univariate representation* is a variant of the Kronecker representation. It consists of polynomials

$$\begin{aligned}Q&=z_k^d+Q_{k-1}\,z_k^{d-1}+\cdots+Q_0\\U_i&=U_{i,d-1}\,z_k^{d-1}+\cdots+U_{i,0},\qquad i>k,\end{aligned}$$

such that the coefficients Q_j are computable analytic functions in $z_{<k}$ on $\boldsymbol{\alpha}_{<k}$, the coefficients $V_{i,j}$ are computable meromorphic functions in $z_{<k}$ which are analytic on $\boldsymbol{\alpha}_{<k}\setminus\mathcal{X}$, where $\mathcal{X}=\mathcal{V}_{\boldsymbol{\alpha}_{<k}}(\operatorname{Res}_{z_k}(Q,Q'))$ and $\operatorname{Res}_{z_k}(Q,Q')\neq0$. Finally, the system (34) is assumed to be equivalent to

$$\left\{\begin{aligned}Q(z_{\leqslant k})&=0\\z_i&=U_i(z_{\leqslant k}),\quad i>k,\end{aligned}\right.\tag{37}$$

for all $z\in\boldsymbol{\alpha}\setminus(\mathcal{X}\times\boldsymbol{\alpha}_{\geqslant k})$. Given a Kronecker representation, we may compute a univariate representation as follows. We first compute the inverse $I=I_{d-1}\,z_k^{d-1}+\cdots+I_0$ of Q' modulo Q, whose coefficients are meromorphic and analytic on $\mathcal{U}=\boldsymbol{\alpha}_{<k}\setminus\mathcal{X}$. Taking $U_i=I\,V_i\bmod Q$, we may then retrieve the univariate representation from the Kronecker representation.

9.6. Intersection with a new equation

Given a Kronecker representation satisfying the hypothesis of Proposition 17, let us now study the intersection problem with a new equation $P_{k-1}(z)=0$. We first compute a univariate representation for (34), as detailed in section 9.5. Now consider the resultant

$$R(z_{<k}) \;=\; \mathrm{Res}_{z_k}(Q(z_{\leqslant k}), P_{k-1}(z_{\leqslant k}, U_{>k}(z_{\leqslant k}))).$$

In practice, we may compute R by evaluating P_{k-1} at $(z_{\leqslant k},\, U_{>k}(z_{\leqslant k}))$ in the algebra $\mathbb{B}=\mathbb{A}[z_k]/(Q)$ where $\mathbb{A}$ is the set of computable analytic functions on $\mathcal{U}$, and then take the norm of this evaluation. Let $\phi_1(z_{<k}), \ldots, \phi_d(z_{<k})$ denote the solutions of (34) in $z_{\geqslant k}$ as a function of $z_{<k}$, considered as analytic functions in $z_{<k}$ on some Riemann surface above $\mathcal{U}$. By a classical property of resultants, we have

$$R(z_{<k}) \;=\; \prod_{i=1}^{d} P_{k-1}(z_{<k}, \phi_i(z_{<k})) \tag{38}$$

on $\mathcal{U}$. Whereas the functions $\phi_i(z_{<k})$ generally admit a non trivial monodromy, the value of $R(z_{<k})$ is uniquely determined as a function on $z_{<k}$ on $\mathcal{U}$. Moreover, since $\phi_i(z_{<k})\in\boldsymbol{\alpha}_{\geqslant k}$ for $z_{<k}\in\boldsymbol{\alpha}_{<k}$, the analytic function R is bounded on $\mathcal{U}$. Since the complement $\mathcal{X}$ of $\mathcal{U}$ in $\boldsymbol{\alpha}_{<k}$ is isolated, this means that R extends uniquely into an analytic function on $\boldsymbol{\alpha}_{<k}$.

Lemma 18. *Given a Kronecker representation (36) for $P_{\geqslant k}(z)=0$ and assuming that $P_{\geqslant k}(z)=0$ is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$, the resultant R vanishes on $\mathcal{V}_{\boldsymbol{\alpha}}(P_{\geqslant k-1})$.*

Proof. If $z\in\mathcal{V}_{\boldsymbol{\alpha}}(P_{\geqslant k-1})\cap(\mathcal{U}\times\boldsymbol{\alpha}_{>k})$, then (38) directly yields $R(z)=0$. For general $z\in\mathcal{V}_{\boldsymbol{\alpha}}(P_{\geqslant k-1})$, let $u\in\mathbb{C}^{k-1}$ be a vector such such that $z_{<k}+u\,\varepsilon\in\mathcal{U}$ for all sufficiently small $\varepsilon\neq 0$. Then $f_i(\varepsilon)=P_{k-1}(z_{<k}+u\,\varepsilon,\, \phi_i(z_{<k}+u\,\varepsilon))$ is a bounded function, given by a Puiseux series of valuation $\geqslant 0$ in ε. Since $P_{\geqslant k}(z)$ is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$, one of the curves $(z_{<k}+u\,\varepsilon,\, \phi_i(z_{<k}+u\,\varepsilon))$ tends to z for $\varepsilon\rightarrow 0$. Since $P_{k-1}(z)=0$, it follows that $f_i(\varepsilon)$ and therefore $f_1(\varepsilon)\cdots f_d(\varepsilon)$ tend to 0 for $\varepsilon\rightarrow 0$. We conclude that $R(z)=0$, using the analyticity of R. □

We have shown above how to find an analytic relation R for z_{k-1} which is implied by $P_{\geqslant k}$ on $\boldsymbol{\alpha}$. We still need to show how to express $z_{\geqslant k}$ as a function of $z_{<k}$. For this we use a similar technique as in the proof of Theorem 16. Consider formal variables $\epsilon_k, \ldots, \epsilon_n$ with $\epsilon_i\,\epsilon_j=0$ for all i, j and new coordinates z^{ϵ} such that

$$\begin{aligned} z_{k-1} &= z_{k-1}^{\epsilon} - \epsilon_{\geqslant k}\cdot z_{\geqslant k}^{\epsilon} \\ z_i &= z_i^{\epsilon}, \qquad i\neq k-1. \end{aligned}$$

With respect to these new coordinates, the univariate representation for (34) may be rewritten as

$$\begin{aligned} Q^{\epsilon}(z_{\leqslant k}^{\epsilon}) &= 0 \\ z_i^{\epsilon} &= U_i^{\epsilon}(z_{\leqslant k}^{\epsilon}), \quad i>k. \end{aligned}$$

In a similar way as above above, we next evaluate the resultant

$$\begin{aligned} R^{\epsilon}(z_{<k}^{\epsilon}) &= \mathrm{Res}_{z_k^{\epsilon}}(Q^{\epsilon}(z_{\leqslant k}^{\epsilon}), P_{k-1}(z_{<k-1}^{\epsilon}, z_{k-1}^{\epsilon}-\epsilon_k\, z_k^{\epsilon}-\epsilon_{>k}\cdot U_{>k}^{\epsilon}(z_{\leqslant k}^{\epsilon}), z_k^{\epsilon}, U_{>k}^{\epsilon}(z_{\leqslant k}^{\epsilon}))). \\ &= R(z_{<k})+\epsilon_k\, S_k(z_{<k})+\cdots+\epsilon_n\, S_n(z_{<k}) \end{aligned}$$

and prove that $S_k, \ldots, S_n$ are analytic functions in $z_{<k}$ on $\boldsymbol{\alpha}_{<k}$. The common zeros of (34) and P_{k-1} correspond to the zeros of this resultant with respect to the new coordinates. Setting $z_{k-1}^{\epsilon} = z_{k-1} + \epsilon_{\geqslant} \cdot z_{\geqslant k}$ and $R^{\epsilon}(z_{<k}^{\epsilon}) = 0$ leads to

$$\begin{aligned} R(z_{<k}) &= 0 \\ R'(z_{<k})\, z_i &= S_i(z_{<k}), \quad i \geqslant k, \end{aligned}$$

where $R' = \partial R / \partial z_{k-1}$. Assuming that $\operatorname{Res}_{z_{k-1}}(R, R')$ is not the zero function on $\boldsymbol{\alpha}_{<k-1}$, this almost yields a Kronecker representation for $P_{\geqslant k-1}(z) = 0$.

In order to obtain an actual Kronecker representation, we finally compute the Weierstrass normal form $\tilde{Q}$ of R with respect to z_{k-1}, together with an invertible analytic function W in $z_{<k}$ such that $R = W \tilde{Q}$. Setting $\tilde{U}_i = S_i / W$, we then obtain the Kronecker representation

$$\begin{aligned} \tilde{Q}(z_{<k}) &= 0 \\ \tilde{Q}'(z_{<k})\, z_i &= \tilde{U}_i(z_{<k}), \quad i \geqslant k \end{aligned}$$

for $P_{\geqslant k-1}(z) = 0$ on $\boldsymbol{\alpha}$, where $\tilde{Q}' = \partial \tilde{Q} / \partial z_{k-1}$. If $P_{\geqslant k}(z) = 0$ is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$, then Lemma 18 also implies that $\tilde{Q}$ vanishes on $\mathcal{V}_{\boldsymbol{\alpha}}(P_{\geqslant k-1})$.

9.7. Incremental multiplicity certification

We are now in a position to certify the multiplicity of the system (11) on a sufficiently small ball. We first construct a homotopy H for the system (11) which is in sufficiently general position. For notational reasons, it will be convenient to use z_0 instead of t for the time parameter. We take

$$\begin{aligned} H_i(z_0, \ldots, z_n) &= (1 - z_0)\, P_i(z_1, \ldots, z_n) + z_0\, I_i(z_1, \ldots, z_n) \\ I_i(z_1, \ldots, z_n) &= (z_i + \varepsilon\, z_{i+1} + \cdots + \varepsilon^{n-i}\, z_n)^{d_i} - 1, \end{aligned}$$

where $\varepsilon = 1 / (2\, d_1 \cdots d_n)$.

For $k = n, \ldots, 1$ our main objective is to certify that the system $H_k(z_0, \ldots, z_n) = \cdots = H_n(z_0, \ldots, z_n) = 0$ is equisolvable in $z_k, \ldots, z_n$ on a sufficiently small ball $\boldsymbol{\alpha}$ with $0 \in \boldsymbol{\alpha}_0$ around a numerical solution. For $k = n$, we proceed as in section 8.4.3. Assume now that we have proved equisolvability in $z_{>k}$ for the system $H_{>k}(z) = 0$ on $\boldsymbol{\alpha}$ and that we also have a Kronecker representation

$$\begin{aligned} \hat{Q}(z_{\leqslant k+1}) &= 0 \\ \hat{Q}'(z_{\leqslant k+1})\, z_i &= \hat{V}_i(z_{\leqslant k+1}), \quad i > k+1 \end{aligned}$$

for the same system, where $\hat{Q}' = \partial \hat{Q} / \partial z_{k+1}$. Using the algorithms from section 8.4.3, we may again ensure that $\hat{Q}$ is equisolvable in z_k on $\boldsymbol{\alpha}_{\leqslant k}$. Using the algorithm from section 9.6, we next compute a Kronecker representation

$$\begin{aligned} Q(z_{\leqslant k}) &= 0 \\ Q'(z_{\leqslant k})\, z_i &= V_i(z_{\leqslant k}), \quad i > k \end{aligned}$$

for the intersection $\mathcal{V}_{\boldsymbol{\alpha}}(H_{\geqslant k})$, where $Q' = \partial Q \,/\, \partial z_k$. We will show below that $\operatorname{Res}_{z_k}(Q, Q') \neq 0$ on $\boldsymbol{\alpha}_{<k}$. Furthermore, Q vanishes on $\mathcal{V}_{\boldsymbol{\alpha}}(H_{\geqslant k-1})$, by Lemma 18. We conclude that the system $H_{\geqslant k}(z) = 0$ is equisolvable in $z_{\geqslant k}$ on $\boldsymbol{\alpha}$, by Proposition 17. By induction, this completes our algorithm to certify (almost) multiple roots α on a sufficiently small polydisk $\boldsymbol{\alpha}$ around α.

We still have to check that $\mathrm{Res}_{z_k}(Q, Q')$ is not identical to zero on $\boldsymbol{\alpha}$. Assume the contrary, so that $\mathrm{Res}_{z_k}(Q, Q')$ vanishes everywhere, by analytic continuation. In other words, for any fixed $z_{<k}$, the equation $Q(z) = 0$ admits a multiple solution in z_k. In particular, taking $z_0 = 1$, the generator $q(z_k)$ of the ideal $(I_k, \ldots, I_n) \cap \mathbb{C}[z_k]$ admits a multiple solution. Denoting by $z^1, \ldots, z^s$ the solutions of the system $I_k(z) = \cdots = I_n(z) = 0$, we have $q(z_k) = (z_k - z_k^1) \cdots (z_k - z_k^s)$. Now ε was taken sufficiently small so as to ensure that $z_k^i \neq z_k^j$ for all $i \neq j$, contradicting the assumption that q admits a multiple root.

9.8. Global algebraic certification

Even if the coefficients of the system (11) are all rational or algebraic, then the computed solutions are only numeric approximations. For some applications, it is useful to have exact representations of the solutions. This allows for instance to check whether a given other polynomial with rational coefficients vanishes on the solution set or on some points of the solutions set.

The Kronecker representation provides one useful exact representation for the set of solutions. Modulo a generic linear change of coordinates, we may assume without loss of generality that for any distinct solutions z, z' of (11), their first coordinates z_1 and z_1' are also distinct. Using the algorithm in the proof of Lemma 15, we may then compute a Kronecker representation for the system (11).

Assume now that (11) has rational coefficients and that we have computed numeric approximations $\tilde{z}^1, \ldots, \tilde{z}^k$ of $z^1, \ldots, z^k$ with bit precision p. Using the above method, we may thus compute a numeric approximation of the Kronecker representation with an accuracy of approximately p bits. We apply rational number reconstruction [GG02, Chapter 5] in order to provide a guess for the Kronecker representation with rational coefficients. We may check whether this guess is correct by evaluating P at $z_1 = u, z_2 = V_2(u) / Q'(u), \ldots, z_n = V_n(u) / Q'(u)$ over $\mathbb{A} = \mathbb{Q}[u] / (Q(u))$. If the check fails, then we double the bit precision, use Newton's method to improve the approximations $\tilde{z}^1, \ldots, \tilde{z}^k$, and keep iterating.

In the case that we suspect our system to admit multiple roots, which means that we have at least one herd of numerical solutions, the centers of herds of solutions are still known numerically with a precision of approximately p bits. Consequently, we may use the above method in order to compute the radical of the zero dimensional ideal I generated by the P_i. By evaluating the P_i over suitable algebras with nilpotent elements, we may check in a similar way as above whether the multiplicity of each root matches with the numeric multiplicity and obtain an exact algebraic representation for I. In the projective setting with no solutions at infinity, this strategy can be used to provide a full algebraically certification of all solutions.

10. Systems of analytic equations

It is possible to generalize the techniques of this paper to the local resolution of a system of analytic equations on a polydisk. In this section, we outline some ideas for generalizations of this kind. As a general rule, such generalizations usually work better when the local solutions in the polydisk are well separated from the other solutions outside the polydisk.

10.1. Univariate equations

Consider the equation

$$f(z) = 0,$$

where f is an effective analytic function on the closed unit disk $\mathcal{B}(0, 1)$, without any roots on the unit circle. Assume that we wish to find all roots of f in $\mathcal{B}(0, 1)$. Using the methods from section 8.4.3, the first step is to compute and certify the number d of roots in $\mathcal{B}(0, 1)$.

Having determined the exact number d of roots f inside $\mathcal{B}(0,1)$, most standard numerical methods for finding the d roots of a polynomial of degree d can be mimicked. For instance, we may use the homotopy

$$\lambda\,(z^d-\varepsilon^d)\,t+f(z)\,(1-t)=0$$

for $\lambda=\max|f_k|$ and $0<|\varepsilon|<1$. In the unlucky case that we only find a subset $u_1,\ldots,u_k$ of the roots with $k<d$, we set $P(z)=(z-u_1)\cdots(z-u_k)$ and keep repeating the algorithm using a homotopy of the form

$$\lambda\,(z^{d-k}-\tilde{\varepsilon}^{d-k})\,P(z)\,t+f(z)\,(1-t)=0,$$

and for a new random choice of $\tilde{\varepsilon}$ with $0<|\tilde{\varepsilon}|<1$. This method will ultimately pick random $\tilde{\varepsilon}$ sufficiently close to any of the roots, thereby ensuring the termination of the method. We may also use Aberth's method, while resetting points that fall outside the unit disk to random points inside the disk.

10.2. Direct reduction to the polynomial case

Now consider a system

$$f_1(z)=\cdots=f_n(z)=0 \tag{39}$$

of equations, where $f_1,\ldots,f_n$ are computable analytic functions on the closed unit polydisk $\mathcal{B}(0,1)^n$. From now on, our aim is to determine the solutions of this system in $\mathcal{B}(0,1)^n$.

For any choice of degrees $d_1,\ldots,d_n$, we may consider the truncated polynomials

$$P_i \;=\; \sum_{\|k\|\leqslant d_i}(f_i)_k\,z^k,$$

and consider the truncated system

$$P_1(z)=\cdots=P_n(z)=0. \tag{40}$$

Using the homotopy methods from this paper, we may compute the solutions of this system and keep only those ones which are in $\mathcal{B}(0,1)^n$. At a second stage, we form the homotopy

$$H_i(z,t) \;=\; P_i(z)\,t+f_i(z)\,(1-t)$$

and follow the solutions of the truncated systems from $t=1$ to $t=0$. This yields an uncertified candidate for the set of solutions to (39).

Remark 19. The above method admits several variants. For instance, if we want to avoid the explicit computation of truncated polynomials, then we may directly use the homotopy

$$H_i(z,t) \;=\; \left(z_i^{d_i}-\sigma_i^{d_i}\right)t+\lambda_i\,f_i(z)\,(1-t)$$

for suitable $\sigma_i\in\mathbb{C}$ and $\lambda_i\in\mathbb{R}^{>}$ with $|\sigma_i|=1$. While following the homotopies, we may also decide to drop any paths which lead “too far” outside $\mathcal{B}(0,1)^n$.

Given i, let us now investigate how to pick d_i. For any $d\in\mathbb{N}$, we define

$$\mu_{i,d} \;=\; \max_{\|k\|=d}|(f_i)_k|.$$

Typically, we now take d_i to be the degree d for which $\mu_{i,d}$ is maximal (and in any case larger than this value). We may determine this degree by first computing a bound M_i for $|f_i|$ on a polydisk $\mathcal{B}(0,\rho)^n$ with $\rho>1$, so that $|f_k|\leqslant M_i\,\rho^{-\|k\|}$ for all k, whence $\mu_{i,d}\leqslant M_i\,\rho^{-d}$ for all d. Starting with $\delta:=0$ and $d:=1$, we now repeat the following: if $\mu_{i,d}>\mu_{i,\delta}$, then $\delta:=d$. Else, if $M_i\,\rho^{-d}<\mu_{i,\delta}$, then we stop and take $d_i:=\delta$. Otherwise, set $d:=d+1$ and continue.

Let us now investigate how to certify that we found all solutions in $\mathcal{B}(0,1)^n$. Since (39) is really a perturbation of (40), one idea would be ensure that for each solution of (40), the solutions remain either inside $\mathcal{B}(0,1)^n$ or outside $\mathcal{B}(0,1)^n$ for small perturbations. More precisely, for each i, we start by computing a bound B_i for $|f_i - P_i|$ on $\mathcal{B}(0,1)^n$ (for instance, we may take $B_i = \sum_{\|k\|>d_i} M_i\, \rho^{-k}$ with the above notations, but is sometimes better to take more explicit coefficients for a sharper bound). We next consider the system

$$P_1(z) + \mathcal{B}(0, B_1) = \cdots = P_n(z) + \mathcal{B}(0, B_n) = 0.$$

For each solution z to (40), we now compute a ball enclosure $\mathcal{B}(z, \eta)$ of this solution such that any solution $\tilde{z}$ to $P_1(\tilde{z}) + \varepsilon_1 = \cdots = P_n(\tilde{z}) + \varepsilon_n = 0$ with $|\varepsilon_i| \leqslant B_i$ near z lies in $\mathcal{B}(z, \eta)$. This can be done using the ball version of Newton's method. We next check whether each of the obtained enclosures $\mathcal{B}(z, \eta)$ is either entirely contained in $\mathcal{B}(0,1)^n$ or in its complement. If this is the case, then we have obtained the desired certification. Indeed, consider a solution $\tilde{z} \in \mathcal{B}(0,1)^n$ to (39). Then setting $\varepsilon_i = f_i(\tilde{z}) - P_i(\tilde{z})$, we have $P_i(\tilde{z}) + \varepsilon_i = 0$ and $|\varepsilon_i| \leqslant B_i$, for all i. Hence $\tilde{z} \in \mathcal{B}(z, \eta) \subseteq \mathcal{B}(0,1)^n$ for one of the above enclosures.

Remark 20. One disadvantage of the above certification method is that we have to keep track of both the solutions inside and outside $\mathcal{B}(0,1)^n$ for small perturbations of (40). An alternative idea would be to consider the set $\mathcal{S}$ of all solutions to (39) inside $\mathcal{B}(0,1)^n$ as a generalized point which is the solution of a suitably deflated system, as in we did in section 9.2 for the certification of herd homotopies. This generalized solution is usually unique in a large polydisk, corresponding to large perturbations of the system of equations for $\mathcal{S}$. With some luck, this polydisk actually contains all sets of $|\mathcal{S}|$ points in $\mathcal{B}(0,1)^n$, thereby certifying that $\mathcal{S}$ is the set of all solutions to (39) inside $\mathcal{B}(0,1)^n$.

Unfortunately, this idea only works in dimension 1. For instance, in dimension two, the ball of systems $\mathcal{X}_r = \{x^2 - a\,x - b, y - c\,x - d\}$ with $a, b, c, d \in \mathcal{B}(0,r)$ and $r \in \mathbb{R}^>$ does not contain a system which admits the set $\{(0,0),(0,1)\}$ as its solutions. Nevertheless, it can be checked that any set of two elements in $\mathcal{B}(0,1)^2$ is the solution of a system in either $\mathcal{X}_4$ or $\mathcal{Y}_4$, where $\mathcal{Y}_r = \{y^2 - a\,y - b, x - c\,y - d\}$ with $a, b, c, d \in \mathcal{B}(0,r)$ for any $r \in \mathbb{R}^>$.

10.3. Incremental resolution

As noticed in remark 20, one major disadvantage of the certification method from the previous section is that we need to consider the behaviour of *all* solutions to (40) under small perturbations, and even of those which lie far outside $\mathcal{B}(0,1)^n$.

The incremental geometric resolution technique from the section 9.7 for the certification of multiple roots can also be used for finding and certifying an arbitrary set of roots in a polydisk $\boldsymbol{\alpha}$. For this to work, it is necessary that the system (40) is equisolvable on $\boldsymbol{\alpha}$, which means in practice that we chose a sufficiently general position and that the other roots of (40) outside $\boldsymbol{\alpha}$ are sufficiently far away from $\boldsymbol{\alpha}$.

Remark 21. In the analytic setting, it can be interesting to compute the numeric solutions to (40) in an incremental way as well. A heuristic way to do the intersection step numerically goes as follows. We assume general position and consider the last intersection (the other intersections are done similarly, in fibers with $z_1 = \cdots = z_{k-1} = 0$). Suppose that we are given a Kronecker representation

$$\begin{aligned} Q(0, z_2) &= 0 \\ Q'(0, z_2)\, z_i &= V_i(0, z_2), \quad i > 2 \end{aligned}$$

for the system $f_2(z) = \cdots = f_n(z) = 0$, in the fiber where $z_1 = 0$. Let μ be an upper bound for the number of roots of $f_1(z_1, 0, \ldots, 0)$. For each of the roots ω of the equation $z_1^{\mu} = 1$, we continue our Kronecker representation to a new one in the fiber with $z_1 = \omega$. We finally use the homotopy

$$\begin{aligned} (z_1^{\mu} - 1)\, t + f_1(z)\, (1 - t) &= 0 \\ Q(z_1, z_2) &= 0 \\ Q'(z_1, z_2)\, z_i &= V_i(z_1, z_2), \quad i > 2, \end{aligned}$$

and perform the analytic continuation of the μ known solutions at $t = 1$ until $t = 0$.

In the case when our method fails to prove equisolvability, we need to decompose the system (39) into simpler systems which are equisolvable. This can be done using the following techniques:

- By covering $\mathcal{B}(0,1)^n$ by a finite number of polydisks, each on which the system is equisolvable. For instance in the case of the function Q from example 14, we may cover $\mathcal{B}(0,1)^2$ by the polydisk $\mathcal{B}(0,1) \times \mathcal{B}(0,3)$, or by the polydisk $\mathcal{B}(0, 1/2) \times \mathcal{B}(0, 3/2)$ and several other polydisks of the form $\mathcal{B}(c, r) \times \mathcal{B}(0,1)$.
- By applying a permutation of the coordinates. For instance the function Q from example 14 is equisolvable in $x \in \mathcal{B}(0,1)$ with respect to $y \in \mathcal{B}(0,1)$. We may also consider other linear changes of coordinates modulo adjustment of the regions. For instance, the equation $x\, y = 0$ cannot be made equisolvable by one of the above means. Nevertheless, after setting $x = u + v$ and $y = u - v$, the equation becomes equisolvable on $\mathcal{B}(0,2)^2$.
- By factoring the equation. Especially when an analytic function is restricted to a small region, it frequently occurs that it can be factored into simpler functions on this region. For instance, the equation $x^4 - y^3 + 100\, x\, y = 0$ can be factored as $100\, (x + \cdots)\, (y + \cdots) = 0$ on $\mathcal{B}(0,1)^2$ into two equations $x + \cdots = 0$ and $y + \cdots = 0$ which are equisolvable in x resp. y.

In principle, it is always possible to reduce to the equisolvable case using these techniques, but the number of required subdivisions may quickly grow out of hands. Indeed, for each solution (which we assumed to be single), there exists a sufficiently small neighbourhood where the f_i are essentially linear functionals. The design of a good algorithm for keeping the number of subdivisions as small as possible is beyond the scope of this paper.

Bibliography

[AH83] G. Alefeld and J. Herzberger. *Introduction to interval analysis*. Academic Press, New York, 1983.

[ANS08] ANSI/IEEE. IEEE standard for binary floating-point arithmetic. Technical Report, ANSI/IEEE, New York, 2008. ANSI-IEEE Standard 754-2008. Revision of IEEE 754-1985, approved on June 12, 2008 by IEEE Standards Board.

[BHSW06] D. Bates, J. Hauenstein, A. Sommese and C. Wampler. Bertini: software for numerical algebraic geometry. http://www.nd.edu/~sommese/bertini/, 2006.

[BSHW08] D.J. Bates, A.J. Sommese, J.D. Hauenstein and C.W. Wampler. Adaptive multiprecision path tracking. *SIAM Journal on Numerical Mathematics*, 46(2):722–746, 2008.

[Dur08] C. Durvye. *Algorithmes pour la décomposition primaire des idéaux polynomiaux de dimension nulle donnés en évaluation*. PhD thesis, Univ. de Versailles (France), 2008.

[GG02] J. von zur Gathen and J. Gerhard. *Modern Computer Algebra*. Cambridge University Press, 2-nd edition, 2002.

[GHH+97] M. Giusti, K. Hägele, J. Heintz, J.E. Morais, J.L Montaña and L.M. Pardo. Lower bounds for diophantine approximation. *Journal of Pure and Applied Algebra*, 117–118:277–317, 1997.

[GHMP95] M. Giusti, J. Heintz, J.E. Morais and L.M. Pardo. When polynomial equation systems can be solved fast? In G. Cohen, M. Giusti and T. Mora, editors, *Proc. AAECC'11*, volume 948 of *Lecture Notes in Computer Science*. Springer Verlag, 1995.

[GLSY05] M. Giusti, G. Lecerf, B. Salvy and J.-C. Yakoubsohn. On location and approximation of clusters of zeros of analytic functions. *Foundations of Computational Mathematics*, 5(3):257–311, 2005.

[GLSY07] M. Giusti, G. Lecerf, B. Salvy and J.-C. Yakoubsohn. On location and approximation of clusters of zeros: case of embedding dimension one. *Foundations of Computational Mathematics*, 7(1):1–49, 2007.

[HLRZ00] G. Hanrot, V. Lefèvre, K. Ryde and P. Zimmermann. MPFR, a C library for multiple-precision floating-point computations with exact rounding. http://www.mpfr.org, 2000.

[Hoe02] J. van der Hoeven. Relax, but don't be too lazy. *JSC*, 34:479–542, 2002.

[Hoe05] J. van der Hoeven. Effective complex analysis. *JSC*, 39:433–449, 2005.

[Hoe07] J. van der Hoeven. On effective analytic continuation. *MCS*, 1(1):111–175, 2007.

[Hoe09] J. van der Hoeven. Ball arithmetic. Technical Report, HAL, 2009. Http://hal.archives-ouvertes.fr/hal-00432152.

[Hoe11] J. van der Hoeven. Efficient root counting for analytic functions on a disk. Technical Report, HAL, 2011. Http://hal.archives-ouvertes.fr/hal-00583139.

[JKDW01] L. Jaulin, M. Kieffer, O. Didrit and E. Walter. *Applied interval analysis*. Springer, London, 2001.

[Kea94] R.B. Kearfott. An interval step control for continuation methods. *SIAM J. Numer. Anal.*, 31(3):892–914, 1994.

[Kra69] R. Krawczyk. Newton-algorithmen zur bestimmung von nullstellen mit fehler-schranken. *Computing*, 4:187–201, 1969.

[Kul08] U.W. Kulisch. *Computer Arithmetic and Validity. Theory, Implementation, and Applications*. Number 33 in Studies in Mathematics. De Gruyter, 2008.

[Lan76] S. Lang. *Complex analysis*. Addison-Wesley, 1976.

[Lec01] G. Lecerf. *Une alternative aux méthodes de réécriture pour la résolution des systm̀es algébriques*. PhD thesis, École polytechnique, 2001.

[Lec02] G. Lecerf. Quadratic Newton iteration for systems with multiplicity. *Foundations of Computational Mathematics*, 2(3):247–293, 2002.

[Ley09] A. Leykin. NAG. http://www.math.uic.edu/~leykin/NAG4M2, 2009. Macaulay 2 package for numerical algebraic geometry.

[LVZ06] A. Leykin, J. Verschelde and A. Zhao. Newton's method with deflation for isolated singularities of polynomial systems. *TCS*, 359(1–3):111–122, 2006.

[LVZ08] A. Leykin, J. Verschelde and A. Zhao. *Algorithms in algebraic geometry*, chapter Higher-order deflation for polynomial systems with isolated singular solutions, pages 79–97. Springer, New York, 2008.

[MB96] K. Makino and M. Berz. Remainder differential algebras and their applications. In M. Berz, C. Bischof, G. Corliss and A. Griewank, editors, *Computational differentiation: techniques, applications and tools*, pages 63–74. SIAM, Philadelphia, 1996.

[MB04] K. Makino and M. Berz. Suppression of the wrapping effect by Taylor model-based validated integrators. Technical Report MSU Report MSUHEP 40910, Michigan State University, 2004.

[MKC09] R.E. Moore, R.B. Kearfott and M.J. Cloud. *Introduction to Interval Analysis*. SIAM Press, 2009.

[MM11] A. Mantzaflaris and B. Mourrain. Deflation and certified isolation of singular zeros of polynomial systems. In A. Leykin, editor, *Proc. ISSAC'11*, pages 249–256. San Jose, CA, US, Jun 2011. ACM New York.

[Moo66] R.E. Moore. *Interval Analysis*. Prentice Hall, Englewood Cliffs, N.J., 1966.

[Mor87] A.P. Morgan. *Solving polynomial systems using continuation for] engineering and scientific problems*. Prentice-Hall, Englewood Cliffs, N.J., 1987.

[Neu90] A. Neumaier. *Interval methods for systems of equations*. Cambridge university press, Cambridge, 1990.

[OWM83] T. Ojika, S. Watanabe and T. Mitsui. Deflation algorithm for multiple roots of a system of nonlinear equations. *J. of Math. An. and Appls.*, 96(2):463–479, 1983.

[RG10] S. Rump and S. Graillat. Verified error bounds for multiple roots of systems of nonlinear equations. *Num. Algs.*, 54:359–377, 2010.

[Rum80] S.M. Rump. *Kleine Fehlerschranken bei Matrixproblemen*. PhD thesis, Universität Karlsruhe, 1980.

[Rum10] S.M. Rump. Verification methods: rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010.

[Sch82] A. Schönhage. The fundamental theorem of algebra in terms of computational complexity. Technical Report, Math. Inst. Univ. of Tübingen, 1982.

[SW05] A.J. Sommese and C.W. Wampler. *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*. World Scientific, 2005.

[Ver96] J. Verschelde. *Homotopy continuation methods for solving polynomial systems*. PhD thesis, Katholieke Universiteit Leuven, 1996.

[Ver99] J. Verschelde. PHCpack: a general-purpose solver for polynomial systems by homotopy continuation. *ACM Transactions on Mathematical Software*, 25(2):251–276, 1999. Algorithm 795.

[Wei00] K. Weihrauch. *Computable analysis*. Springer-Verlag, Berlin/Heidelberg, 2000.